



UNIVERSIDAD DE GRANADA

**PROGRAMA OFICIAL DE POSGRADO EN ESTUDIOS AVANZADOS DE TRADUCCIÓN
E INTERPRETACIÓN**

**RECURSOS Y HERRAMIENTAS LINGÜÍSTICOS
PARA LOS SISTEMAS DE BÚSQUEDAS DE
RESPUESTAS MONOLINGÜES Y
MULTILINGÜES**

JUNCAL GUTIÉRREZ ARTACHO

**TESIS DOCTORAL
JUNIO, 2014**

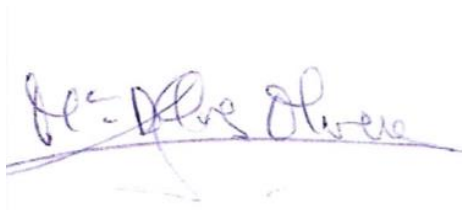
Editor: Universidad de Granada. Tesis Doctorales
Autora: Juncal Gutiérrez Artacho
ISBN: 978-84-9125-160-6
URI: <http://hdl.handle.net/10481/40371>

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

La doctoranda D^a **Juncal Gutiérrez Artacho** y los directores de la tesis **Dr. D^a María Dolores Olvera Lobo** y **Dr. D. Bryan John Robinson Fryer** garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctoranda bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, a 25 de abril de 2015

Director/es de la Tesis



Fdo.: María Dolores Olvera Lobo

Doctoranda



Fdo.: Juncal Gutiérrez Artacho



Fdo.: Bryan John Robinson Fryer

0

PRESENTACIÓN

Normativa

La presente tesis doctoral ha sido elaborada de acuerdo con la modalidad "compendio de publicaciones". Se ha preparado, pues, una memoria de trabajos previamente publicados en revistas científicas internacionales indexadas en las bases de datos del Journal Citation Report de Thomson Reuters y/o del SCImago Journal & Country Rank de Scopus, y un capítulo de libro en una editorial de prestigio nacional.

La memoria ha sido redactada de acuerdo con lo establecido en las "Normas reguladoras de las enseñanzas oficiales de Doctorado y del título de Doctor por la Universidad de Granada", recogidas en el documento NCG61/1, que fue aprobado en la sesión extraordinaria del Consejo de Gobierno de 2 de mayo de 2012.

Estas normas establecen:

"[La tesis doctoral] debe contar, al menos, con los siguientes contenidos: título, resumen, introducción, objetivos, metodología, resultados, conclusiones y bibliografía. Una tesis doctoral puede también consistir en el reagrupamiento en una memoria de trabajos de investigación publicados por el doctorando en medios científicos relevantes en su ámbito de conocimiento.

Los artículos que configuren la tesis doctoral deberán estar publicados o aceptados con fecha posterior a la obtención del título de grado y del máster universitario, no podrán haber sido utilizados en ninguna tesis anterior y se deberá hacer mención a la Universidad de Granada a través de la afiliación del doctorando.

Si la publicación ha sido realizada por varios autores, además del doctorando, se debe adjuntar la declaración de los restantes autores de no haber presentado dicha publicación en otra tesis

doctoral o la renuncia a hacerlo. Los coautores señalarán el trabajo del doctorando en los mencionados trabajos.

En esta modalidad, la tesis debe tener además de los apartados mencionados en el punto anterior, los artículos que la componen, bien integrados como capítulos de la tesis o bien como un Anexo.

El Comité de Dirección de la Escuela de Doctorado correspondiente podrá establecer el número mínimo de artículos necesarios para presentar una tesis en esta modalidad y las condiciones adicionales sobre la calidad de los trabajos".

Por otra parte, la Escuela Internacional de Posgrado de la Universidad de Granada ha dictado que:

"en la cubierta y en la portada de la memoria figurará: "Universidad de Granada", junto con el escudo institucional.

Programa de Doctorado.

Título de la Tesis.

Autor.

La tesis debe contar, al menos, con los siguientes contenidos: título, compromiso de respeto derechos de autor, resumen, introducción, objetivos, metodología, resultados, conclusiones y bibliografía.

En la tesis doctoral deberá figurar en las páginas iniciales una hoja donde tanto el doctorando como sus directores garanticen que se han respetados los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones".

Fuente:

http://escuelaposgrado.ugr.es/pages/doctorado/tramites_administrativos_alumnos_doctorado/tesis_doctoral [consultado: 10 de mayo de 2015].

En cuanto a la modalidad "compendio de publicaciones" la Escuela Internacional de Posgrado de la Universidad de Granada ha publicado estas directrices:

"La recomendación del Consejo Asesor de Doctorado es que para utilizar este tipo de formato se utilice un mínimo de tres artículos y que se incluya un informe con el factor de impacto de las publicaciones presentadas. En aquellas áreas en las que no sea aplicable este criterio se sustituirá por las bases relacionadas por la Comisión Nacional Evaluadora de la Actividad Investigadora (CNEAI) para estos campos científicos.

Las Comisiones Académicas de los Programas de Doctorado podrían requerir condiciones adicionales a las reflejadas en esta página web, debería ponerse en contacto con el programa para conocer si existen estas condiciones.

El doctorando debe respetar los derechos de propiedad intelectual relativos a la difusión de los artículos utilizados en la tesis doctoral".

Fuente:

http://escuelaposgrado.ugr.es/pages/doctorado/tramites_administrativos_alumnos_doctorado/tesis_doctoral [consultado: 10 de abril de 2015].

En cuanto a la modalidad "Mención internacional" la Escuela Internacional de Posgrado de la Universidad de Granada publica las siguientes directrices:

- **Idioma de presentación de la tesis:** parte de la tesis doctoral, al menos un resumen extenso y las conclusiones, debe redactarse y presentarse en una de las lenguas habituales para la comunicación científica en el campo de conocimiento en el que se enmarca el trabajo de investigación, distinta a cualquiera de las lenguas oficiales en España. Esta norma no será de aplicación cuando las estancias, informes y expertos/as procedan de un país de habla hispana.
- **Informes de expertos/as:** la tesis debe ser informada por un mínimo de dos expertas/os doctoras/es pertenecientes a alguna institución de educación superior o instituto de investigación no española (Informe de Experto (word-odt)). El informe se redactará en español o en inglés. La solicitud de estos informes se realizará una vez que el/la director/a o directores/as hayan autorizado la presentación de la tesis. Las/los expertas/os deberán cumplir los mismos requisitos que se exigen a los/las miembros del tribunal (se adjuntará informe de idoneidad).
- **Composición del tribunal:** debe formar parte del tribunal al menos un/a experto/a perteneciente a alguna institución de educación superior o centro de investigación no española, con el título de doctor/a, distinto del/de la responsable de la estancia y distinto de las/los expertas/os firmantes de los informes.

Las posteriores actuaciones con respecto a la mención de doctorado internacional, incluida la Solicitud de la misma forman parte del procedimiento de Depósito de Tesis Doctoral que podrá consultar en esta web.

FUENTE: <http://escuelaposgrado.ugr.es/pages/internacional/mencioninternacional/mencion> [consultado: 10 de abril de 2015].

Estancias

Para cumplir los requisitos de la normativa para la mención internacional, D^a Juncal Gutiérrez Artacho realizó una estancia de investigación de una duración de tres meses en el Instituto Harriman de la Universidad de Columbia [Nueva York, Estados Unidos], financiada por el programa de “Ayudas complementarias para beneficiarios de ayudas (FPU): Estancias Breves y Traslados Temporales (Convocatoria 2013)” del Ministerio de Educación, Cultura y Deportes. La Universidad de Columbia cuenta con uno de los grupos más punteros en el ámbito objeto de esta tesis doctoral, el de *Natural Language Processing*, y alberga a los principales investigadores en esta área –como K. McKeown, D.R. Radev, S. Blair-Goldensohn, M. Lee, J. Ely y J. Cimino–, siendo algunos de los precursores del principal foro en Recuperación de Información, la muy prestigiosa *Text REtrieval Conference* (TREC). Además, esta Universidad se encuentra en el puesto 8 en el listado mundial de Shangai, según este mismo ranking, en esta área la Universidad se encuentra en el puesto 11.

Publicaciones

A la vista de la citada normativa, hemos preparado una memoria con los siguientes contenidos: título, autorización de los directores de la tesis doctoral, compromiso de respeto de los derechos de autor, agradecimientos, resumen (en castellano y en inglés), presentación, introducción – donde se incluye los objetivos y la metodología–, resultados, conclusiones y bibliografía. Dentro del bloque de resultados se incluyen los textos íntegros de los ocho artículos originales ya publicados en revistas de impacto, que abordan desde distintas perspectivas los materiales que son objeto de estudio en esta investigación. En los siguientes párrafos se puede encontrar información a fecha de hoy sobre el factor de impacto de las ocho publicaciones presentadas:

1. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011) “Language resources used in multi-lingual Question Answering Systems”. *Online Information Preview*. 35 (4): 543 – 557

Citas recibidas:

Según Web of Science: 1

Según Google Scholar: 5

Según la revista: 3

La revista *Online Information Review* (ISSN: 1468-4527) es publicada por la editorial Emerald Group Publishing Limited (Reino Unido).

H Index: 30

Factor de impacto: 0,939 en el año 2011 dentro del ISI Journal Citation Reports.

Factor de impacto: 0,828 en el año 2011 dentro del Scimago Journal & Country Rank.

Posición en su categoría:

32 (Q2) de 83 en la categoría “Information Science & Library Science” dentro del ISI Social Sciences Citation Index en 2011.

30 (Q1) de 189 en la categoría “Library and Information Science” dentro del Scimago Journal & Country Rank en 2011.

Está indexada en las siguientes bases de datos:

- Scopus
- Thomson Reuters Journal Citation Report
- Google Scholar
- Journal Citation Reports (JCR)
- Web of Science: Social Sciences Citation Index (SSCI), entre otras bases de datos nacionales e internacionales de relevancia.

2. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2012) "Language Resources for Translation in Multi-lingual Question Answering Systems." *Translation Journal*, 16 (2).

La revista *Translation Journal* (ISSN: 1536-7207) es una revista *online* publicada desde 1997.

3. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2015) "Question Answering Track Evaluation in TREC, CLEF and NTCIR". In: *New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing*, 353 (1): 13-22, Berlín: Springer International Publishing.

Publicado en abril de 2015 (sin citas).

La serie *Advances in Intelligent Systems and Computing* (ISSN: 2194-5357) es una de las series con mayor prestigio en su ámbito publicada por la editorial Springer International Publishing. Desde 1997, anteriormente llamada *Advances in Intelligent and Soft Computing*, contiene publicaciones sobre teoría, aplicaciones y métodos de diseño en Sistemas de Inteligencia e Inteligencia Computacional.

H Index: 12

Factor de impacto: 0,139 en el año 2013 dentro del Scimago Journal & Country Rank.

Posición en su categoría:

251 (Q4) de 435 en la categoría "Computer Science" dentro del Scimago Journal & Country Rank en 2013.

Está indexada en las siguientes bases de datos:

- ISI Proceedings,
- EI-Compendex,
- DBLP,
- Scopus,
- Google Scholar
- Springerlink, entre otras bases de datos nacionales e internacionales de relevancia.

4. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2010) "Question-Answering Systems as Efficient Source of Terminological Information: Evaluation". *Health Information and Library Journal*. 27 (4): 268 – 276.

Citas recibidas:

Según Web of Science: 5

Según Scopus: 7

Según Google Scholar: 13

La revista *Health Information and Library Journal* (ISSN: 0165-5515, 1741-6485) es publicada por la editorial Wiley Online Library (Reino Unido).

H Index: 24

Factor de impacto: 0,761 en el año 2010 dentro del ISI Journal Citation Reports.

Factor de impacto: 0,619 en el año 2010 dentro del Scimago Journal & Country Rank.

Posición en su categoría:

39 (Q2) de 77 en la categoría "Information Science & Library Science" dentro del ISI Social Sciences Citation Index en 2010.

47 (Q2) de 193 en la categoría "Library and Information Science" dentro del Scimago Journal & Country Rank en 2010.

Está indexada en las siguientes bases de datos:

- Scopus
- Thomson Reuters Journal Citation Report
- Google Scholar
- MEDLINE/PubMed (NLM)
- Journal Citation Reports (JCR)
- Web of Science: Social Sciences Citation Index (SSCI), entre otras bases de datos nacionales e internacionales de relevancia.

5. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011) "Open- vs. Restricted- Domain Question Answering Systems in the Biomedical Field". *Journal of Information Science*. 37 (2): 152-162

Citas recibidas:

Según Web of Science: 2

Según Scopus: 5

Según Google Scholar: 8

La *Journal of Information Science* (ISSN: 0165-5515, 1741-6485) es publicada por la editorial SAGE (Reino Unido).

H Index: 40

Factor de impacto: 1,299 en el año 2011 dentro del ISI Journal Citation Reports.

Factor de impacto: 0,978 en el año 2011 dentro del Scimago Journal & Country Rank.

Posición en su categoría:

24 (Q2) de 83 en la categoría "Information Science & Library Science" dentro del ISI Social Sciences Citation Index en 2011.

21 (Q1) de 189 en la categoría "Library and Information Science" dentro del Scimago Journal & Country Rank en 2011.

Está indexada en las siguientes bases de datos:

- Scopus

- Thomson Reuters Journal Citation Report
- Google Scholar
- Journal Citation Reports (JCR)
- Web of Science: Social Sciences Citation Index (SSCI), entre otras bases de datos nacionales e internacionales de relevancia.

6. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2013) "Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general". *Revista Española de Documentación Científica*, 36(2), abril-junio 2013, e009 ISSN-L:0210-0614.

doi: <http://dx.doi.org/10.3989/redc.2013.2.921>

Citas recibidas:

Según Google Scholar: 2

Según la revista: 3

La *Revista Española de Documentación Científica*, también conocida como la REDC, (ISSN: 0210-0614) es publicada por el Consejo Superior de Investigaciones Científicas (España). La REDC publica artículos originales de investigación experimental o teórica, previamente sujetos a un proceso de selección y evaluación por pares.

H Index: 12

Factor de impacto: 0,879 en el año 2013 dentro del ISI Journal Citation Reports.

Factor de impacto: 0,385 en el año 2013 dentro del Scimago Journal & Country Rank.

Posición en su categoría:

40 (Q2) de 84 en la categoría "Information Science & Library Science" dentro del ISI Social Sciences Citation Index en 2013.

90 (Q2) de 205 en la categoría "Library and Information Science" dentro del Scimago Journal & Country Rank en 2013.

Está indexada en las siguientes bases de datos:

- Scopus
- Journal Citation Reports (JCR)

- Thomson Reuters Journal Citation Report
- Google Scholar
- Web of Science: Social Sciences Citation Index (SSCI), entre otras bases de datos nacionales e internacionales de relevancia.

7. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011) "Multilingual Question-Answering System in biomedical domain on the Web: an evaluation". *Multilingual and Multimodal Information Access Evaluation. Lecture Notes in Computer Science*, 6941: 83-88, Berlín: Springer International Publishing.

Citas recibidas:

Según Web of Science: 1

Según Scopus: 1

Según Google Scholar: 7

La serie *Lecture Notes in Computer Science* (ISSN: 1611-3349, 0302-9743) es publicada por la editorial Springer Verlag (Alemania).

H Index: 118

Factor de impacto: 0,307 en el año 2011 dentro del Scimago Journal & Country Rank.

Posición en su categoría:

123 (Q1) de 405 en la categoría "Computer Science" dentro del Scimago Journal & Country Rank en 2011.

Está indexada en las siguientes bases de datos:

- ISI Conference Proceedings Citation Index - Science (CPCI-S), included in ISI Web of Science
- El Engineering Index (Compendex and Inspec databases)
- ACM Digital Library
- Google Scholar, entre otras bases de datos nacionales e internacionales de reconocido prestigio.

8. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2014) "Satisfacción de usuarios del ámbito de la traducción en el uso de sistemas de búsqueda multilingüe de respuestas como recurso de información terminológica". *TIC, trabajo colaborativo e interacción en Terminología y Traducción*. Interlingua, Granada: Comares

El libro *TIC, trabajo colaborativo e interacción en Terminología y Traducción* (ISBN: 9788490450468) pertenece a la colección *Interlingua* de la editorial Comares (España), una de las colecciones más prestigiosas de Traducción e Interpretación en lengua española (130 títulos publicados entre 1996 y 2015) y está recogida en el MLA International Index desde 2006.

0

RESUMEN

En el entorno de la Web la sobrecarga de información se deja sentir aún más que en otros contextos. De esta forma, en demasiadas ocasiones, al plantear una determinada consulta en las herramientas de búsqueda de información web (buscadores, directorios o metabuscadores) el número de páginas web recuperadas resulta excesivo y no todas ellas son relevantes ni útiles para los objetivos del usuario. Los sistemas de búsqueda de respuestas (o SBR) se presentan como una alternativa a los tradicionales sistemas de Recuperación de Información (o RI) tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de mostrar al usuario una lista de documentos relacionados con su búsqueda (Jackson y Schilder, 2005). El funcionamiento de los SBR se basa en los modelos de respuestas cortas (Blair-Goldensohn *et al.*, 2004), y la ventaja principal que ofrece al usuario es que éste no ha de consultar documentos completos para obtener la información requerida puesto que el sistema ofrece la respuesta correcta en forma de un número, un sustantivo, una frase corta o un fragmento breve de texto (Pérez-Coutiño *et al.*, 2004).

Dentro de los tipos de SBR el presente trabajo centra parte de su estudio al análisis y evaluación de los SBR multilingües y translingües. Estos sistemas necesitan la incorporación de algún tipo de recurso lingüístico, herramienta o técnica de traducción para la correcta recuperación del resultado, ya que el SBR puede extraer la respuesta de una colección de documentos escritos en una lengua distinta a la que se plantea la pregunta.

Puesto que la búsqueda de respuestas monolingüe y multilingüe se presenta como un avance destacado en la mejora de la RI (Kolomiyets y Moens, 2011) se hace necesario determinar su eficacia para el usuario final. Con este objetivo se han realizado ocho estudios donde se evalúa: a) en los dos primeros los recursos y herramientas lingüísticos utilizados por estos sistemas para la recuperación multilingüe o translingüe (*artículo 1; artículo 2*); b) el rendimiento y la calidad de las respuestas de los principales SBR monolingües y multilingües de dominio general y dominio especializado disponibles en la Web (QuALiM, SEMOTE, START, TrueKnowledge y HONqa) ante preguntas de diversos tipos (de definición, factuales y de lista) y temas (Arte y Literatura, Biología, Medicina, Personajes, Historia, Economía o Deportes, entre otros), para lo que se

aplican diferentes medidas de evaluación (*artículo 3, artículo 4, artículo 5, artículo 6, artículo 7*); y finalmente, c) la satisfacción y las necesidades de los usuarios finales sobre los SBR multilingües como recurso terminológico. Todos los estudios presentados en la presente tesis doctoral han sido evaluados y aceptados por la comunidad científica, estando disponibles en las principales revistas y editoriales internacionales con revisión a pares.

A continuación se detalla el análisis realizado en cada uno de los trabajos presentados:

1ª Publicación: LANGUAGE RESOURCES USED IN MULTI-LINGUAL QUESTION ANSWERING SYSTEMS

Propósito de este artículo: En el campo de la RI, algunas herramientas multilingües se han creado para ayudar a los usuarios a superar las barreras lingüísticas.

Aunque, estas herramientas no están desarrolladas completamente y es necesario investigar más en su mejora y aplicación. Uno de los principales problemas que presentan es la elección de un correcto recurso lingüístico que ofrezca una mejor cobertura y solucione los problemas de traducción en el contexto de la RI translingüe.

Diseño: La investigación se centra en el análisis de los recursos utilizados por los SBR multilingües, que responden a las preguntas de los usuarios con respuestas cortas, en lugar de ofrecerles una lista de documentos relacionados con la búsqueda. Se ha realizado un análisis de las principales publicaciones sobre los SBR multilingües con el objetivo de identificar la tipología, las ventajas y desventajas, y el uso real y tendencia de cada uno de los recursos y herramientas lingüísticos usados en este nuevo tipo de sistemas.

Resultados: Se han identificado y estudiado cinco de los recursos más usados en los SBR multilingües: base de datos, diccionarios, corpus, ontologías y tesauros. Los tres recursos más populares tradicionalmente (traductores automáticos, diccionarios, y corpus) han ido gradualmente dejando un espacio abierto a otros, como las ontologías y la enciclopedia *online* Wikipedia.

¿Qué aporta este artículo?: La perspectiva ofrecida por las técnicas de traducción puede mejorar la efectividad de los SBR.

2ª Publicación: LANGUAGE RESOURCES IN MULTI-LINGUAL QUESTION ANSWERING SYSTEMS

En el campo de la RI, las herramientas monolingües, multilingües y translingües se han creado para que puedan ayudar a los especialistas en sus tareas; así como ayudar al resto de usuarios a encontrar una variedad amplia de información. Las herramientas translingües van evolucionando

aunque todavía es necesario varios años de estudio e investigación para mejorar sus funcionalidades. Una de las principales dificultades a las que se enfrentan es la tarea de la traducción de las preguntas planteadas (Diekema, 2003). Dado la actual expansión en investigación, desarrollo y creación de sistemas de RI multilingüe o translingüe, también denominados CLIR por sus siglas en inglés, se ha considerado como esencial el analizar y evaluar los recursos usados por un tipo de estos sistemas, los SBR multilingües.

Como objetivo general se ha analizado e introducido las técnicas de traducción en el estudio de los SBR multilingües. El segundo objetivo ha sido identificar y analizar los principales tipos de recursos y herramientas lingüísticos utilizados en los procesos de CLIR asociados a los SBR translingües, y determinar cuál es el uso real que hacen estos sistemas de los mismos.

3º Publicación: **QUESTION ANSWERING TRACK EVALUATION IN TREC, CLEF AND NTCIR**

Los SBR se presentan como una alternativa real a los sistemas de RI (SRI), ya que permiten al usuario obtener una respuesta rápida y comprensible a sus necesidades de información. Hace ya 15 años desde que el foro especializado en RI, TREC, introdujo la primera sección o *track* dedicada exclusivamente a la búsqueda de información. En las principales conferencias en evaluación de RI se han creado *tracks* específicas centradas en el desarrollo y evaluación de este tipo de sistemas. El presente estudio realiza una breve revisión de los foros TREC, CLEF y NTCIR desde la perspectiva de la búsqueda de respuestas. Para ello, se presenta un resumen del panorama histórico de los 15 años en las *tracks* sobre evaluación en búsqueda de respuestas mediante el método de revisión sistemática. Hemos identificado los diferentes *tasks* o *labs* específicos creados en cada QA *track*, el tipo de preguntas de evaluación utilizadas, así como las medidas de evaluación utilizadas en las diferentes competiciones analizadas.

Los resultados reflejan que es CLEF el foro que ha aplicado una variedad más amplia de tipos de preguntas de evaluación (factuales, de definición, de lista, causales, sí/no, entre otras). NTCR, celebrada en 13 ocasiones, es el foro en el que se han usado un mayor número de medidas de evaluación diferentes. Las medidas tradicionales de exactitud, precisión y exhaustividad siguen siendo las tres medidas más usadas en las tres competiciones.

4ª Publicación: **QUESTION-ANSWERING SYSTEMS AS EFFICIENT SOURCE OF TERMINOLOGICAL INFORMATION: EVALUATION**

Los SBR se presentan como una nueva alternativa a los SRI. La mayoría de los usuarios a menudo necesitan recuperar información específica a preguntas factuales en lugar de documentos completos. Hemos realizado un estudio para evaluar la eficiencia de los SBR como fuentes terminológicas para médicos, traductores especializados y usuarios en general. Con este

fin se ha analizado el funcionamiento de un SBR de dominio abierto, START, y uno de dominio especializado, MedQA. El estudio se ha realizado con una colección de doscientas preguntas de definición (*What is...?*), tanto especializadas como generales, del sitio web WebMed. Se ha estudiado las fuentes que los SBR usan para recuperar las respuestas, y tras ello se ha aplicado una serie de medidas de evaluación para comprobar la calidad de las respuestas. Ambos sistemas han demostrado ser apropiados para la recuperación de terminología al proporcionar fuentes fiables y respuestas correctas.

5ª Publicación: **OPEN- VS. RESTRICTED-DOMAIN QA SYSTEMS IN THE BIOMEDICAL FIELD**

Los SBR se presentan como una alternativa a los sistemas tradicionales de RI tratando de ofrecer respuestas precisas a preguntas factuales. Se ha realizado un estudio para evaluar la eficiencia de estos sistemas como fuentes terminológicas para los especialistas y para usuarios en general. Con este fin, se ha evaluado el funcionamiento de cuatro SBR, dos especializados en el dominio biomédico (MedQA y HONqa) y dos de dominio general (START y QuALiM). El estudio ha utilizado una colección de 150 preguntas biomédicas definicionales (*What is...?*), obtenidas del sitio web médico WebMD. Para determinar el funcionamiento, se han evaluado las respuestas ofrecidas utilizando una serie de medidas específicas (precisión, MRR, TRR, FHS).

El estudio permite confirmar que los cuatro sistemas son útiles para la recuperación de información definicional en este ámbito, ya que han proporcionado respuestas coherentes y precisas con un grado de aceptabilidad adecuado.

6ª Publicación: **EVALUACIÓN DEL RENDIMIENTO DE LOS SISTEMAS DE BÚSQUEDA DE RESPUESTAS DE DOMINIO GENERAL**

Los SBR son una alternativa a los tradicionales sistemas de RI tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de presentar al usuario una lista de documentos relacionados con su búsqueda. Se ha evaluado la eficacia de cuatro SBR disponibles en la Web –QuaLiM, SEMOTE, START, y TrueKnowledge–, mediante una amplia muestra de preguntas de definición, factuales y de lista, pertenecientes a distintos dominios temáticos. Se utilizó una colección de 500 preguntas cuyas respuestas fueron valoradas por los usuarios y, posteriormente, se aplicaron varias medidas para su evaluación (MRR, TRR, FHS, MAP y precisión). Se observa que START y TrueKnowledge presentan un nivel aceptable de respuestas correctas, precisas y en una secuencia bien ordenada. Los resultados obtenidos revelan el potencial de esta clase de herramientas en el ámbito del acceso y la RI de dominio general.

7ª Publicación: **MULTILINGUAL QUESTION-ANSWERING SYSTEM IN BIOMEDICAL DOMAIN ON THE WEB: AN EVALUATION**

Los SBR son una alternativa a los tradicionales sistemas de RI tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales. El presente estudio evalúa la efectividad de estos sistemas como fuentes terminológicas para especialistas, así como para los usuarios en general, en el contexto de las búsquedas multilingües en un dominio restringido. Con este objetivo, se han analizado los resultados obtenidos por el SBR multilingüe de dominio restringido HONqa, disponible en la Web. En el estudio se utilizó un conjunto de 120 preguntas biomédicas de definición (*What is...?*), obtenidas del sitio web médico WebMD, en donde se pueden formular en inglés, francés e italiano. Las respuestas se han analizado aplicando una serie de medidas específicas (MRR, TRR, FHS, precisión, exhaustividad, MAP). También han sido analizadas las fuentes usadas para obtener las respuestas.

El estudio confirma que para todas las lenguas en las que se analizó el funcionamiento de la efectividad es necesario mejorar, aunque en este contexto multilingüe las preguntas en lengua inglesa obtuvieron mejores resultados en la recuperación de información definicional que en francés y en italiano. Uno de los motivos se puede deber a que las fuentes de información de las que se extraen las respuestas son más numerosas en inglés y tienen una estructura más apropiada para este propósito que en las otras lenguas estudiadas.

8ª Publicación: **SATISFACCIÓN DE USUARIOS DEL ÁMBITO DE LA TRADUCCIÓN EN EL USO DE SISTEMAS DE BÚSQUEDA MULTILINGÜE DE RESPUESTAS COMO RECURSO DE INFORMACIÓN TERMINOLÓGICA**

Con el rápido crecimiento de Internet y el desarrollo de las nuevas tecnologías en los últimos años, los SBR se han convertido en una alternativa a los tradicionales sistemas de RI. Aunque existe una prolífera producción sobre estos sistemas, pocos son los trabajos que se han desarrollado en la evaluación centrada en el usuario. El presente trabajo se centra exclusivamente en la evaluación de los SBR multilingües ya que permite al usuario acceder a información terminológica no disponible en su lengua, y en la evaluación centrada en el usuario para entender las necesidades del usuario e identificar las dimensiones y factores en el desarrollo de un sistema de información con el fin de mejorar su aceptación.

El objetivo es conocer el grado de satisfacción del usuario para el SBR multilingüe HONqa (inglés, francés e italiano) utilizando la herramienta de los investigadores Ong y colegas. (2009) basado en un examen de modelos y teorías. Para ello, se ha analizado el grado de satisfacción de los profesionales de la traducción como recurso de información terminológica.

**Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües:
Análisis y evaluación**

Un total de 122 especialistas en inglés, francés o italiano ha utilizado y evaluado el recurso terminológico desde la perspectiva de la traducción.

Tras el análisis de los resultados obtenidos respecto al SBR HONqa, los cuales han permitido evaluar la satisfacción de los estudiantes de Traducción, se ha comprobado que el sistema les resulta a los usuarios-alumnos fácil y útil para la recuperación de información terminológica en todos los idiomas.

0**ABSTRACT**

Within the sphere of the Web, the overload of information is more notable than in other contexts. Thus, too often, on planning a search with tools from the Web (search engines, directories, or meta-search engines), the number of web pages found proves excessive and not all of them are relevant or useful for the objectives of the user. Question answering systems (QAS) are presented as an alternative to the traditional Information Retrieval (IR) systems, seeking to offer precise and understandable answers to factual questions instead of showing the user a list of documents related to a given search (Jackson and Schilder, 2005). The functioning of QAS is based on short-answer models (Blair-Goldensohn *et al.*, 2004), and the main advantage that it offers is that the user does not have to consult complete documents to collect the information needed, as the system provides the correct answer in the form of a number, a noun, a short phrase, or a brief fragment of text (Pérez-Coutiño *et al.*, 2004a).

In relation to the types of QAS, the present work focuses on the analysis and assessment of multilingual and cross-language QAS. These systems need to include some type of linguistic translation resource, tool, or technique for the proper retrieval of the result, since the QAS can retrieve the answer from a collection of documents written in languages differing from the one in which the question is formulated.

Given that the QAS is presented as a substantial advance in the improvement of IR (Kolomiyets and Moens, 2011), it becomes necessary to determine its effectiveness for the final user. With this aim, 8 studies were undertaken to evaluate: a) in the first two, the linguistic resources and tools used in these systems for multilingual retrieval (*article 1; article 2*); b) the performance and quality of the answers of the main monolingual and multilingual QA of general domain and specialized domain in the Web (QuALiM, SEMOTE, START, TrueKnowledge, and HONqa) in response to different types of questions (of definition, of facts, and of lists) and subjects (e.g. art, literature, biology, medicine, names, history, economy, or sports), so that different evaluation means can be applied (*article 3, article 4, article 5, article 6, article 7*); and finally c) the satisfaction and needs of the final users on the multilingual QAS as a terminological resource. All the studies presented in this doctoral thesis have been evaluated and

accepted by the scientific community, being available in the main international journals and publishing houses with peer reviews.

Below, the analysis made in each of the works presented is discussed in detail.

Paper 1: LANGUAGE RESOURCES USED IN MULTI-LINGUAL QUESTION ANSWERING SYSTEMS

Purpose of this paper: In the field of IR, some multilingual tools are being created to help the users to overcome the language barriers.

Nevertheless, these tools are not developed completely and it is necessary to investigate more for their improvement and application. One of their main problems is the choice of the linguistic resources to offer better coverage and to solve the translation problems in the context of the multilingual IR.

Designs: This research is focused in the analysis of resources used by the multilingual QAS, which respond to users' queries with short answers, rather than just offering a list of documents related to the search. We have carried out an analysis of the main publications about the multilingual QAS with the aim of identifying the typology, the vantages and advantages, and the real use and trend of each one of linguistic resources and tools used in this new kind of system.

Findings: We identified and studies five of the most used resources used in the cross-languages QAS: databases, dictionaries, corpora, ontologies and thesauri. The three most popular traditional resources (automatic translators, dictionaries, and corpora) are gradually leaving a widening gap for others – such as ontologies and the free encyclopaedia Wikipedia.

What is value of paper: The perspective offered by the translation discipline can improve the effectiveness of QA systems.

Paper 2: LANGUAGE RESOURCES IN MULTI-LINGUAL QUESTION ANSWERING SYSTEMS

In the field of IR monolingual and multilingual tools are being created that can greatly assist specialists in their work; as well as helping other users find a wide variety of information. Multilingual tools are evolving but several years of study and research are still needed to improve implementations. One of the main difficulties facing these tools is the task of translating queries made by users and the documentary sources found in response (Diekema, 2003). Given the current expansion in research, development, and the creation of multilingual IR systems, it was considered worthwhile analysing and evaluating the resources used by one type of these systems, the multilingual QAS.

Our article is primarily intended as a general purpose analysis and aims to encompass translation in the study of multilingual QAS. The second general aim is to identify and analyse the linguistic resources and tools found in these systems. Specific objectives include identifying the main types of language resources and tools useful in the multilingual IR processes associated with multilingual QAS, and establishing how much use is made of these tools by multilingual QAS.

Paper 3: QUESTION ANSWERING TRACK EVALUATION IN TREC, CLEF AND NTCIR

QAS are put forward as a real alternative to IR systems as they provide the user with a fast and comprehensible answer to his or her information need. It has been 15 years since TREC introduced the first QA track. The principal campaigns in the evaluation of IR have been specific tracks focusing on the development and evaluation of this type of system. This study is a brief review of the TREC, CLEF and NTCIR Conferences from the QA perspective. We present a historical overview of 15 years of QA evaluation tracks using the method of systematic review. We have examined identified the different tasks or specific labs created in each QA track, the types of evaluation question used, as well as the evaluation measures used in the different competitions analyzed. Of the conferences, it is CLEF that has applied the greater variety of types of test question (factoid, definition, list, causal, yes/no, amongst others). NTCIR, held on 13 occasions, is the conference which has made use of a greater number of different evaluation measures. Accuracy, precision and recall have been the three most used evaluation measures in the three campaigns.

Paper 4: QUESTION-ANSWERING SYSTEMS AS EFFICIENT SOURCE OF TERMINOLOGICAL INFORMATION: EVALUATION

QAS stand as a new alternative for IR Systems. Most users often need to retrieve specific information about a factual question in order to obtain a whole document. We conducted a study to evaluate the efficiency of QAS as terminological sources for physicians, specialized translators, and users in general. To this end we analyzed the performance of one open-domain QA system, START, and one restricted-domain QAS, MedQA. The research entailed a collection of two hundred definitional questions (What is...?), either general or specialized, from WebMed. We studied the sources that QAS used to retrieve the answers, and later applied a range of evaluation measures to mark the quality of answers. Both QAS were determined to be appropriate for the retrieval of terminology, proving reliable as sources and supplying correct answers.

Paper 5: OPEN- VS. RESTRICTED-DOMAIN QA SYSTEMS IN THE BIOMEDICAL FIELD

QAS stand as a new alternative for information retrieval systems. We conducted a study to evaluate the efficiency of QAS as terminological sources for physicians, specialized translators and users in general. To this end we analysed the performance of two open-domain and two restricted-domain QAS. The research entailed a collection of 150 definitional questions from WebMed. We studied the sources that QAS used to retrieve the answers, and later applied a range of evaluation measures to mark the quality of answers. Through analysing the results obtained by asking the 150 questions in the QAS MedQA, START, QuALiM and HONqa, it was possible to evaluate the systems' operation through applying specific metrics (MRR, FHS, TRR, Precision, Recall). Despite the limitations demonstrated by these systems, it has been confirmed that these four QA systems are valid and useful for obtaining definitional medical information in that they offer coherent and precise answers.

Paper 6: EVALUACIÓN DEL RENDIMIENTO DE LOS SISTEMAS DE BÚSQUEDA DE RESPUESTAS DE DOMINIO GENERAL

QAS constitute an alternative to the conventional systems of IR, in an attempt to offer precise and understandable answers to factual questions instead of presenting the user with a list of related documents for a search. The effectiveness of four QAS has been evaluated for four QA available on the Web (QuaLiM, SEMOTE, START, and TrueKnowledge), by a broad sample of questions concerning definitions, facts, and lists, pertaining to different thematic domains. A collection of 500 questions was used with answers that were assessed by the users, and afterwards several measures were applied for the evaluation (MRR, TRR, FHS, MAP, and precision). It was found that START and TrueKnowledge presented acceptable levels of answers that were correct, precise, and in a well-ordered sequence. The results revealed the potential of these kinds of tools in the sphere of information access and recovery in the general domain.

Paper 7: MULTILINGUAL QUESTION-ANSWERING SYSTEM IN BIOMEDICAL DOMAIN ON THE WEB: AN EVALUATION

QAS are presented as an alternative to traditional systems of information retrieval, intended to offer precise responses to factual questions. This study assesses the effectiveness of these systems with terminological sources for specialists, as well as general users, in the context of multilingual searches within a restricted domain. With this aim, an analysis has been made of the results offered by the QA multilingual biomedical system HONqa, available in the Web. The study has used a set of 120 biomedical definitional questions (What is...?), taken from the medical

website WebMD, which were formulated in English, French, and Italian. The answers have been analysed using a series of specific measures (MRR, TRR, FHS, precision, recall, MAP), as well as sources used to get the responses.

The study confirms that for all the languages analysed the functioning effectiveness needs to be improved, although in the multilingual context analysed the questions in the English language achieve better results for retrieving definitional information than in French and Italian. This may be because the information sources from which the answers are elicited are more numerous in English and have a more appropriate structure for this purpose than in the other languages studied.

Paper 8: SATISFACCIÓN DE USUARIOS DEL ÁMBITO DE LA TRADUCCIÓN EN EL USO DE SISTEMAS DE BÚSQUEDA MULTILINGÜE DE RESPUESTAS COMO RECURSO DE INFORMACIÓN TERMINOLÓGICA

With the rapid growth of Internet and the development of new technologies in recent years, QAS have become an alternative to the traditional systems of information recovery. Although many studies have examined these systems, few have evaluated the user. The present work centres exclusively on the evolution of multilingual and cross-language QA, as it permits access to terminological information that is not in the user's language, and on the evaluation focusing user needs in order to identify the dimensions and factors in the development of an information system with the aim of improving its acceptance.

The goal is to determine the degree of user satisfaction for the multilingual QAS HONqa (English, French, and Italian), using the tool of the researchers ONG et al. (2009) based on an examination of models and theories. For this, the degree of satisfaction of translation professionals as a resource of terminological information was analysed.

A total of 122 specialists in English, French, or Italian has used and assessed the terminological resource from the standpoint of translation.

After the analysis of the results found with respect to the QAS HONqa, which have enabled an evaluation of the satisfaction of Translation students, it was confirmed that the system proved easy and useful for users/students for the recovery of terminological information in all three languages.

Presentación (normativa, estancias y publicaciones)

Resumen

Abstract

Índice

Lista de siglas y abreviaturas

Lista de figuras

Lista de tablas

1. INTRODUCCIÓN

1.1. Antecedentes

1.1.1. Recuperación de Información

1.1.1.1. Recuperación de información monolingüe

1.1.1.2. Recuperación de información multilingüe

1.1.1.2.1. SRI monolingües ensamblados

1.1.1.2.2. Recuperación de información translingües

1.1.2. Sistemas de búsqueda de respuestas

1.1.2.1. Evolución de los SBR

1.1.2.2. Arquitectura de los SBR

1.1.2.3. Análisis de las preguntas

1.1.2.4. Selección de pasajes y documentos relevantes

1.1.2.5. SBR multilingües

1.1.2.6. SBR en la Web

1.1.3. La Traducción en CLIR

1.1.3.1. Recursos utilizados para la traducción en los SBR multilingües

1.1.3.1.1. Recursos lingüísticos

1.1.3.1.1.1. Bases de datos

- 1.1.3.1.1.2. Corpus
 - 1.1.3.1.1.2.1. Corpus clasificados según idioma
 - 1.1.3.1.1.2.2. Corpus clasificados según estructura
 - 1.1.3.1.1.3. Diccionarios
 - 1.1.3.1.1.4. Enciclopedias
 - 1.1.3.1.1.4.1. Wikipedia y Wiktionary
 - 1.1.3.1.1.5. EuroWordNet
 - 1.1.3.1.1.6. Ontologías
 - 1.1.3.1.1.7. Páginas web
 - 1.1.3.1.1.8. Tesoros
- 1.1.3.1.2. Herramientas lingüísticas
 - 1.1.3.1.2.1. Gramáticas computacionales
 - 1.1.3.1.2.2. Traductores automáticos
- 1.2. Justificación
 - 1.2.1. Motivación, hipótesis iniciales y objetivos
 - 1.2.2. Metodología
 - 1.2.2.1. Análisis orientado a los recursos y herramientas lingüísticos en los SBR multilingües
 - 1.2.2.2. Evaluación orientada al sistema
 - 1.2.2.3. Evaluación orientada al usuario: satisfacción de usuarios en SBR multilingües

II. ESTUDIOS REALIZADOS

- 2.1. Language resources used in multi-lingual Question Answering Systems (Publicación 1)
 - 2.1.1. Introduction
 - 2.1.2. State of Art about Question Answering Systems
 - 2.1.3. Method Section
 - 2.1.4. Results and Discussion
 - 2.1.5. Conclusions
- 2.2. Language Resources for Translation in Multi-lingual Question Answering Systems (Publicación 2)
 - 2.2.1. Introduction
 - 2.2.2. Information Retrieval
 - 2.2.3. Question Answering Systems
 - 2.2.4. Multi-lingual QA Systems
 - 2.2.5. Method section
 - 2.2.6. Results and Discussion
 - 2.2.7. Conclusion

- 2.3. Question Answering Track Evaluation in TREC, CLEF and NTCIR (Publicación 3)
 - 2.3.1. Introduction
 - 2.3.2. Question Answering in Information Retrieval Evaluation Campaigns
 - 2.3.2.1. Text REtrieval Conference (TREC)
 - 2.3.2.2. Conference and Labs Evaluation Forum (CLEF)
 - 2.3.2.3. NTCIR Conference
 - 2.3.3. Test Questions
 - 2.3.4. Answer Assessment
 - 2.3.5. Evaluation measures
 - 2.3.6. Conclusions

- 2.4. Question-Answering Systems as Efficient Source of Terminological Information: Evaluation (Publicación 4)
 - 2.4.1. Introduction
 - 2.4.2. Background
 - 2.4.3. Methodology
 - 2.4.4. Results
 - 2.4.5. Discussion
 - 2.4.6. Conclusions

- 2.5. Open- vs. Restricted- Domain Question Answering Systems in the Biomedical Field (Publicación 5)
 - 2.5.1. Introduction
 - 2.5.2. Open- vs. Restricted- Domain QA Systems on the Web
 - 2.5.3. Methodology
 - 2.5.4. Results and discussion
 - 2.5.5. Conclusions

- 2.6. Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general (Publicación 6)
 - 2.6.1. Introducción
 - 2.6.2. Sistemas de búsqueda de respuestas
 - 2.6.3. Método y materiales
 - 2.6.4. Resultados y discusión
 - 2.6.5. Conclusiones

- 2.7. Multilingual Question-Answering System in biomedical domain on the Web: an evaluation (Publicación 7)
 - 2.7.1. Introduction
 - 2.7.2. Question-Answering System: Beyond Information Retrieval
 - 2.7.3. Method section
 - 2.7.4. Results section
 - 2.7.4.1. Evaluation measures
 - 2.7.4.2. Information sources

2.7.5. Conclusions

2.8. Satisfacción de usuarios del ámbito de la traducción en el uso de sistemas de búsqueda multilingüe de respuestas como recurso de información terminológica (Publicación 8)

2.8.1. Introducción

2.8.2. Sistemas de búsqueda de respuestas: Más allá de la Recuperación de Información

2.8.3. Interactividad de los Sistemas de búsqueda de respuestas

2.8.3.1. Evaluación centrada en el usuario

2.8.4. Metodología

2.8.5. Resultados y discusión

2.8.5.1. Datos demográficos

2.8.5.2. Autopercepción sobre experiencia y conocimientos

2.8.5.3. Satisfacción de los usuarios respecto al sistema

2.8.6. Conclusiones

III. CONCLUSIONES

3.1. Conclusiones generales

3.2. Conclusiones versus hipótesis

3.3. Conclusiones particulares

3.3.1. Sobre la identificación, el análisis y la evaluación de los recursos y herramientas lingüísticos en los SBR (Objetivos 1 y 2)

3.3.2. Sobre la evaluación de los SBR (Objetivo 3)

3.3.3. Sobre la evaluación de la satisfacción de usuarios (Objetivo 4)

3.4. Consideraciones finales y líneas de investigación futuras

Referencias

Anexos

Glosario

O LISTA DE SIGLAS Y ABREVIATURAS

A continuación se detallan las siglas y abreviaturas citadas en el presente trabajo. Se especifican aquellas que presentan una forma diferente en inglés y español.

ACM		<i>Association for Computing Machinery</i>
AVEP		
Español	AVEP	Precisión promedio
English	AVEP	<i>Average Precision</i>
CLEF		<i>Cross-Language Evaluation Forum</i>
CLIA		
Español	CLIA	Acceso a la Información translingüe
English	CLIA	<i>Cross-language Information Access</i> <i>Cross-lingual Information Access</i>
CLIR		
Español	CLIR	Recuperación de Información translingüe
English	CLIR	<i>Cross-language information retrieval</i> <i>Cross-lingual information retrieval</i>
ELRA		<i>European Language Resources Association</i>
EWN		EuroWordNet
FHS		<i>First Hit Success</i>
HONQA		<i>Health On the Net foundation Question-Answering</i>
AI		Acceso a la Información
LREC		<i>Language Resources and Evaluation Conference</i>

MAP		<i>Mean Average Precision</i>
MEDQA		<i>Medical Question Answering system</i>
MLIA		
Español	MLIA	Acceso a la información multilingüe
English	MLIA	<i>MultiLingual Information Access</i>
MLIR		
Español	MLIR	Recuperación de Información multilingüe
English	MLIR	<i>MultiLingual Information Retrieval</i>
MRR		<i>Mean Reciprocal Rank</i>
NSIR		<i>NSIR Question Answering System</i>
NTCIR		<i>the NII Text Collection for IR systems</i>
SRI		
Español	SRI	Sistemas de Recuperación de Información
English	IRS	<i>Information Retrieval systems</i>
PLN		
Español	PLN	Procesamiento del lenguaje natural
English	NLP	<i>Natural Language Processing</i>
RI		
Español	RI	Recuperación de Información
English	IR	<i>Information Retrieval</i>
SBR		
Español	SBR	Sistema de búsqueda de respuestas
English	QAS	<i>Question-answering system</i>
	QA	<i>Question-Answering</i>
SERA		Sociedad Española de Recuperación de Información
START		<i>SynTactic Analysis using Reversible Transformations</i> <i>START Natural Language Question Answering System</i>

TA		
Español	TA	Traducción Automática
English	MT	<i>Machine Translation</i>
TREC		<i>Text REtrieval Conference</i>
TRR		<i>Total Reciprocal Rank</i>

0

LISTA DE FIGURAS

- FIGURA 1** Arquitectura de un SRI (Vallez y Pedraza-Jiménez, 2007)
- FIGURA 2** Proceso de Recuperación de Información (Baeza-Yates y Ribeiro-Neto, 2011)
- FIGURA 3** Taxonomía de los Modelos de RI según Baeza-Yates y Ribeiro-Neto (2011)
- FIGURA 4** Taxonomía de los sistemas MLIR
- FIGURA 5** Ejemplo de un sistema MLIR monolingüe ensamblado
- FIGURA 6** Ejemplo de un sistema CLIR (Aceves, 2008)
- FIGURA 7** Ejemplo de otro sistema CLIR (Magnini, 2002)
- FIGURA 8** Ejemplo de un sistema de PLN
- FIGURA 9** General architecture of a Textual QA system (Monz, 2003)
- FIGURA 10** Proceso que tiene lugar en un SBR
- FIGURA 11** Esquema básico de un SBR translingüe que incorpora traducción automática (Olvera-Lobo y García-Santiago, 2010)
- FIGURA 12** Recursos utilizados por los SBR multilingües para solucionar el problema de la traducción (Diekema, 2003)
- FIGURA 13** Arquitectura de EuroWordNet (Díez Orzas, 1999)
- FIGURA 14** Representación de una ontología conceptual (Gatius, 2000)
- FIGURA 15** Ejemplo de un SBR cuya fuente documental se documenta en páginas web (Ramamonjisoa, 2003)
- FIGURA 16** Objetivos de TAUS
- FIGURA 17** Idiomas utilizados en los SBR multilingües de las publicaciones analizadas

- FIGURA 18** An example of a basic architecture of a monolingual QA system (Cui *et al.*, 2004)
- FIGURA 19** NSIR answer search interface
- FIGURA 20** Answers provided by MedQA to the query: What is narcolepsy?
- FIGURA 21** An example of a basic architecture of a multi-lingual QA system (Aceves Pérez, 2008)
- FIGURA 22** Groups participating in each of the subject areas of CLEF
- FIGURA 23** Papers analysed by Conference
- FIGURA 24** Papers analysed by year
- FIGURA 25** Resources used by multi-lingual QA systems for translation (Diekema, 2003)
- FIGURA 26** Resources used in the papers reviewed
- FIGURA 27** Use of automatic translators per year
- FIGURA 28** Use of corpora per year
- FIGURA 29** Use of dictionaries per year
- FIGURA 30** Use of ontologies per year
- FIGURA 31** Type of translation used by the multi-lingual QA systems
- FIGURA 32** Resources and tools per year
- FIGURA 33** Papers analysed by Conference
- FIGURA 34** Papers analysed by year
- FIGURA 35** Resources used in the papers reviewed
- FIGURA 36** Frequency each type of test questions is been used.
- FIGURA 37** The intermediating 'window' of an inexact answer
- FIGURA 38** Analysis of frequencies according to the response time in MedQA
- FIGURA 39** Category of question reference
- FIGURA 40** Incorrect, inexact and correct answer
- FIGURA 41** Arquitectura general de un SBR
- FIGURA 42** Interfaz de QuALiM: Página de resultados

- FIGURA 43** Interfaz de SEMOTE: Página principal
- FIGURA 44** Interfaz de START: Página principal
- FIGURA 45** Interfaz de TrueKnowledge: Página principal
- FIGURA 46** Correct, inexact and incorrect answers
- FIGURA 47** Recall-precision relationship
- FIGURA 48** Página de resultados de sistema de búsqueda de respuestas HONqa
- FIGURA 49** Histograma del grado de conocimiento de los estudiantes
- FIGURA 50** Interfaz de Google al buscar información sobre Virginia Woolf

0

LISTA DE TABLAS

TABLA 1	Taxonomía de los modelos en RI
TABLA 2	Lista de respuesta para una hipotética pregunta de lista
TABLA 3	Análisis de las palabras que forman una pregunta en un SBR
TABLA 4	Lista de SBR disponibles en la Web
TABLA 5	Tabla resumen de los contenidos de las publicaciones realizadas
TABLA 6	Colecciones utilizadas en los test realizados
TABLA 7	Origen de la colección de preguntas de evaluación
TABLA 8	Preguntas extraídas por años
TABLA 9	Preguntas según la especialidad temática
TABLA 10	Publicaciones sobre SBR en Web of Science
TABLA 11	SBR utilizados en la investigación realizada
TABLA 12	Características de cada test para la evaluación orientada al sistema
TABLA 13	List of the most important multi-lingual QA systems
TABLA 14	Tasks in QA Track in TREC
TABLA 15	Tasks in CLEF QA Track
TABLA 16	Tasks in NTCIR
TABLA 17	Average of test questions per conference
TABLA 18	Type of test questions per conference
TABLA 19	Measures used by the conferences
TABLA 20	Frequency each evaluation measure used
TABLA 21	Categories of reference of definitional questions

TABLA 22	Sources used by START
TABLA 23	Sources used by MedQA
TABLA 24	Answers provided by MedQA
TABLA 25	Measures for evaluating the quality of answers
TABLA 26	Evaluation measures for QA systems
TABLA 27	Precision considering correct and inexact answers
TABLA 28	Correlation of measures
TABLA 29	Procedencia de las preguntas de la muestra
TABLA 30	Preguntas por año
TABLA 31	Temas a los que se refieren las preguntas de la muestra
TABLA 32	Preguntas según el tipo de respuesta esperada
TABLA 33	Respuestas recuperadas en los cuatro SBR
TABLA 34	Medidas de evaluación
TABLA 35	Correlación entre medidas
TABLA 36	Answers retrieved by HONqa in the three languages
TABLA 37	Evaluation measures
TABLA 38	Sources used in English
TABLA 39	Sources used in French
TABLA 40	Sources used in Italian
TABLA 41	Herramienta de evaluación para medir la satisfacción de los usuarios
TABLA 42	Tabla de ítems del cuestionario agrupada en cuatro factores
TABLA 43	Estadísticas de la edad
TABLA 44	Estadísticas de los alumnos según curso

PARTE I. INTRODUCCIÓN

CAPÍTULO 1

1.1. ANTECEDENTES

El desarrollo de la Web semántica requiere un gran esfuerzo económico y humano. Consecuentemente, es más útil crear mecanismos y herramientas que faciliten tal expansión. Desde el punto de vista de la Recuperación de Información (en adelante RI), el uso del lenguaje natural favorece el acceso a los contenidos de la Web semántica, así como simplifica y agiliza a los usuarios la gestión de los contenidos en su forma habitual de expresión. Aunque la necesidad del hombre de superar las barreras lingüísticas para poder comunicarse entre sí es casi tan antigua como la historia de la humanidad, la creciente popularidad de Internet y la amplia disponibilidad de recursos informativos en red para un público cada vez más numeroso es un fenómeno relativamente reciente. La *World Wide Web*, junto con la creciente globalización de las empresas y de las organizaciones, y la existencia de cada vez más usuarios no angloparlantes, conlleva una fuerte demanda de herramientas que permitan a estos usuarios obtener información de las más variadas fuentes. Las últimas estadísticas del sitio web *Internet World Stats*¹ demuestran que aunque el inglés sigue siendo el lenguaje predominante en Internet, hay un crecimiento progresivo en la incorporación de las lenguas neolatinas en la red, de modo que tenemos un aumento no sólo en la cantidad de información disponible sino también en las lenguas usadas en esas fuentes de documentación. Sin embargo, aún investigadores y desarrolladores subestiman las restricciones lingüísticas ya subrayadas. Una característica clave para evaluar en términos de eficiencia a los sistemas de RI (en adelante SRI) es la capacidad que poseen para permitir a los usuarios encontrar un corpus de documentos en diferentes lenguas, y facilitar la recuperación de información relevante a pesar de las limitaciones en la competencia lingüística.

Desde los años cuarenta, el problema del almacenamiento y recuperación de información ha propiciado un aumento de la atención (Rijsbergen, 1979). Hasta entonces los investigadores de

¹ <http://www.internetworldstats.com/stats7.htm> [Consultado 08/01/2015]

diferentes disciplinas han ayudado a desarrollar los métodos más eficientes y sofisticados para procesar, tratar y recuperar la información que satisface las necesidades de los usuarios.

En demasiadas ocasiones, al plantear una determinada consulta en las herramientas de búsqueda de información web (buscadores, directorios o metabuscadores) el número de páginas web recuperadas resulta excesivo y no todas ellas son relevantes ni útiles para los objetivos del usuario. Por ello, en diversos ámbitos se ha comenzado a reconocer la utilidad de otros tipos de sistemas, como los sistemas de búsqueda de respuestas (en adelante SBR), para recuperar información especializada de forma rápida y efectiva (Crouch *et al.*, 2005; Lee *et al.*, 2006; Yu *et al.*, 2007).

Los SBR se presentan como una alternativa a los tradicionales SRI tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de mostrar al usuario una lista de documentos relacionados con su búsqueda (Jackson y Schilder, 2005). El beneficio es que los usuarios no tienen que leer el documento completo para encontrar la información deseada. Los SBR pretenden emular el comportamiento del lenguaje humano, por ello permiten responder preguntas del lenguaje natural. Como resultado, la máquina intenta entender la respuesta y responder adecuadamente. En otras palabras, la comprensión del lenguaje natural por la máquina es un proceso esencial en el desarrollo de los SBR (Belkin y Vickery, 1985; Sultan, 2006). En el campo de la RI, se vienen desarrollando diferentes tipos de herramientas monolingües y multilingües para facilitar a los especialistas su trabajo, así como para ayudar a otros usuarios a encontrar una amplia variedad de información.

Este capítulo trata de proporcionar un análisis general del objeto de estudio y ofrece una visión amplia de los SBR monolingües, multilingües y translingües. Además, se han identificado y analizado las principales características y arquitectura de los diferentes tipos de SBR.

1.1.1. RECUPERACIÓN DE INFORMACIÓN

La RI es una disciplina centrada en los problemas de selección de información desde un sistema de almacenamiento para facilitar la recuperación de las necesidades de los usuarios (Baeza-Yates y Ribeiro-Neto, 1999, 2011; Korfhage, 1997; Salton, 1989; Salton y McGill, 1983; Rijsbergen, 1979). Tradicionalmente, la RI se ha entendido como el proceso, totalmente automático, en el que dada una consulta (que, supuestamente, expresa la necesidad de información del usuario) y una colección de documentos, el sistema devuelve una lista ordenada de documentos potencialmente relevantes para esa consulta (Meadow, 1992).

Sin embargo, esta sencillez sólo es aparente ya que muchos de los conceptos relativos a la RI son difíciles de definir, tanto operativa como conceptualmente. Es decir, cuando hablamos de RI nos referimos a un concepto que suele ser definido en un sentido muy amplio (Rijsbergen, 1999) y desde distintas perspectivas. Se trata de un proceso complejo durante el que se realizan actividades que incluyen, entre otras: el reconocimiento de la estrategia de búsqueda planteada, la aplicación de diferentes algoritmos de recuperación y de ordenación de los resultados según su relevancia o utilidad probable para el usuario, o la selección de *ítems* de información (documentos o su representación) que el sistema intenta localizar y recuperar tan veloz y económicamente como sea posible.

Un SRI es un sistema usado para almacenar *ítems* de información que necesitan ser procesados, buscados, recuperados, y proporcionado a los usuarios (Salton y Mc Gill, 1983). Efectivamente, pueden verse como una “caja negra” que acepta *inputs* y produce *outputs* (Harter y Hert, 1997). Un SRI con funcionamiento óptimo recuperaría todos los documentos concurrentes (lo que implica una cobertura completa) y sólo aquellos documentos que son relevantes (precisión perfecta) (Baeza-Yates y Ribeiro-Neto, 2011). Por tanto, puede afirmarse que el valor de estos sistemas dependen de su capacidad para identificar rápida y correctamente la información útil, de su facilidad para rechazar los *ítems* extraños o irrelevantes y de la versatilidad de los métodos que emplea (Salton y Mc Gill, 1983).

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

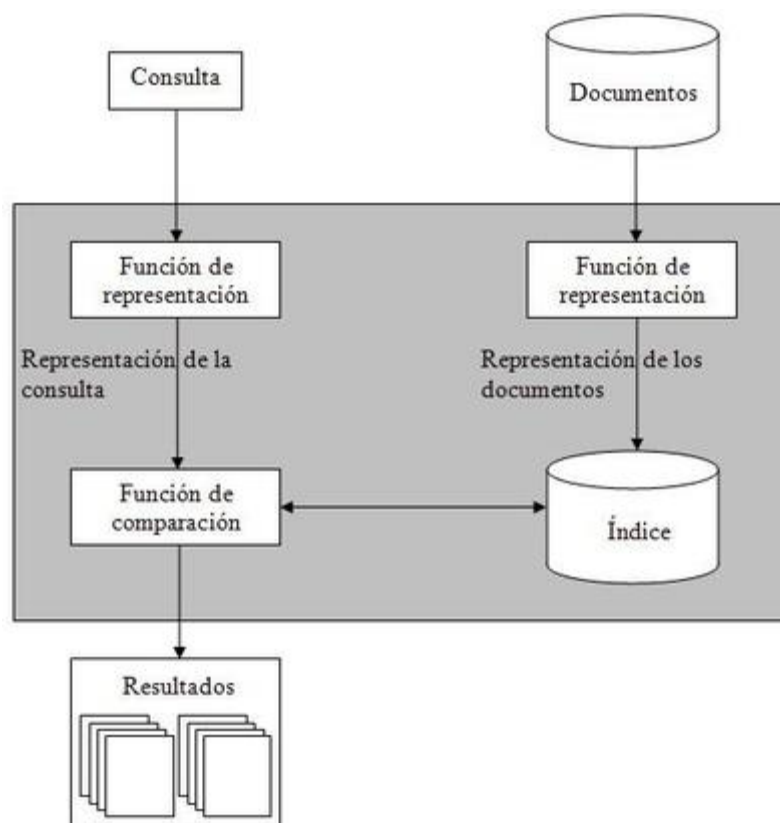


Figura 1. Arquitectura de un SRI (Vallez y Pedraza-Jiménez, 2007)

Cuando una necesidad de información aparece, se desencadena un proceso llamado “estrategia de búsqueda”, que permite proporcionar documentos por el sistema (Belkin y Croft, 1987). El proceso debe entablar: *a)* definición de la necesidad de información; *b)* selección de las fuentes de información utilizadas; *c)* traducción de la consulta expresada por el usuario en lenguaje natural a un lenguaje indexado de esa fuente de información, si fuera necesario; *d)* traducción de dicha expresión desde un lenguaje indexado al lenguaje de interrogación; *e)* planteamiento de las expresiones desde el lenguaje de la pregunta; *f)* evaluación de los resultados por el usuario y redefinición de la consulta, en caso necesario; y *g)* selección y recuperación de documentos que respondan a las necesidades del usuario (Chowdhury, 1999).

Un SRI está compuesto por tres componentes principales: la base de datos documental, el subsistema de consultas y el subsistema de *matching* o evaluación. La base de datos documental almacena los documentos representados por los términos índices extraídos de los mismos mediante un mecanismo de indización, habitualmente automático. El subsistema de consultas está compuesto por un lenguaje de consultas que permite al usuario formular sus necesidades de información a través de una interfaz. El subsistema de evaluación evalúa el grado en el que los documentos satisfacen los requisitos expresados en las consultas, tras lo cual asigna a cada

documento un valor de relevancia y recupera los documentos encontrados más relevantes. Los subsistemas de consultas y evaluación están íntimamente ligados, es decir, la definición del segundo depende del tipo de consultas que soporte el primero.

En los últimos años, los SRI han evolucionado persiguiendo una mayor afinidad con los usuarios, es decir, intentando adecuar los resultados a sus necesidades de información. La perspectiva de la RI orientada a usuarios se centra principalmente en la representación de los documentos y de los problemas o necesidades de información, en el comportamiento en las búsquedas y en los componentes humanos de los sistemas en situaciones reales. Se nutre principalmente de la psicología cognitiva y emplea métodos de las ciencias sociales (Olvera-Lobo, 1999). En este sentido una de las primeras y más destacadas aportaciones es la de Belkin (1980) –y su ASK o *Anomalous State of Knowledge*– que afirma que el proceso de RI se inicia cuando un usuario reconoce algo anómalo en su estado de conocimiento y desea resolver esa incoherencia. El marco conceptual en el que se elabora es explícitamente cognitivo y estructural. Además, el éxito del proceso de RI –y, por tanto, del funcionamiento del SRI– también dependerá del usuario y de lo que Blair (1990) denomina “umbral de futilidad” o punto en el que el usuario cesa de examinar la lista de documentos recuperados y determina su satisfacción respecto a la necesidad de información suscitada.

Este modelo tradicional lleva consigo muchas restricciones implícitas tales como: a) la suposición de que los usuarios del sistema buscan documentos (textos completos), no respuestas, y que son los documentos, como tales, los que responden y satisfacen una consulta; b) que el proceso debe ser directo y unidireccional en lugar de interactivo; c) y por último, que la consulta y el documento están escritos en la misma lengua (Salton y Mc Gill, 1983). Además, hay que tener en cuenta que la *World Wide Web* es el ámbito donde los diversos modelos de SRI encuentran su máxima aplicación favoreciéndose de esta manera el aumento de usuarios de los mismos (Baeza-Yates y Ribeiro-Neto, 2011). Ante este contexto general surgen como alternativa los SBR que tratamos en esta tesis doctoral.

1.1.1.1. RECUPERACIÓN DE INFORMACIÓN MONOLINGÜE

En los últimos años los SRI han evolucionado en la línea de mejorar los criterios de afinidad con los usuarios para presentarles sus respuestas adecuándose a sus necesidades de información (Martínez Méndez, 2002).

Como hemos mencionado, un momento importante para los SRI fue el desarrollo y creación de la Web, pues se considera el punto clave para su consolidación gracias al vertiginoso abaratamiento de la tecnología informática, al desarrollo cada vez mayor de las telecomunicaciones y a la facilidad de publicación de cualquier documento o material por cualquier autor, sin necesidad de pasar por el filtro de editores o revisores. Como es comúnmente sabido en Internet “todo vale” (Martínez Méndez, 2002).

Para entender los SRI monolingües hay que mencionar los sistemas más comunes y por tanto, los que han suscitado más interés entre los investigadores (Baeza-Yates y Ribeiro-Neto, 2011; Villena Román, 1997; Grossman y Frieder, 1998).

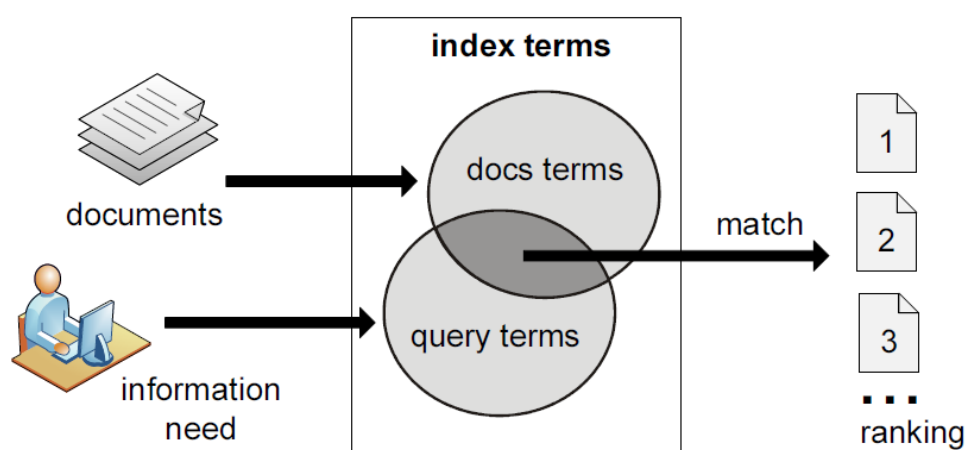


Figura 1. Proceso de Recuperación de Información (Baeza-Yates y Ribeiro-Neto, 2011)

Tradicionalmente, los sistemas más utilizados eran aquellos basados en los denominados modelos tradicionales de RI (modelo booleano, modelo espacio vectorial y modelo probabilístico). Durante décadas el más conocido y usado ha sido el modelo booleano basado en la lógica clásica (Salton, 1989; Salton y McGill, 1983; Rijsbergen, 1979), el cual presenta una serie de limitaciones importantes. Para mejorarlo, se propuso modelos más avanzados como el vectorial –basado en espacios vectoriales– cuyo primer ejemplo más representativo fue el sistema SMART (Salton, 1971; Salton y McGill, 1983), el probabilístico –basado en la teoría de la probabilidad y la inferencia bayesiana– cuyo sistema más representativo fue el sistema INQUERY (Allan *et al.*, 1997; Callan *et al.*, 1992) y las distintas variantes de los modelos difusos –basados en lógica difusa– (Crestani y Pasi, 2000; Miyamoto, 1990). No obstante, progresivamente el desarrollo de nuevos modelos más versátiles ha cambiado el uso y tendencia de los modelos tradicionales.

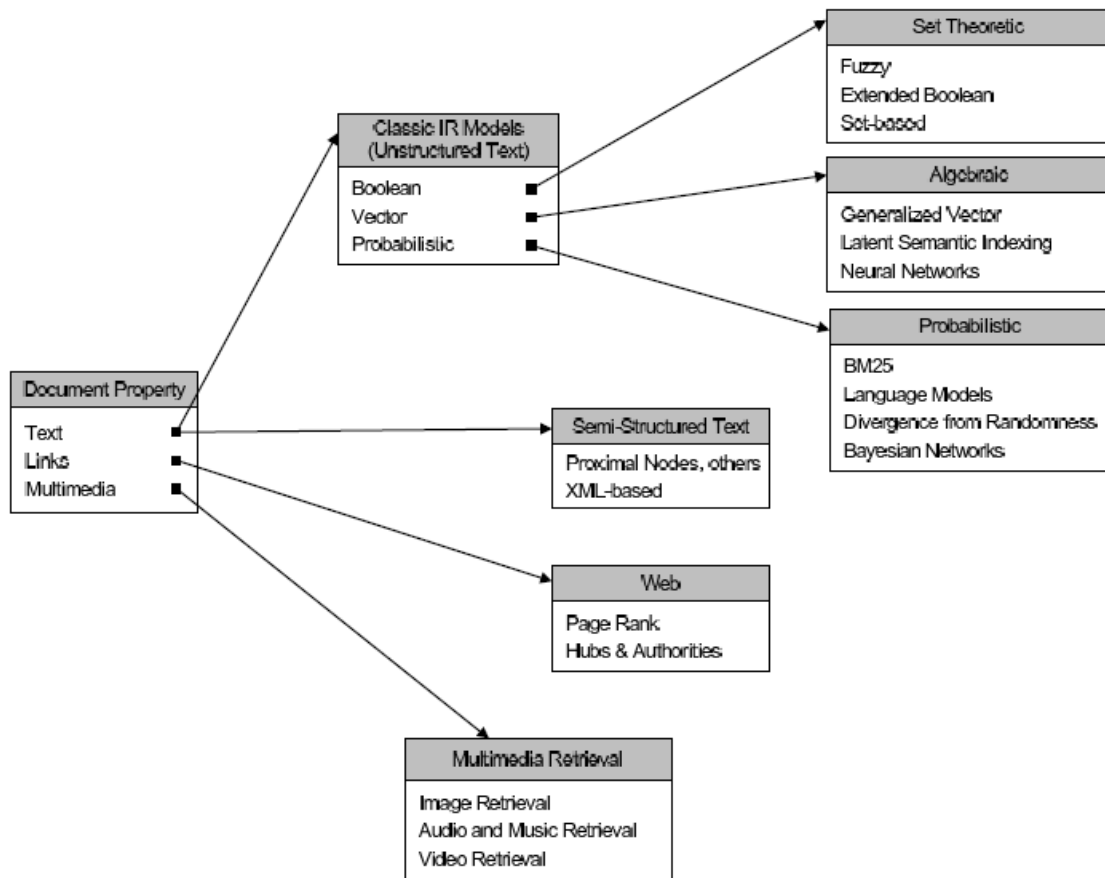


Figura 2. Taxonomía de los Modelos de RI según Baeza-Yates y Ribeiro-Neto (2011)

Una taxonomía bastante completa de los modelos de RI es la desarrollada por Baeza-Yates y Ribeiro-Neto (2011) en su obra *Modern Information Retrieval: the Concepts and Technology behind search (2nd Edition)*. Los autores amplían en esta segunda edición una primera taxonomía más actualizada que divide los modelos en cuatro grandes grupos: *classical models*, *alternative Set Theoretic Models*, *alternative Algebraic Models*, *alternative Probabilistic Models* y *other Models*. En la Tabla 1, aparece la taxonomía creada por estos autores con las principales características de cada modelo.

Modelos clásicos	
Modelo booleano	<ul style="list-style-type: none"> - Basado en la lógica clásica (Salton, 1989; Salton y McGill, 1983; Rijsbergen, 1979) y en la teoría y Algebra booleana (Baeza-Yates y Ribeiro-Neto, 2011) - Se requiere especificar expresiones booleanas para las consultas - El modelo booleano predice si cada documento es o no relevante. (Baeza-Yates y Ribeiro-Neto, 2011)

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

Modelo vectorial	<ul style="list-style-type: none"> - Basado en el espacio vectorial – que fue muy reconocido por su uso en SMART (<i>System for the Mechanical Analysis and Retrieval of Text</i>) (Salton, 1989; Salton and McGill, 1983). - Propone un marco en la que la correspondencia parcial es posible. - Los documentos son ordenados en orden decreciente según su grado de similaridad. (Baeza-Yates y Ribeiro-Neto, 2011) - Calculan la probabilidad de un documento de ser relevante a una pregunta dada (García Broncano, 2006).
Modelo probabilístico	<ul style="list-style-type: none"> - Basado en la teoría de probabilidad e inferencia bayesiana – cuyo sistema más representativo es INQUERY (Callan et al., 1992). - El modelo probabilístico supera el problema de RI usando un marco probabilístico. - El SRI usa esta información para redefinir la descripción del conjunto de respuestas ideales (Baeza-Yates y Ribeiro-Neto, 2011).
Modelos teóricos de conjuntos alternativos	
Modelo basado en conjuntos	<ul style="list-style-type: none"> - Enfoque reciente que combina la teoría de conjuntos con la clasificación vectorial. - La idea fundamental es usar las dependencias comunes en los términos indexados para mejorar los resultados. - El enfoque, que pretende mejorar los resultados con varias colecciones, constituye el primer modelo de RI que se aprovecha en términos de dependencia de forma efectiva de las colecciones generales (Baeza-Yates y Ribeiro-Neto, 2011).
Modelo booleano extendido	<ul style="list-style-type: none"> - En el modelo booleano no se generaba ninguna clasificación del conjunto de las respuestas. Esta estrategia permite combinar las características del modelo del espacio vectorial con las propiedades del algebra booleana. - El modelo es bastante potente, las propiedades son muy interesantes y pueden ser útiles, si bien la computación es, a menudo, compleja (Baeza-Yates y Ribeiro-Neto, 2011).
Modelo de conjunto difuso	<ul style="list-style-type: none"> - Modelo propuesto por Ogawa, Morita y Kobayashi (1991). - La teoría de conjuntos difusos trata con la representación de clases cuyas fronteras no están bien definidas. - Las tres operaciones más comúnmente usadas en los conjuntos difusos son: el complemento de un conjunto difuso, la unión de dos o más conjuntos y la intersección de dos o más conjuntos difusos.
Modelos algebraicos alternativos	
Modelo vectorial	<ul style="list-style-type: none"> - Este modelo propone un marco interesante que personifica la noción de dependencias de los términos.

generalizado	<ul style="list-style-type: none"> - El modelo adopta la idea de que cuando concurren varios términos, se induce dependencias entre ellos (Baeza-Yates y Ribeiro-Neto, 2011).
Modelo de indexación semántica latente	<ul style="list-style-type: none"> - Mapeo de documentos y preguntas en un espacio multidimensional compuesto de conceptos. - El indexado semántico latente permite una conceptualización interesante del problema de la RI. - Desde un punto de vista práctico, este modelo no ha obtenido resultados alentadores (Baeza-Yates y Ribeiro-Neto, 2011).
Modelo de redes neuronales	<ul style="list-style-type: none"> - Representación de las interconexiones neuronales en el cerebro humano. - Este modelo todavía no ha sido testado extensivamente y no son claras las mejoras que ofrece (Baeza-Yates y Ribeiro-Neto, 2011).
Modelos probabilísticos alternativos	
BM25	<ul style="list-style-type: none"> - Creado como resultado de una serie de experimentos en variaciones del modelo probabilístico. - Basado en el principio de frecuencia de los documentos inversos. - Usado como base para evaluar las nuevas funciones de clasificación, en sustitución del modelo vectorial clásico (Baeza-Yates y Ribeiro-Neto, 2011).
Modelos de lenguas	<ul style="list-style-type: none"> - Se compone de lo siguiente: <ul style="list-style-type: none"> + Un conjunto de modelos con documentos en diversas lenguas, una por documento de la colección. + Una función de distribución probabilística que permite estimar la probabilidad de documentos generados para cada término de la pregunta. + Una función de clasificación que combina estas probabilidades generadas para cada término en una lista de documentos relacionados con la pregunta.
Divergencia desde la aleatoriedad	<ul style="list-style-type: none"> - Propuesto por Amati y Rijsbergen (2002). - La idea es calcular la importancia de cada término midiendo la divergencia entre la posición del término producida por un proceso aleatorio y la posición actual del término en cuestión.
Modelos de redes bayesianas	<ul style="list-style-type: none"> - Propuesto por Turtle y Croft (1991) - Usando un modelo previo de conocimiento sobre la semántica de aplicación.

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

Modelo de redes de inferencia	<ul style="list-style-type: none"> - Propuesto por Ribeiro-Neto y Muntz (1996) - Variables aleatorias asociadas con documentos, términos indexados y preguntas. - Solamente se tienen en cuenta como nodos activos los estados que tiene un solo documento.
Modelos de redes de confianza	- Una variante del modelo de redes de inferencias con una ligera variación en la tipología de redes.
Otros modelos	
Modelo hipertextual	<ul style="list-style-type: none"> - La solución es definir una nueva estructura organizativa a partir de una ya existente. - Un hipertexto se compone básicamente de nodos que se encuentran correlacionados con enlaces directos en una estructura gráfica.
Modelos basados en la Web	<ul style="list-style-type: none"> - Las colecciones se componen de páginas web, en lugar de documentos. - Las páginas deben haber sido recopiladas. - Las colecciones deben ser amplias.
Recuperación textual estructurada	Se trata de estructuras textuales de documentos para mejorar la recuperación.
Recuperación multimedia	Las estrategias de recuperación que se aplican son bastante distintas a las estrategias de recuperación textual.
La búsqueda empresarial y vertical	<ul style="list-style-type: none"> - La búsqueda empresarial es la tarea de buscar información de interés en colecciones de documentos corporativos. - En la búsqueda empresarial tiene una gran importancia aquella información que no está presente en la Web, como privacidad, propiedades, permisos. - Las colecciones verticales presentan cambios específicos en las búsquedas y recuperación.

Tabla 1. Taxonomía de los modelos RI

1.1.1.2. RECUPERACIÓN DE INFORMACIÓN MULTILINGÜE

La RI debe adaptarse a las necesidades de búsqueda de los usuarios, una audiencia cada vez más heterogénea. Hoy en día nos encontramos países con varias lenguas cooficiales, usuarios que pueden escribir una consulta en una lengua y entender perfectamente la respuesta dada en otra, diferentes lenguajes naturales en la Web, un creciente auge de fuentes de información en diversos idiomas, así como otras situaciones mencionadas anteriormente. Este panorama ha puesto en relieve la necesidad de investigar y desarrollar SRI multilingües (en adelante MLIR, *Multi-Language or MultiLingual Information Retrieval*).

En la literatura disponible existe cierta confusión con el concepto de MLIR y es posible encontrar definiciones contradictorias. Los sistemas MLIR son una extensión de los SRI tradicionales que están capacitados para operar sobre una colección de documentos multilingües.

Al igual que en cualquier SRI monolingüe, los sistemas MLIR presentan tres dimensiones que, aunque no son exactamente independientes, nos permiten entender las características y diferencias de los distintos modelos y sistemas. La primera dimensión en MLIR es conocida como MLIA (Acceso a la información multilingüe, *Multi-Language Information Access*), que hace referencia a todos los temas de acceso, búsqueda y recuperación de cualquier tipo de información en el idioma que se haya expresado la consulta, así como en los idiomas que se encuentre disponible la colección del SRI. La segunda dimensión se refiere al tipo de información que ofrecen los SRI, o sea, tanto al nivel de especialización de los documentos de los SRI como a la tipología de documentos que localiza y rastrea ese sistema. Los actuales foros centrados en evaluación dividen sus *tracks* o sesiones basándose en esta dimensión. Por último, la tercera dimensión importante en MLIR guarda relación con la colección de documentos que pueden ser buscados.

MLIR es un concepto muy genérico que abarca un amplio número de sistemas. Estos sistemas presentan dos tipos de arquitectura, una que no necesita la traducción y otra que siempre implica la traducción de la consulta, palabras claves, documentos y/o respuestas (véase figura 4).

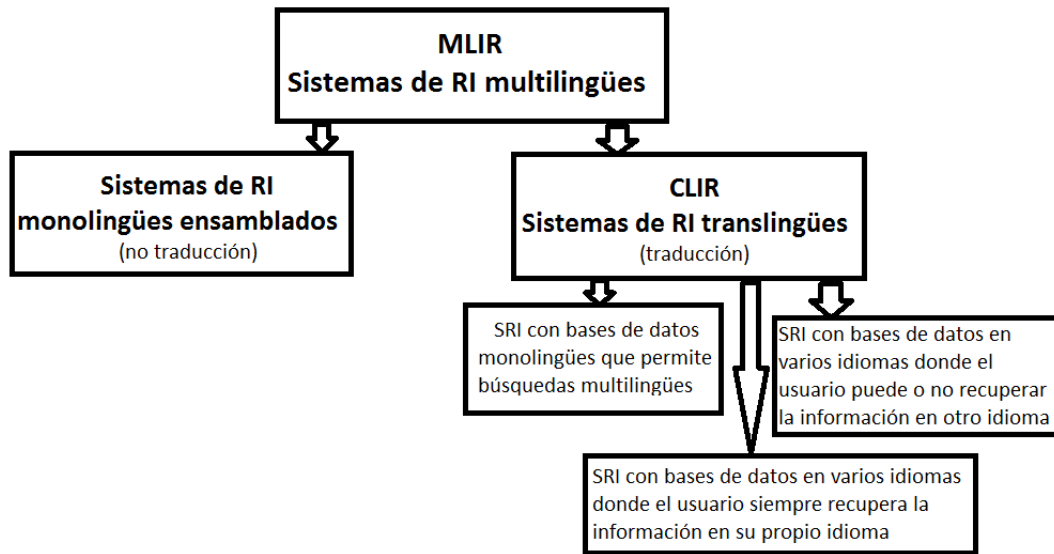


Figura 4. Taxonomía de los sistemas MLIR

Aunque ambos tipos de sistemas MLIR trabajan con varios idiomas, su funcionamiento es muy distinto. Los primeros están formados por dos o más SRI monolingües ensamblados, de modo que el usuario solamente accede a la base de datos del idioma de la consulta. No existe ningún proceso de traducción implicado. Los segundos, denominados sistemas CLIR (por sus siglas en inglés, *Crosslingual or Cross-Language Information Retrieval*) son muy numerosos y presentan diversos tipos de arquitecturas. El único nexo en común entre todos los sistemas CLIR es que siempre contienen un recurso o herramienta lingüístico para el proceso de traducción.

1.1.1.2.1. SRI MONOLINGÜES ENSAMBLADOS

Dado que esta tesis doctoral se centra en los recursos y herramientas lingüísticos aplicados a un tipo de SRI, se explicará brevemente las características de estos sistemas. Los sistemas MLIR monolingües ensamblados se caracterizan por mostrar un SRI en varios idiomas, de modo que el usuario puede plantear su consulta en cualquiera de las lenguas ofertadas por el sistema, el cual ofrecerá una lista ordenada de documentos redactados exclusivamente en el idioma de la consulta. Aunque el sistema se encuentra disponible en varios idiomas, el usuario solamente puede realizar búsquedas monolingües en cada una de las colecciones a través del motor de

búsqueda disponible para cada idioma. Ningún proceso de traducción está implicado en estos sistemas ya que las bases de datos son independientes.

Un conocido ejemplo de SRI monolingüe ensamblado es Google, donde los usuarios pueden consultar documentos hasta en 129 idiomas diferentes pero no es posible una recuperación translingüe.

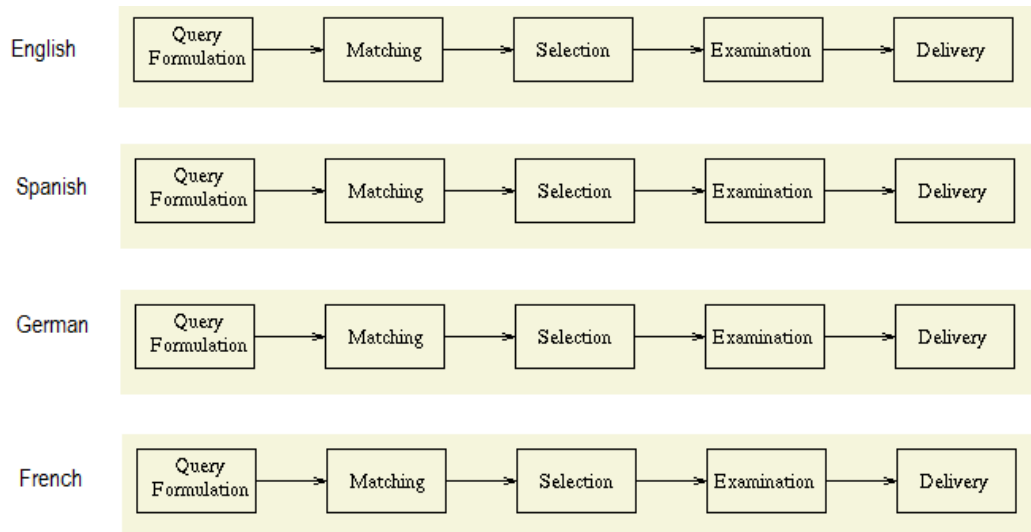


Figura 5. Ejemplo de un sistema MLIR monolingüe ensamblado.

1.1.1.2.2. RECUPERACIÓN DE INFORMACIÓN TRANSLINGÜE

Los sistemas CLIR son un tipo de sistemas MLIR. La principal diferencia de CLIR con los sistemas MLIR monolingües ensamblados es que estos sistemas implican, al menos, la participación de dos lenguas y por tanto, siempre se aplica una técnica de traducción.

CLIR es entendida como planteada una pregunta determinada, estos sistemas se componen de una colección multilingüe de documentos y recuperan información relevante sin importar que lengua sea usada (Grefenstette, 1998). Los sistemas CLIR son muy variados y existen numerosas tipologías, el único nexo en común en todos ellos es la aplicación de uno o varios recursos o herramientas lingüísticos para la traducción de las preguntas, palabras claves, documentos y/o respuestas.

Un tipo muy frecuente de sistemas CLIR es aquellos donde el usuario al plantear su consulta en el sistema en un determinado idioma, podrá recuperar documentos escritos en todos los idiomas disponibles en el SRI. Los documentos recuperados podrán ser, o no ser, útiles para el

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

usuario dependiendo de su competencia lingüística en esos idiomas. En la figura 6 se puede ver un ejemplo de este tipo, en donde el usuario al plantear su necesidad de búsqueda en español el sistema CLIR podrá proporcionarle una lista conjunta de respuestas en español, inglés y francés que son los idiomas disponibles.

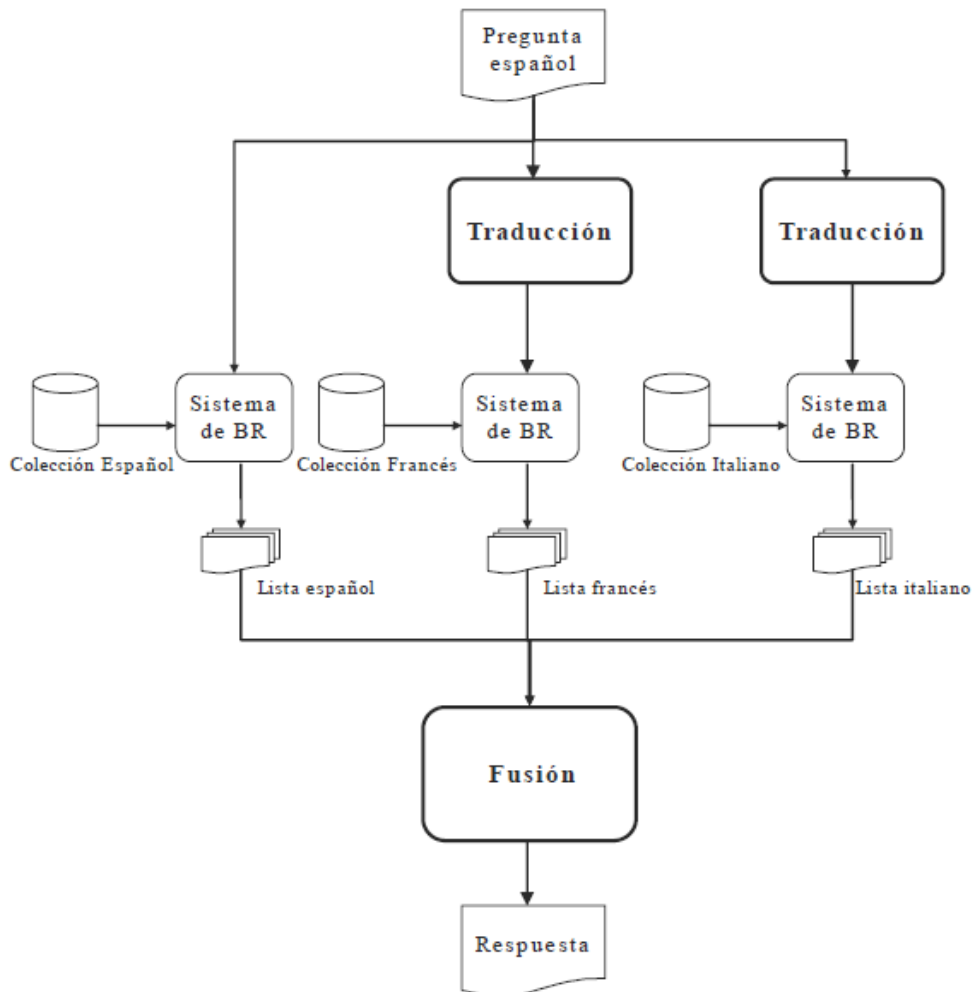


Figura 6. Ejemplo de un sistema CLIR (Aceves, 2008)

En algunos sistemas CLIR el proceso de la traducción es doble, de modo que el usuario al plantear su consulta en su idioma podrá recuperar la información en el mismo idioma de su consulta, aunque la fuente de información estuviera disponible en una lengua diferente (véase figura 7).

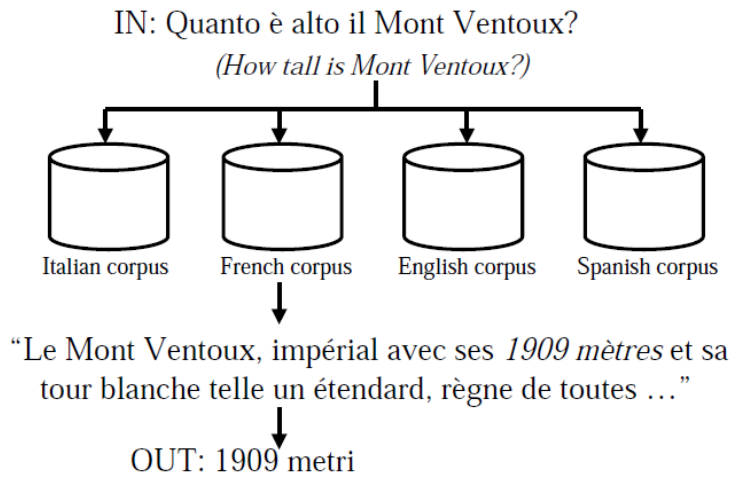


Figura 7. Ejemplo de otro sistema CLIR (Magnini, 2002)

También se entienden como sistemas CLIR aquellos que satisfacen la necesidad de búsqueda del usuario en su idioma o en otro, a pesar de que la base de datos de documentos se encuentre en un solo idioma, distinto al de la consulta (véase figura 8).

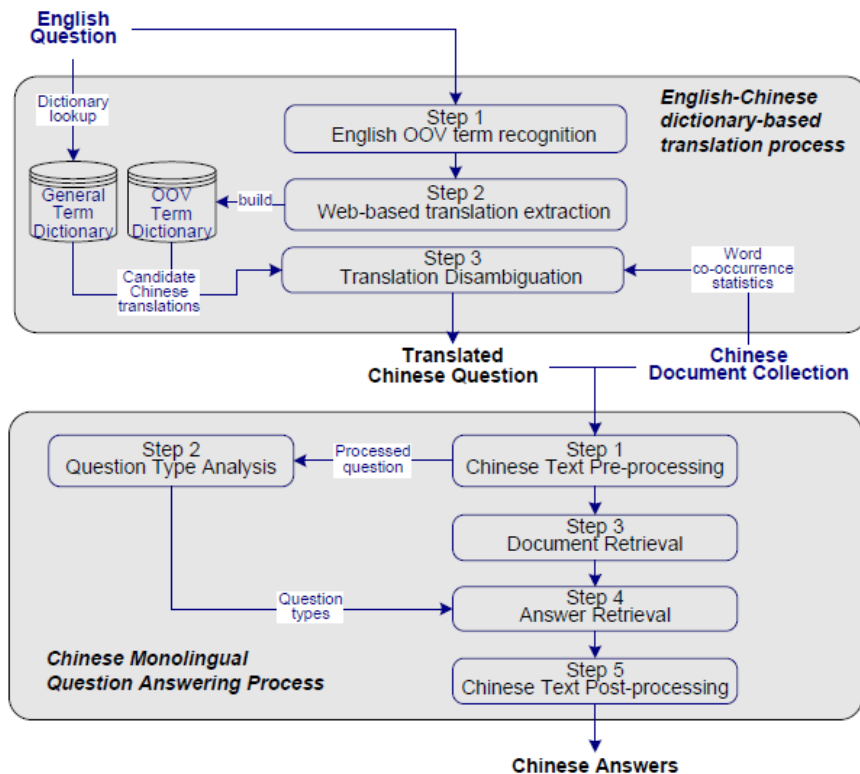


Figura 8. Ejemplo de sistema CLIR con base de datos monolingüe (Zhang *et al.*, 2007)

Tradicionalmente, se ha atribuido a CLIR el problema de ofrecer documentos a los usuarios que ellos no pueden leer (Oard y Gonzalo, 2001). Sin embargo, esto no es del todo verdad. Uno de los primeros trabajos en CLIR lo realizó Salton en 1970, donde comparaba la efectividad de las preguntas en inglés y alemán con preguntas obtenidas usando un tesoro bilingüe para recuperar documentos en ambas lenguas. Salton demostró empíricamente que CLIR, usando un tesoro bilingüe *ad hoc*, es comparable con la RI monolingüe. Usualmente en CLIR un tesoro multilingüe de cualquier tipo es creado para presentar una lista de descriptores para cada documento en una colección y las relaciones semánticas entre ellos, y cada término en el tesoro debe ser traducido para cada idioma implicado. Los descriptores se pueden incluir en el tesoro manualmente o automáticamente (López Ostenero, 2002).

Aunque, como ya hemos mencionado, las primeras iniciativas investigadoras en el relación a CLIR se le atribuyen a Salton (1970), el primer *Workshop* dedicado enteramente a CLIR se celebró en Zürich y fue organizado por la *Association for Computing Machinery* (ACM) durante la conferencia SIGIR-96 (Grefenstette, 1998). En la actualidad, hay cuatro grandes foros internacionales dedicados a la evaluación de sistemas de recuperación que centran su atención en técnicas y procedimientos relacionados con CLIR: *Text REtrieval Conference* (TREC)², *Conference and Labs of the Evaluation Forum*, anteriormente conocida como *Cross-Language Evaluation Forum* (CLEF)³, *NII Text Collection for IR Systems* (NTCIR)⁴ y *Language Resources and Evaluation Conference* (LREC)⁵ (Olvera-Lobo, 2009).

En noviembre de 1992 tuvo lugar un acontecimiento clave: la primera conferencia TREC, patrocinada por los estadounidenses *National Institute of Standards and Technology* (NIST) y *Advanced Research Projects Agency* (ARPA) (Harman, 1993). Reunió a investigadores del ámbito de la RI para discutir los resultados que ofrecían cada uno de sus sistemas, tomando como base una nueva gran colección de evaluación. En este primer encuentro participaron 22 grupos (entre universidades y empresas) de los Estados Unidos, el Reino Unido, Alemania y Australia. Muy pronto, se constituyó en una iniciativa sin precedentes que, en gran parte, hereda la filosofía de Cranfield y es la principal y más significativa contribución a la evaluación de la RI en varios sentidos.

En la conferencia se estableció la creación de una colección constituida por un conjunto de documentos —que incluía el *Wall Street Journal*, *San Jose Mercury News*, artículos de informática de Ziff, informes del Registro Federal, patentes estadounidenses, resúmenes del

² Disponible en trec.nist.gov/

³ Disponible en www.clef-campaign.org/

⁴ Disponible en research.nii.ac.jp/ntcir/index-en.html

⁵ Disponible en www.lrec-conf.org/

Departamento de Comercio, entre otros—, por un conjunto de temas y por un conjunto de valoraciones de relevancia. Lo novedoso del experimento fue su tamaño y variedad —dos gigabytes que, para la época, era una escala hasta entonces desconocida, y que desde entonces no ha dejado de crecer—.

TREC, que se celebra anualmente, realiza un papel muy importante al ofrecer un paradigma de referencia para la investigación en RI. Lo más significativo es su filosofía global (Ellis, 96) en cuanto que hay una vuelta al “realismo” en la evaluación. Además, a pesar de que el “concepto TREC” se basa en una colección ideal y emplea aproximaciones estadísticas y probabilísticas, la gama de técnicas exploradas es extraordinariamente amplia y refleja la creciente heterogeneidad de las aproximaciones y propuestas en la investigación en RI. Las tareas de investigación llevadas a cabo por los grupos participantes en la Conferencia se organizan en torno a diferentes secciones o apartados temáticos que se denominan *tracks* y que se van actualizando cada año.

Durante los años que TREC ha adoptado idiomas europeos en CLIR, se abordan cuatro enfoques principales: traducción automática (TA) (cada tema o documentos se traducen en las lenguas de destino); el uso de diccionarios bilingües electrónicos u otros recursos lingüísticos existentes; el uso de técnicas N-grama; y el uso de recursos de corpus para entrenar o capacitar a los mecanismos de MLIR.

Un año después de la creación de TREC se celebró por primera vez *NII Test Collection for IR Systems* (NTCIR). NTCIR está patrocinado por *Japan Society for Promotion of Science* (JSPS) y *National Center for Science Information Systems* (NACSIS), y se dedica a la evaluación de sistemas translingües entre el inglés y algunas lenguas asiáticas.

Los objetivos son: (i) animar la investigación en tecnologías de acceso a la información proporcionando una gran colección de *test* reutilizables para experimentos y una infraestructura de evaluación común que permite la comparación de sistemas; (ii) proveer un fórum para grupos de investigación interesados en la comparación de sistemas e intercambio de ideas en una atmósfera informal; e (iii) investigar los métodos de evaluación para mejorar las técnicas y procedimiento de acceso a la información a partir de un conjunto de datos para experimentos a gran escala.

Como indican Gey *et al.* (2005) los esfuerzos realizados por las campañas de evaluación de NTCIR han tenido un impacto significativo al facilitarles a los investigadores y a los desarrolladores de sistemas CLIR la infraestructura necesaria para realizar evaluaciones y ajustes, creando un prestigioso marco internacional que les brinde la oportunidad de examinar y comparar ideas y enfoques. El resultado ha sido un progreso considerable en el diseño de sistemas multilingües y en el desarrollo y construcción de una consolidada comunidad de investigadores en torno a este tema.

En 2000, se crea una nueva conferencia, CLEF (*Conference and Labs of the Evaluation Forum*), el fórum europeo más importante para la evaluación de los SRI multimedia y multilingües. CLEF pretende promover la investigación y el desarrollo en MLIA mediante *a)* el desarrollo de una infraestructura para el testeo, la puesta en marcha y evaluación de SRI en los idiomas europeos tanto en contextos monolingües como translingües, y *b)* la creación de *tests* reutilizables que los sistemas desarrollados puedan usar en diferentes bancos de pruebas. El objetivo final es animar y fomentar las tecnologías de RI desarrolladas en Europa para garantizar su competitividad a esfera global.

Otra iniciativa destacada es *International Conference on Language Resources and Evaluation* (LREC), organizado cada dos años por la ELRA (*European Language Resources Association*) con el apoyo de las instituciones y organizaciones relacionadas con las tecnologías del lenguaje humano. El foro LREC pretende convertirse en el motor directivo para crear recursos lingüísticos útiles para las tecnologías disponibles de la ingeniería lingüística, y para su posterior evaluación. Para alcanzar este objetivo, ELRA es una asociación activa en la identificación, distribución, colección, validación, normalización, mejora y promoción de la producción de recursos lingüísticos, favoreciendo el desarrollo de un campo científico de recursos lingüísticos y evaluación.

Aunque numerosas conferencias y foros internacionales abordan CLIR, cabe mencionar una alternativa nacional: la Sociedad Española de Recuperación de Información (SERI), creada en 2009 a partir de la celebración en junio de ese mismo año del primer Congreso Español de Recuperación de Información (CERI). Uno de sus principales cometidos es el impulso y desarrollo de investigaciones realizadas por estudiosos españoles en cualquier campo relacionado con los objetivos y ámbito de trabajo de esta asociación.

1.1.2. SISTEMAS DE BÚSQUEDA DE RESPUESTAS

Durante toda la historia de humanidad el conocimiento, se comunica, se guarda y se maneja en forma de lenguaje natural. Actualmente, con el desarrollo de procesamiento del lenguaje natural (PLN), las computadoras pueden ayudar al tratamiento y acceso a este conocimiento.

Sin embargo, lo que es conocimiento para nosotros –los seres humanos– no lo es para las computadoras. Para combatir esta situación, se dedica mucho esfuerzo al desarrollo de técnicas destinadas a habilitar a las computadoras a entender el texto.

El PLN es la tarea de estudiar, diseñar e implementar sistemas computacionales capaces de utilizar y comprender el lenguaje natural tal y como lo usamos los humanos, es decir, posibilitando una comunicación fluida y eficaz (García Cumbreras, 2009).

El lenguaje humano ha sido estudiado y analizado por numerosos investigadores de disciplinas muy distintas. Todos ellos se han centrado en diferentes aspectos del lenguaje según los objetivos de su disciplina y por tanto, se les han planteado problemas de diferente índole y origen, sin embargo, todos sus descubrimientos han servido de base para el desarrollo y evolución del PLN actual.

El esquema general de la mayoría de los sistemas y métodos que involucran el procesamiento de lenguaje es el siguiente: *a)* el texto no se procesa directamente sino se transforma en una representación formal que preserva sus características relevantes para la tarea específica; *b)* el programa principal manipula esta representación, transformándola según la tarea; *c)* los cambios hechos a la representación formal (o la respuesta generada en esta forma) se transforman en el lenguaje natural.

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

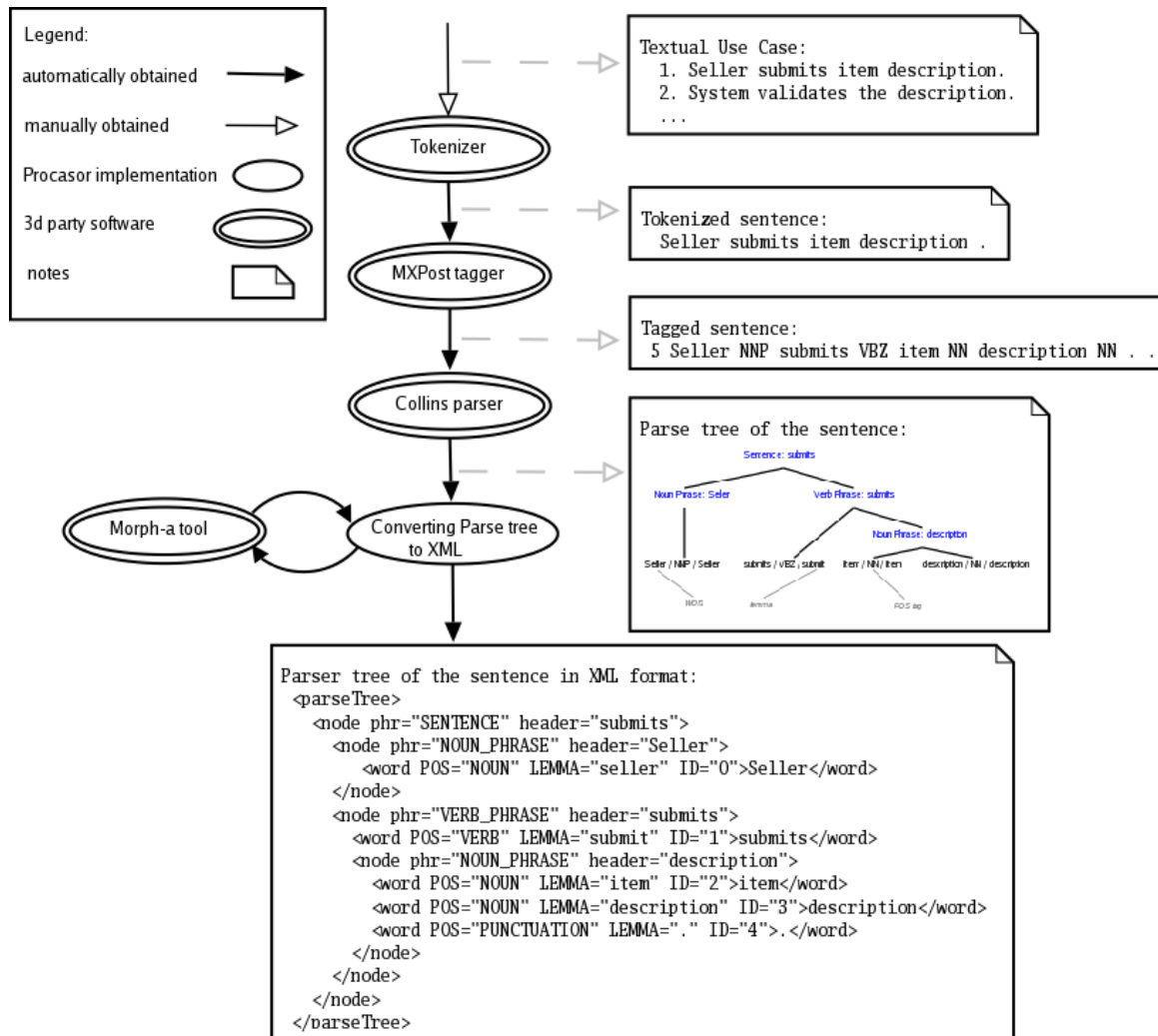


Figura 8. Ejemplo de un sistema de PLN⁶

Todas estas técnicas han servido de base para el desarrollo de PLN complejos, los cuales fundamentan los actuales SRI, y por ende, los SBR.

La búsqueda de información es una de las actividades vitales del ser humano. No es una actividad vinculada exclusivamente con Internet y ni siquiera es una actividad nueva, sino que es tan antigua como la Humanidad. Lo que marca la diferencia con nuestro tiempo es la súper abundancia de información, pero cuanto mayor oferta de información existe, más necesario son los sistemas eficientes de búsqueda (Codina Bonilla, 2011).

Un paso en la evolución hacia la mejora de la RI son los SBR. La búsqueda de respuestas implica a diferentes disciplinas tales como la psicología, la filosofía, la lingüística, la educación, la traducción, la documentación, la computación y las ciencias de la información. Como

⁶ Disponible en: <http://aiya.ms.mff.cuni.cz/procasor/current/developerDocumentation.html>

consecuencia, los estudios de inteligencia artificial, en particular el PLN, y los aspectos de RI en esta área se benefician del conocimiento adquirido en otras disciplinas.

Los SBR se presentan como una alternativa a los tradicionales SRI tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de presentar al usuario una lista de documentos relacionados con la búsqueda. Estos sistemas están abriendo un nuevo campo de investigación y cada día están adquiriendo mayor relevancia. Los SBR multilingües son aquellos sistemas donde el idioma en el que se plantea la pregunta puede ser diferente a la lengua en la que está escrito el documento recuperado, pero se diferencian del resto de sistemas MLIR integrados en que éstos no recuperan documentos completos sino que responden con una respuesta corta a la consulta planteada.

Algunos autores señalan que ya en los años 30 se hicieron aproximaciones psicológicas al problema de las respuestas a preguntas y que, incluso podemos remontarnos a la época Aristotélica para descubrir reflexiones sobre este asunto desde un punto de vista filosófico (Monz, 2003). Efectivamente, si bien el interés por la tecnología vinculada a los SBR data de hace relativamente poco (Greenwood, 2006), los primeros SBR –monolingües y en lengua inglesa– surgieron en los años 60 (Simmons, 1965) y utilizaban bases de datos de dominio restringido con información estructurada. No obstante, el avance en las técnicas de PLN fue una importante razón para el desarrollo de los actuales SBR (Belkin and Vickery, 1985)

1.1.2.1. EVOLUCIÓN DE LOS SISTEMAS DE BÚSQUEDA DE RESPUESTAS

El comienzo de los SBR dentro del campo de la Inteligencia Artificial fue a comienzos de los años sesenta cuando los investigadores utilizaban técnicas basadas en el conocimiento para el almacenamiento y recuperación de la información. Los primeros sistemas desarrollados por estos investigadores crearon reglas que permitían derivar expresiones lógicas en segmentos naturales de manera automática (Belkin y Vickery, 1985).

Los primeros SBR se pueden clasificar en tres grandes grupos Monz (2003): los sistemas basados en bases de datos, aquellos basados en textos y por último, los basados en inferencia de conocimiento.

Uno de los primeros SBR creado fue ORACLE (Phillips, 1960). Este sistema se caracterizaba porque producía un análisis sintáctico de la pregunta y del corpus de texto que podía contener la respuesta. Cada pregunta es relacionada analógicamente con unos fragmentos textuales en el corpus, pero además, es también transformada en una frase declarativa, que reordena las

palabras y combina los verbos auxiliares con aquellos principales. Tres años más tarde, se crea el SBR denominado PROTOSYNTHESIS (Simmons *et al.*, 1963).

Otro de los primeros SBR se denominó BASEBALL (Green *et al.*, 1961, 1963). Este sistema se caracterizaba por responder preguntas en inglés sobre cualquier tema relacionado con el deporte del béisbol. El SBR funcionaba con un simple buscador donde se planteaban preguntas simples sin conectores, ni superlativos o frase subordinadas -*How many games did the Yankees play in July?*-. Se trataba de un sistema de acceso abierto cuya base de información se obtenía de una base de datos realizada manualmente, por tanto, forma parte de los SBR con información estructurada.

En 1977 se crea LUNAR, desarrollado por *Bolt Beranek and Newman Inc.* (Woods *et al.*, 1972; Woods, 1977). Este sistema, también del grupo basado en bases de datos, se creó para poder ofrecer análisis geológico de las piedras lunares recogidas por las misiones Apolo. El SBR está formado por dos bases de datos que mostraban respuesta automáticamente a las preguntas de los usuarios -*What is the average concentration of aluminium in high alkali rocks?*-. Al contrario que el anterior este sistema era de acceso restringido. Tanto el sistema BASEBALL como éste tuvieron una muy buena acogida ya que su funcionamiento era óptimo.

Otros de los primeros SBR fueron "Chat-80" (Warren, 1981), una base de datos geográficos -*Which is the largest African country?*- con una versión moderna que convierte la pregunta al lenguaje SQL, y "Eliza" (Weizenbaum, 1966), un sistema de diálogo que simulaba un psicoanalista y que puede considerarse precursor de los actuales *chatbot* -software diseñado para simular una conversación inteligente con uno o más humanos por medio de texto y/o audio-. Por último, los antecesores más inmediatos de los sistemas web de búsqueda de respuestas, en los que aquí nos centramos, son los sistemas de búsqueda en documentos de texto, los cuáles tomaron un importante impulso a partir de la conferencia TREC-8 (Voorhees, 1999).

Las técnicas utilizadas en los SBR para el tratamiento y análisis de la información son enormemente variadas y, según el enfoque adoptado por los diseñadores del sistema, éstas van desde la utilización de métodos estadísticos hasta la aplicación de técnicas de PLN -si bien suele ser ésta última la aproximación más frecuente-. Los diferentes criterios y decisiones que pueden tomarse en el diseño y desarrollo de la arquitectura de un SBR dan lugar a una heterogénea tipología de los mismos. Las lenguas de trabajo del sistema, el contenido temático de los documentos de la base de datos, el nivel de organización de la información (estructurada o no) contenida en los mismos o el grado de interactividad con el usuario, entre otros aspectos, determinan distintos tipos de SBR.

El criterio de la lengua es muy popular en un gran número de las clasificaciones propuestas (Adiwibowo y Adriani, 2007; Izquierdo *et al.*, 2007; Roger *et al.*, 2007; Solorio *et al.*, 2005). Permite distinguir principalmente entre sistemas monolingües y multilingües, aunque en algunos casos puede incorporarse un cuarto tipo relativo a los SBR multilingües que utilizan el inglés como idioma pivote (García-Cumbreras *et al.*, 2006).

Por otro lado, la cobertura temática de los documentos de la base de datos también es un aspecto tenido en cuenta (Harabagiu *et al.*, 2000; Magnini *et al.*, 2001; Moldovan *et al.*, 2003; Roger *et al.*, 2007). Así, mientras que unos son SBR de dominio abierto, es decir, de tema general o multidisciplinar, no restringido, otros cuentan con una colección especializada en un ámbito temático determinado. Como es de esperar, los sistemas que mejores resultados obtienen son los segundos ya que una colección de documentos de tema homogéneo facilita el uso de herramientas y de recursos léxico-semánticos específicos para el PLN (Wedgwood, 2005) lo que contribuye a evitar la ambigüedad y a obtener una mayor eficacia en el tratamiento lingüístico.

El nivel de organización de la información contenida en los documentos indizados por el sistema origina una tipología que distingue entre SBR que trabajan con información estructurada, aquéllos que lo hacen con documentos semiestructurados (Katz *et al.*, 2007) y los que gestionan una colección no estructurada (Cucerzan y Agichtein, 2005).

Además, en el diseño de los SBR se pueden establecer distintos niveles de interactividad con el usuario. En este contexto, la capacidad de interacción conversacional es un aspecto que cobra aún mayor relevancia. Aunque es notable el interés creciente por los SBR capaces de mantener un diálogo continuado con los usuarios –y a éstos, a su vez, les resultan mucho más atractivos– la mayor parte de los sistemas aún no alcanzan un grado de comunicación que permita el desarrollo de una secuencia de preguntas-respuestas a modo de conversación real. Mientras que los asistentes virtuales cuentan con una extensa base de datos de la que extraen las respuestas, los sistemas basados en textos cuentan con una colección de documentos (entendido el documento en su sentido más amplio) a partir de los que extraen la información y elaboran la respuesta. La ventaja de los SBR plenamente interactivos es que utilizan cada nuevo *input* para conseguir cierta retroalimentación y afinar en la búsqueda de la respuesta más adecuada. Por ejemplo, están más preparados para resolver de manera exitosa las preguntas enlazadas, que incluyen referencias a preguntas o a respuestas anteriores (¿quién descubrió la penicilina? ¿dónde nació?). Estos SBR interactivos –aún muy escasos, como hemos indicado– encajan perfectamente con el enfoque de los tradicionales servicios de referencia. Además, las aplicaciones en este ámbito de referencia (Pomerantz, 2005) parten de la premisa de que la consulta primera del usuario suele ser errónea, ya que éste puede tener cierta dificultad para expresar su necesidad de información. Se cuestiona la validez de la pregunta inicial y se busca

una reformulación más exacta ofreciendo mayor orientación durante la consulta. Por el contrario, la inmensa mayoría de los SBR consideran que el usuario refleja correctamente en la pregunta su necesidad de información. Asimismo, establecen que hay una única respuesta posible -que puede estar contenida o no en el corpus de documentos- de modo que, disponiendo de la base de conocimientos adecuada, siempre se podrá satisfacer la necesidad de información.

Pero, además de los mencionados, se están desarrollando otros novedosos tipos de SBR como los que aplican datos enlazados (QALD, *Question Answering over Linked Data*) (López *et al.*, 2013), los centrados en testar a los SBR para que actúen como evaluadores en los exámenes de entrada a la universidad (Miyashita *et al.*, 2014), entre otros.

1.1.2.2. ARQUITECTURA DE LOS SISTEMAS DE BÚSQUEDA DE RESPUESTAS

El funcionamiento de los SBR se basa en los modelos de respuestas cortas (Blair-Goldensohn *et al.*, 2004) que dividen las preguntas asignándole una palabra clave o etiqueta que indica el tipo de preguntas que pueden ser respondidas. La ventaja principal que ofrece al usuario es que éste no ha de consultar documentos completos para obtener la información requerida puesto que el sistema ofrece la respuesta correcta en forma de un número, un sustantivo, una frase corta o un fragmento breve de texto (Pérez-Coutiño *et al.* 2004b). Cuando la respuesta aparece mencionada explícitamente en la base de datos documental, los SBR suelen responder a las preguntas planteadas aplicando procedimientos relativamente sencillos como el conteo de frecuencias, la aplicación de expresiones regulares o patrones, o los análisis sintácticos superficiales (*shadow parsing*) (Alfonseca, 2011).

Mientras el desarrollo de los SBR representa un progreso en RI, todavía muestran ciertas restricciones. Muchos de ellos han sido solamente desarrollados como prototipos, o versiones de prueba, y algunos pocos han sido comercializados. Algunos investigadores han diseñado y creado sistemas que son presentados y evaluados en varios foros o conferencias. Sin embargo, debido a que su utilidad está limitada a contextos muy específicos, o por problemas en su implementación, solo unos pocos están disponibles para todos los usuarios.

Los SBR pueden ser de dominio abierto y responder preguntas de diversos campos, como es el caso de START⁷ *Natural Language Question Answering System*, o NSIR *Question Answering*

⁷ Disponible en: start.csail.mit.edu/

*System*⁸. El otro tipo de SBR son aquellos de dominio especializado, centrados en un tema especializado, como es el caso de HONQA⁹ (*Health On the Net Foundation*) or MedQA¹⁰, (*Medical Question Answering System*) (Frank, *et al.*, 2006). Los sistemas de dominio específico utilizan recursos lingüísticos específicos que permiten dar respuestas más precisas.

Los SBR normalmente presentan una sencilla interfaz con un motor de búsqueda mediante el que los usuarios pueden formular su pregunta, e incluso algunos proporcionan un listado de las últimas cuestiones introducidas para ayudarles a entender cómo han de plantearlas. Ciertamente, estos sistemas intentan emular el comportamiento del lenguaje humano por lo que tratan de entender la pregunta formulada en lenguaje natural y proporcionar respuestas adecuadas. En otras palabras, la interpretación del lenguaje natural por el sistema es un proceso esencial en el desarrollo de los SBR (Belkin y Vickery, 1985; Sultan, 2006). Tanto es así que el análisis de la pregunta, así como la búsqueda y la extracción de las respuestas son tres importantes tareas llevadas a cabo por los SBR, las cuáles implican, al menos, el procesamiento de las preguntas, el procesamiento de los documentos y el procesamiento de las respuestas (Kangavari *et al.*, 2008).

Los SBR buscan la interacción entre el usuario y el sistema, ya que la relación sistema-usuario no es unidireccional. La interacción ayuda al SBR a encontrar mejores respuestas, y a su vez, el sistema ayuda al usuario a encontrar la respuesta más rápidamente (Chang *et al.*, 2001). No obstante, todavía es necesario profundizar en el diseño de estos sistemas interactivos que hagan posible la existencia de un verdadero *feedback* entre preguntas y respuestas, y que el usuario se comunique a nivel conversacional con el sistema.

Otra de las características de estos sistemas es que facilitan la búsqueda de la información al usuario mediante la alineación de los documentos. El alineado constituye uno, sino el que más, de los procesos críticos a la hora de valorar la efectividad de un SBR, ya que se refiere al orden en el que el sistema presenta los resultados a los usuarios, quienes, como es lógico, esperan encontrar las respuestas más relevantes a sus necesidades situadas entre las primeras (Martínez Méndez, 2002). Aunque supuestamente los SBR deberían presentar el conjunto de respuestas ordenadas por orden de relevancia atendiendo a las necesidades del usuario, en anteriores trabajos se ha comprobado que los SBR actuales suelen tener un índice de relevancia interno del cual basan la ordenación de las respuestas (Olvera-Lobo y Gutiérrez-Artacho, 2010). Además, no siempre todas las respuestas que respondan correctamente la pregunta del usuario pueden ser adecuadas a las necesidades del usuario, ya que algunas respuestas pueden ser más específicas o contener más información que otras (Monz, 2003).

⁸ Disponible en: tangra.si.umich.edu/clair/NSIR/html/nsir.cgi

⁹ Disponible en: services.hon.ch/cgi-bin/QA10/qa.pl

¹⁰ Disponible en: monkey.ims.uwm.edu:8080/MedQA/

Dentro de los SBR se están desarrollando actualmente arquitecturas especiales que se equiparan a los procesos mentales realizados por el ser humano. Algunos autores, como Graesser y Murachver (1985), desarrollaron un extenso espectro de la psicología de la respuesta de preguntas, sin embargo, todavía los desarrolladores de estos sistemas no han implementado con éxito esos estudios en los SBR.

Como el principal objetivo de los SBR es encontrar respuestas exactas y correctas para las consultas de los usuarios, los tres principales pasos que siguen estos sistemas son el análisis de las preguntas, la búsqueda y la elección de las respuestas. Para ello, se componen de tres módulos: análisis de la pregunta, procesamiento de los documentos y procesamiento de la respuesta (Kangavari *et al.*, 2008).

Las preguntas formuladas por el usuario son procesadas inicialmente por el módulo de análisis o reformulación de la pregunta. Este módulo es esencial puesto que de la cantidad y calidad de la información extraída en este análisis dependerá en gran medida el rendimiento de los restantes módulos y por ende, el resultado final del sistema (Aceves Pérez, 2008). La reformulación se usa a menudo en los SBR para recuperar respuestas en colecciones amplias de documentos (Verona y Motta, 2004). En este módulo, el SBR debe detectar todas las palabras claves, que aportan información relevante acerca del tema de la pregunta, e identificar el tipo de respuesta esperada.

El módulo de procesamiento de los documentos, también denominado por algunos autores como el módulo de selección de pasajes relevantes, constituye una parte esencial de la arquitectura de los SBR. En este módulo se utilizan las palabras claves identificadas en el primer módulo para extraer una primera selección de párrafos que se consideran relevantes a las preguntas. Cada pasaje recuperado es acompañado de un peso numérico, el cual indica la relevancia del texto respecto a las palabras claves de la consulta (Aceves Pérez, 2008).

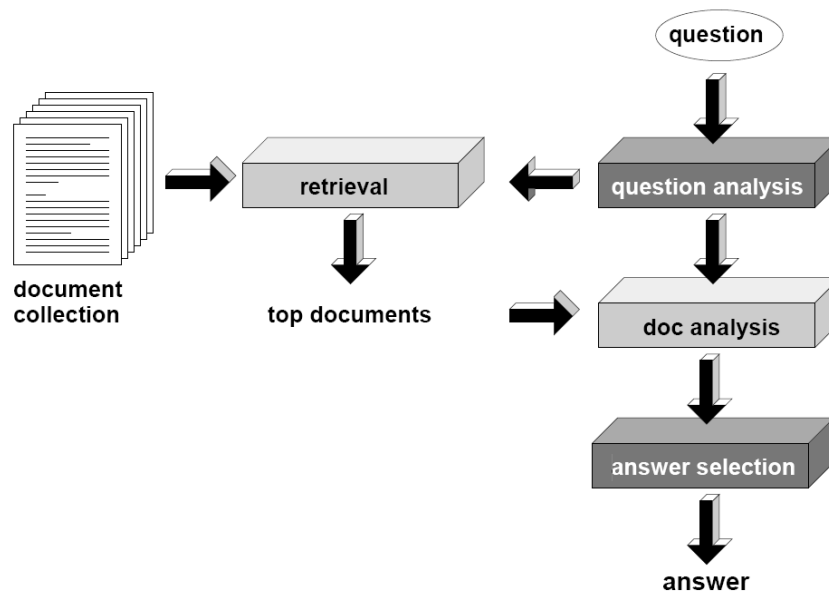


Figura 9. Arquitectura general de un SBR textual (Monz, 2003)

Y finalmente, el módulo de procesamiento de las respuestas donde se realizan dos acciones fundamentales: la extracción y la validación de las respuestas. Primero, se extrae la respuesta candidata de los documentos que han sido recuperados por el motor de búsqueda en el módulo anterior. Tras ello, se valida la respuesta con el filtrado y clasificación de las respuestas, y el sistema final sugiere la respuesta o las respuestas que evaluará el usuario (Kangavari *et al.*, 2008). El objetivo de esta etapa es eliminar los posibles párrafos incorrectos contenidos en la lista devuelta por el SBR (Rodrigo *et al.*, 2010).

1.1.2.3. ANÁLISIS DE LAS PREGUNTAS

El primer paso en todos los SBR es la identificación del tipo de pregunta planteada por el usuario. Para realizar esta primera tarea, el SBR en cuestión tiene que realizar un tratamiento lingüístico de las preguntas planteados por los usuarios, clasificándolas en función del tipo de respuesta que se espera obtener.

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

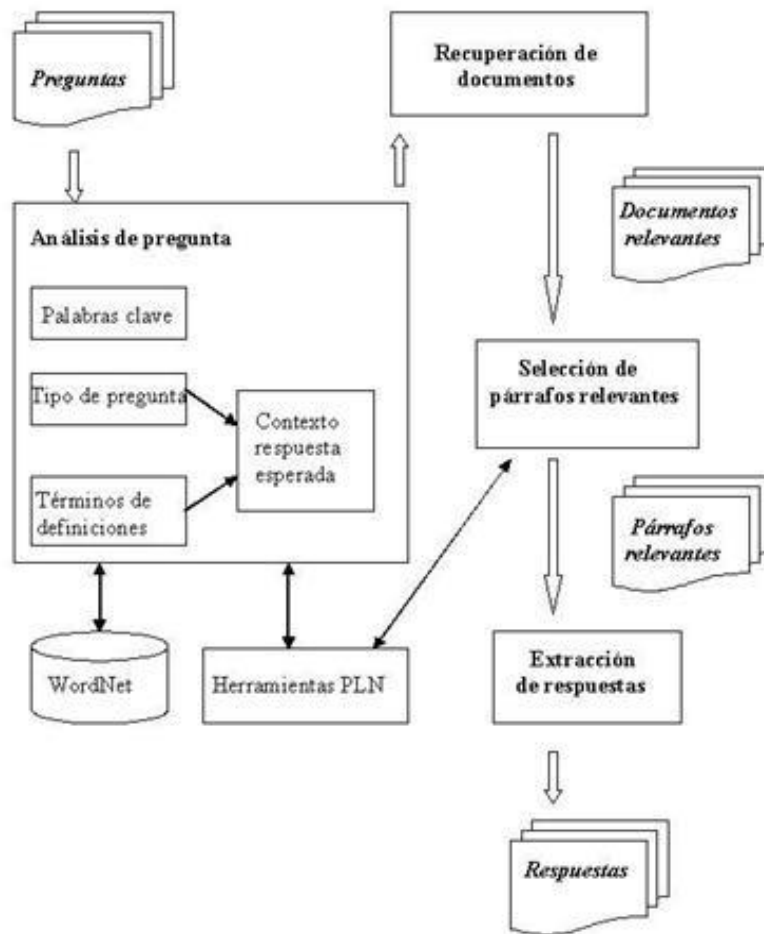


Figura 10. Proceso que tiene lugar en un SBR

La elaboración de un buen sistema de análisis y clasificación de las preguntas será clave para el éxito o fracaso del sistema. Para ello, los SBR han desarrollado unas taxonomías para la clasificación de las preguntas que faciliten los siguientes pasos y que se adaptan con facilidad a las posibles nuevas preguntas que puedan plantearse en el sistema (Olvera-Lobo y Robinson-García, 2009). Este análisis implica su tratamiento lingüístico, y su clasificación se hace en función del tipo de respuesta que se espera obtener.

Las principales taxonomías que se han empleado para la clasificación de preguntas (Pomerantz, 2005) son las basadas en *Wh-words*, es decir, a *who*, *what*, *when*, *where*, *why* y *how*. Es una taxonomía muy usada debido a su simplicidad (Hovy *et al.*, 2001; Moldovan *et al.*, 1999). Sin embargo no es muy efectiva en los SBR ya que no engloba todos los tipos de preguntas posibles como veremos a continuación.

Son numerosas las preguntas que se pueden plantear a los SBR, ya sean monolingües o multilingües. Estos sistemas suelen aceptar todo tipo de preguntas, aunque hay algunos como por ejemplo EAGLi¹¹ que facilitan una lista de posibles preguntas para orientar al usuario. No obstante, aunque los sistemas no suelen restringir el tipo de preguntas, aquellas preguntas más sofisticadas con una carga superior de información no suele ser procesada de igual modo por todos los sistemas, y la mayoría no ofrecen respuesta válida.

Aunque como ya veremos en los apartados posteriores las técnicas usadas por los SBR para el procesamiento y análisis de la información varían sustancialmente de un sistema a otro, sí que existe cierto consenso a la hora de clasificar los tipos de preguntas posibles atendiendo a razones de finalidad para su división. Existen varias clasificaciones dependiendo de cada autor (Hernández-Rubio, 2008), pero la clasificación más generalizada es la siguiente: preguntas de definición, preguntas factuales, preguntas de lista, y un grupo de preguntas que forman parte de cualquier de los otros tipos, preguntas con restricción de tiempo.

Las preguntas de definición solicitan información abierta en términos de sinonimia y formulación. Estas preguntas responden al tipo “*what/who is X?*” en referencia a una institución, persona, cosa o concepto. Las respuestas más relevantes serán aquellas que respondan con un menor número de palabras la necesidad de información de manera eficaz. Las buenas respuestas deberían probablemente ser muy similares en naturaleza a las entradas de una enciclopedia (Greenwood y Saggion, 2004). En el contexto de los SBR, estas preguntas se caracterizan por responder a preguntas simples cuya respuesta satisfactoria dependerá de las intenciones del usuario y de la colección de documentos que disponga el SBR. En el caso de los sistemas CLIR, además de los problemas mencionados anteriormente, también cumplirá un papel esencial la herramienta o el recurso lingüístico encargado de facilitar la traducción de la consulta, documentos o respuestas. Tanto en los foros CLEF y TREC han puesto especial atención a la evaluación de las preguntas de definición, por ello en las baterías de preguntas desarrolladas por ambos foros un tanto por ciento elevado de las preguntas pertenece a este tipo.

Con las preguntas factuales el usuario está buscando obtener un dato muy concreto, como el nombre de una persona, una fecha, un lugar o una cantidad. Estas preguntas se pueden dividir en varios subtipos, dependiendo del pronombre interrogativo usado (*who, what, when, where, why, how much, how many, which, whose*).

En las preguntas de lista cerradas, las respuestas solicitadas se limitan a un conjunto de información, pueden ser formuladas en pronombres interrogativos (*which, how*) o un imperativo (“*name all of London’s airports*”). Por ejemplo, ante la siguiente pregunta a un hipotético SBR “¿qué partidos políticos forman el Congreso de los Diputados en España?”, se espera que el

¹¹ Disponible en <http://eagl.unige.ch/EAGLi/>

sistema nos devuelva la siguiente lista de valores (véase tabla 2), es decir, una lista de catorce valores.

<i>Pregunta</i>	<i>¿Qué partidos políticos forman el Congreso de los Diputados en España?</i>
Respuesta (ordenada de forma descendente por el número de diputados)	Partido Popular (PP)
	Partido Socialista Obrero Español (PSOE)
	Convergència i Unió (CiU)
	Izquierda Unida-Los Verdes: La Izquierda Plural
	Unión Progreso y Democracia (UPyD)
	Partido Nacionalista Vasco (EAJ-PNV)
	Amaiur (AMAIUR)
	Esquerra Republicana de Catalunya (ERC)
	Bloque Nacionalista Gallego (BNG)
	Coalición Canaria-Nuevas Canarias (CC-NC-PNC)
	Compromís-Q (COMPROMÍS)
	Unión del Pueblo Navarro (UPN)
	Foro de ciudadanos (FAC)
Geroa Bai	

Tabla 2. Lista de respuesta para una hipotética pregunta de lista

Por último, hay que destacar las denominadas preguntas con restricción de tiempo, o con restricciones temporales. Estas preguntas pueden incluirse dentro de cualquier grupo de preguntas mencionadas anteriormente, ya que la única peculiaridad que les caracteriza es que buscan una información determinada que se encuentra restringida en un cierto periodo de tiempo (Pustejovsky et al, 2002). El uso de este tipo de preguntas permite evaluar la exactitud de las respuestas del sistema, y de ese modo, evaluar el funcionamiento y calidad del mismo.

En el tratamiento o análisis de las preguntas, incluido dentro del primer módulo de los SBR, los sistemas identifican cada una de las palabras que forman parte de la consulta planteada descartando las palabras vacías de contenido. Si se plantease la siguiente pregunta a un hipotético SBR, el sistema seguiría los siguientes pasos:

¿Cuándo y dónde nació el presidente Obama?

1. El primer paso que realizaría el sistema sería identificar cada una de las palabras que componen la consulta.

	Descripción	¿Da información?	Lexema de análisis
Cuándo	Pronombre interrogativo	Sí	Cuándo
Y	Nexo copulativo	No	
Dónde	Pronombre interrogativo	Sí	Dónde
Nació	Verbo (Pret. Perfecto Simple)	Sí	Nacer
El	Artículo determinado	No	
Presidente	Sustantivo	Sí	Presidente
Obama	Nombre común	Sí	Obama

Tabla 3. Análisis de las palabras que forman una pregunta en un SBR

2. Tras la identificación, el sistema procede a evaluar la relevancia de la palabra para la RI. Normalmente, el SBR desecha todos los artículos determinados o indeterminados, posesivos, demostrativos, conjunciones, preposiciones, etc. El sistema solamente va a ser capaz de identificar los lexemas de las palabras. Tras la selección de las palabras relevantes para la RI, el sistema pasará a recuperar la información correspondiente.

1.1.2.4. SELECCIÓN DE DOCUMENTOS Y PASAJES RELEVANTES

Tras reconocer las entidades nombradas, el SBR va a dividir en pasajes la información recuperada con las palabras seleccionadas en la etapa anterior y va a analizar mediante reconocedores de entidades con el fin de identificar los pasajes susceptibles de contener la respuesta (Clarke *et al.*, 2000). En esta etapa radica la particularidad de los SBR dentro de los SRI, ya que aunque utilizan técnicas similares a las usadas en la RI, se recupera solamente trozos de textos o pasajes en lugar de documentos completos.

El SBR selecciona una serie de pasajes que supuestamente satisfacen las necesidades de la respuesta planteada, tras ello se calcula la relevancia de cada uno de estos pasajes para pasar a la siguiente y última etapa que consiste en la extracción de la información.

La clasificación de las preguntas es uno de los aspectos que más atención requieren, ya que un tercio de los errores se producen en esa etapa (Moldovan *et al.*, 2003). En este sentido, el proceso de clasificación de las preguntas tiene dos objetivos (Li *et al.*, 2004): por un lado, se reducen los tipos de respuestas esperadas. Por otro, se identifica el enfoque, estableciendo el tipo

de pregunta en función del tipo de información que debe contener la respuesta esperada. Así por ejemplo, ante la pregunta “¿cuánto mide el monte Everest?”, distinguiremos entre “monte Everest” (el objeto sobre el cual se solicita una información) y la altura del mismo (el enfoque desde el cual se solicita la información). Identificar tanto el objeto de la pregunta como el enfoque de la misma constituye una importante ayuda para localizar la respuesta adecuada.

1.1.2.5. SISTEMAS DE BÚSQUEDA DE RESPUESTAS MULTILINGÜES

Los SBR monolingües, que como su propio nombre indica son aquellos que solamente trabajan en una sola lengua, juegan un papel muy importante en el desarrollo de los SBR multilingües y translingües, ya que estos sistemas son la base de cualquier tipo de SBR más sofisticado. No se pueden estudiar los sistemas multilingües sin comprender verdaderamente a estos, ya que fueron los pioneros y actualmente son los más comunes en Internet.

Hoy en día, los SBR monolingües siguen siendo un reclamo en la industria de RI, sin embargo, los nuevos modelos desarrollados han evolucionado radicalmente en comparación con los primeros monolingües. Actualmente los investigadores y desarrolladores de esta área se centran en SBR monolingües especializados en un ámbito determinado, ya sean dedicados exclusivamente a un área de especialización determinada, como por ejemplo los SBR monolingües dedicados exclusivamente a recuperar pasajes de la Unión Europea, o ya sean especializados en otro tipo de géneros textuales como son los SBR orales.

La aparición de los SBR multilingües es relativamente reciente, y muchos de los sistemas existentes se encuentran en una etapa experimental. La diferencia de los SBR multilingües y translingües es la misma que existe entre los sistemas MLIR y CLIR, con la diferencia de que en este caso se extraen exclusivamente respuestas cortas. Como hemos mencionado en CLIR, en la literatura disponible se percibe cierta confusión entre ambos tipos, multilingües y translingües, ya que los propios investigadores y desarrolladores mezclan ambos conceptos y confunden un término con el otro.

Se denominan SBR multilingües a los SBR monolingües ensamblados disponibles en varias lenguas, y a todos los SBR translingües, sistemas que emplean alguna técnica de traducción para la recuperación de las respuestas. El primer tipo se caracteriza por presentar un conjunto de sistemas monolingües coordinados en los que cada uno extrae las respuestas de una colección propia de documentos, como es el caso del SBR multilingüe HONqa. El segundo tipo es más

numeroso y, al igual que los sistemas CLIR, ofrece una amplia gama de sistemas muy diferentes entre sí. Normalmente, la arquitectura de los SBR translingües es muy similar a los monolingües, con la única diferencia de que incorporan un módulo de traducción y/o herramienta lingüística para la recuperación multilingüe. Este tipo de sistemas es menos frecuente, ya que entraña más problemas de funcionamiento.

Ambos tipos de SBR destacan frente al resto de sistemas MLIR en que permiten tener acceso a información más concreta de manera más rápida y sencilla –lo que permite un aumento en la posibilidad de éxito en la búsqueda– (Aceves Pérez, 2008).

La traducción es crucial en CLIR porque las preguntas y los documentos no siempre comparten el mismo idioma. Los principales problemas de traducción identificados en los SBR translingües son la ambigüedad léxica, la falta de cobertura de traducción, los lexemas multimodales, y los errores en recursos léxicos (Diekema, 2003). Sin embargo, los aspectos relacionados con la traducción se han abandonado relativamente durante el desarrollo de estos sistemas. La mayoría de los trabajos en este campo son realizados por especialistas en Inteligencia Artificial y Computación, así que a los problemas mencionados anteriormente no se les dan la prioridad adecuada.

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

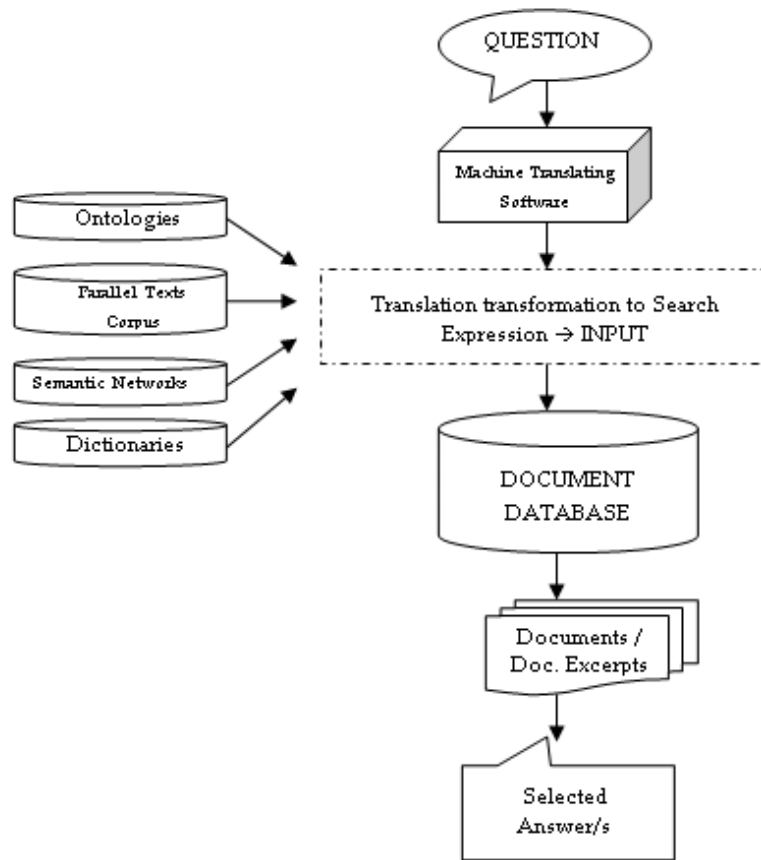


Figura 11. Esquema básico de un SBR translingüe que incorpora traducción automática (Olvera-Lobo y García-Santiago, 2010)

Los SBR translingües no se enfrentan a los problemas de traducción del mismo modo, ya que existen grandes diferencias en la gestión de los documentos y preguntas. Por ello, en los siguientes apartados se explicará la relación que se establece entre CLIR y la Traducción, y las características de los diferentes recursos y herramientas lingüísticos incorporados para la traducción de las consultas, las palabras clave, la colección de documentos –creando un corpus paralelo de traducciones–; y/o las respuestas.

1.1.3. SISTEMAS DE BÚSQUEDA DE RESPUESTAS EN LA WEB

En la actualidad el número de SBR disponibles en la Web de manera gratuita, ya sean sistemas completamente desarrollados o demos, no es demasiado amplio. Como puede verse en la Tabla

4, no todos los sistemas expuestos se encuentran disponibles en la actualidad, bien porque simplemente han desaparecido, o porque llevan varios años sin actualizarse o porque dependen de otros para poder ofrecer una cobertura máxima. Además, los idiomas utilizados son bastante reducidos y solamente funcionan correctamente aquellos que trabajan con fuentes documentales en inglés.

Nombre del sistema	Monolingüe o Multilingüe	Especializado o no	Idiomas	Otros datos de interés
ACQUILEX ¹²	Multilingüe	No	EN, DU	
AFMED ¹³	Monolingüe	Sí	EN	
AquaLog ¹⁴	Monolingüe	No	EN	No disponible.
Ask Hermes ¹⁵	Monolingüe	Sí	EN	No funciona porque trabaja con MedQA.
Auskunft.org ¹⁶	Monolingüe	no	DE	No disponible
BBLIP ¹⁷	Monolingüe	No	EN	
CiceroCustom ¹⁸	Monolingüe	No	EN	
DFKI ¹⁹	Multilingüe	No	EN, SP, DE	No disponible
EAGLI ²⁰	Monolingüe	Sí	EN	
ExtrAns ²¹	Monolingüe	Sí	EN	No disponible
Ferrer ²²	Monolingüe	No	EN	
Gigablast ²³	Monolingüe	No	EN	
HonQA ²⁴	Multilingüe	Sí	EN, IT, FR	
Inferret ²⁵	Multilingüe	No	EN, CH, PO, FR, JA, SV, ES	
Jeopardy ²⁶	Monolingüe	No	EN	
MedQA ²⁷	Monolingüe	Sí	EN	No disponible
NSIR: Question Answering System ²⁸	Monolingüe	No	EN	En la actualidad no funciona correctamente
OASQA ²⁹	Monolingüe	NO	EN	Hay que descargárselo.
PRIBERAM ³⁰	Monolingüe	No	PT	
QALL-ME ³¹	Multilingüe	No		
Qristal ³²	Monolingüe	no	FR	No disponible

12 <http://www.cl.cam.ac.uk/research/nl/acquilex/>

13 <http://eagl.unige.ch/EAGLm/>

14 <http://technologies.kmi.open.ac.uk/aqualog/>

15 <http://www.askhermes.org/>

16 <http://auskunft.org/>

17 <http://blip.cs.brown.edu/>

18 <http://www.languagecomputer.com/>

19 <http://experimental-quetal.dfki.de/>

20 <http://eagl.unige.ch/EAGLi/>

21 <http://www.ifi.uzh.ch/arvo/cl/extrans/>

22 <http://www.languagecomputer.com/>

23 <http://www.gigablast.com/ask.html>

24 <http://services.hon.ch/cgi-bin/QA10/qa.pl>

25 <http://asked.jp/edw/pc/>

26 <http://www.research.ibm.com/deepqa/>

27 <http://monkey.ims.uwm.edu:8080/MedQA/>

28 <http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi>

29 <http://www.osqa.net/>

30 <http://www.priberam.pt/dlpo/consultar.aspx>

31 <http://qallme.fb.ku.edu/index.php?location=home>

32 http://www.qristal.fr/Telechargement_demo.htm

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

Nombre del sistema	Monolingüe o Multilingüe	Especializado o no	Idiomas	Otros datos de interés
Qualim ³³	Monolingüe	No	EN	No disponible
QAS ³⁴	Monolingüe	No	EN	
QUAST ³⁵	Monolingüe	No	EN	
QUETAL ³⁶	Multilingüe	No	EN, DE	No disponible
SEMOTE ³⁷	Monolingüe	No	EN	No disponible
SNOCat ³⁸	Monolingüe	Sí	EN	
START: Natural Language Question Answering System ³⁹	Monolingüe	No	EN	
TextMap Questions Answering System ⁴⁰	Monolingüe	No	EN	No disponible
VSIL ⁴¹	Monolingüe	NO	PT	
Watson ⁴²	Monolingüe	No	EN	
Website Term Browser ⁴³	Multilingüe	No	EN, SP	No disponible
Wolfram Alpha ⁴⁴	Monolingüe	Sí	EN	

Tabla 4. Lista de SBR disponibles en la Web

33 <http://demos.inf.ed.ac.uk:8080/qualim/>

34 <http://qas.intellexer.com/index.html>

35 <http://www.lsi.upc.edu/~qast/2009/>

36 <http://www.dfki.de/pas/f2w.cgi?ltp/quetal-e>

37 <http://www.semote.com/>

38 <http://eagl.unige.ch/SNOCat/>

39 <http://start.csail.mit.edu/>

40 <http://brahms.isi.edu:8080/textmap/>

41 <http://visl.sdu.dk/QA.html>

42 <http://www.research.ibm.com/deepqa/>

43 <http://nlp.uned.es/wtb/uned/query-uned.html>

44 <http://www.wolframalpha.com/>

1.1.3. LA TRADUCCIÓN EN CLIR

La importancia de la documentación y de las tecnologías en la disciplina de la Traducción es un hecho aceptado por todos. Desde hace algo más de una década, los traductores emplean más tiempo de su trabajo realizando tareas no lingüísticas que traduciendo (Muñoz Martínez, 2000). Se puede definir a los traductores como comunicadores profesionales multilingües con un amplio dominio de un gran conjunto de herramientas. Tener maestría y soltura en el manejo y gestión de estas es esencial para realizar una correcta traducción. Sin embargo, decidir qué herramientas son realmente útiles en el trabajo no es una tarea fácil.

Por otra parte, los profesionales de la Traducción se caracterizan por trabajar con temas especializados o semiespecializados que no siempre confluyen en los campos interdisciplinarios de la traducción. Se enfrentan a textos de diferente grado de especialización que versan sobre temas que no siempre conocen con antelación, o a listas de palabras o de frases donde no hay ni referencias textuales externas ni coherencia interna (Wright y Wright, 1997). Su labor consiste en realizar la correcta transcodificación del texto desde la lengua origen a una determinada lengua meta. Como normalmente no son especialistas en los temas a tratar, deben suplir sus carencias de conocimiento documentándose lo máximo posible, de ahí la importancia de la disciplina de la Documentación en la profesión del traductor y en la disciplina en sí. De acuerdo con Olvera-Lobo *et al.* (2007), se puede definir a los profesionales de la Traducción como comunicadores bilingües o multilingües que necesitan tener un amplio dominio de un conjunto de herramientas que son tan imprescindibles como el conocimiento de sus lenguas de trabajo, es decir, las técnicas y fuentes documentales. Además, hoy en día también los recursos y programas informáticos son tan importantes como los idiomas en su formación y profesión. Unas nociones básicas de la disciplina de la Documentación proveen al traductor de unos conocimientos que mejoran y aligeran el proceso documental de cualquier encargo, uno de los más tediosos e importantes para la realización de un trabajo correcto. Se puede afirmar que para estos profesionales es más importante tener maestría en la adquisición y gestión de la información que ser experto en el tema sobre el que se trabaje (Wright y Wright, 1997).

Existe una estrecha relación entre la disciplina de la Traducción y la de Documentación, ya que la primera necesita aplicar los avances y técnicas de documentación en la práctica de sus profesionales. Según Mayoral Asensio (2001), la disciplina de la Traducción presenta una doble interdisciplinariedad, por una parte interna en relación a las disciplinas de nivel superior de las que derivan los Estudios de Traducción, como la Lingüística, la Sociología, la Semiótica, etc., y por otra una interdisciplinariedad externa que se manifiesta en las relaciones establecidas con la ejecución de la disciplina, como la documentación, la localización, la informática, entre otras. La relación entre documentación y traducción se refleja tanto en su dimensión de actividad como en la de disciplina, ya que se trata de dos segmentos contiguos y solapados dentro del continuo de la comunicación, dando lugar a una relación bidireccional entre ambas (Mayoral Asensio, 2005). Los traductores necesitan la documentación para adquirir una serie de destrezas que les permitan conocer y usar ciertas herramientas y fuentes de información, pero la disciplina de la

Documentación también necesita a la propia de la Traducción para desarrollar aspectos de los suyos puesto que ambas son disciplinas auxiliares (Recoder y Cid, 2004).

Otra disciplina que cumple un importante cometido en la traducción es la Terminología que realiza, junto con la Documentación, un papel clave para el desarrollo y éxito del producto traducido. El traductor necesita la terminología cuando se enfrenta a un texto especializado, ya que la especificidad temática de un texto se refleja en la terminología que contiene y, por tanto, el traductor debe ser capaz de manejar correctamente los términos (Cabré Castellví, 2000). Al mismo tiempo, se establece una relación bidireccional entre estas dos disciplinas, Documentación y Terminología, que se explica en el carácter transdisciplinario de la documentación y el trabajo documental previo de cualquier trabajo terminológico. Por su parte, la Terminología necesita de la Documentación, puesto que los documentos técnicos son la base del trabajo terminológico (Cabré Castellví, 1993). Como afirma la misma autora (2000), la terminología es una pieza necesaria para el trabajo documental y los documentos son imprescindibles para el trabajo terminológico.

En las últimas décadas, estas disciplinas han sufrido fuertes cambios que han repercutido en la forma de procesamiento y gestión de la información. El almacenamiento y acceso a los recursos en papel han evolucionado a nuevas formas de procesamiento electrónico que permiten acceder a un mayor número de recursos así como a una mayor cobertura para los profesionales de la Traducción (Pérez Hernández, 2002). No obstante, y aunque el desarrollo en el procesamiento y gestión de la información documental y terminológica ha aumentado el número de recursos disponibles y ha permitido la creación de unos perfiles más versátiles que encajan con diferentes usos y usuarios (Sager, 2001), estos cambios han supuesto un esfuerzo añadido en el aprendizaje y desarrollo de las competencias y destrezas traductoras. El conocimiento y utilización de las fuentes de información y de los recursos disponibles adecuados a cada problema, así como un buen manejo de los sistemas automatizados de acceso, gestión y recuperación de los mismos son dos de las principales tareas requeridas en el marco de la competencia documental de los traductores. Esta competencia, también denominada por Kelly (2002) como subcompetencia instrumental profesional, se completa con las destrezas adquiridas mediante otras competencias o subcompetencias relacionadas. Una de las competencias relacionadas con la disciplina de la Terminología es aquella que la autora denomina subcompetencia temática y que permite el conocimiento de los lenguajes de especialidad así como de la temática especializada del texto.

Sin embargo, no son solamente los profesionales y la disciplina de la Traducción los que necesitan a estas disciplinas para su desarrollo, sino que, como ya se ha mencionado, existe una relación bidireccional. En el caso de la Documentación, debido a los nuevos avances en la RI y a la globalización de la Web, se han creado sistemas multilingües de búsqueda, en los que la traducción y los recursos lingüísticos ocupan un lugar tan importante como las herramientas documentales e informáticas. Este tipo de sistemas ha abierto un nuevo campo de investigación, donde no solamente se cuestionan cuáles son los métodos más eficaces para la RI, sino que también cuáles son los recursos traductológicos necesarios para la correcta transcodificación de la misma.

1.1.3.1. RECURSOS UTILIZADOS PARA LA TRADUCCIÓN EN LOS SBR MULTILINGÜES

En el ámbito de la RI se están creando herramientas documentales e informáticas (monolingües, multilingües y translingües) que pueden ayudar substancialmente a los traductores en su trabajo –además de resultar útiles para otros usuarios con necesidades de información de lo más diversas. El desarrollo de las herramientas multilingües se encuentra todavía en evolución y necesitan varios años de estudios e investigación para su mejora y aplicación. Uno de los principales problemas a los que se enfrentan estas herramientas es la traducción (Diekema, 2003), tanto de las consultas planteadas por los usuarios como de las fuentes documentales que responden a las mismas.

Aunque las investigaciones en esta área se iniciaron hace algo más de una década, estos sistemas son unos auténticos desconocidos fuera del ámbito de CLIR. Los actuales investigadores en el área intentan buscar nuevos métodos para que la RI sea lo más eficiente posible pero sin atender detenidamente a los problemas lingüísticos. Sin embargo, si no se encuentra una solución óptima en relación a la traducción y a los recursos utilizados, difícilmente el sistema podrá recuperar una información relevante para el usuario. Por este motivo, la traducción cobra un protagonismo fundamental en este entorno y permite analizar el problema desde una nueva perspectiva.

Dentro de las técnicas de traducción empleadas por los SBR multilingües se pueden destacar ocho tipos de recursos lingüísticos principales, a saber, las bases de datos, los corpus, los diccionarios, las enciclopedias, EuroWordNet, las ontologías, las páginas web y los tesauros, así como dos tipos de herramientas lingüísticas –los traductores automáticos y las gramáticas computacionales. Estos recursos y herramientas, junto con sus diferentes tipos y subtipos, no funcionan de igual modo y usan diferentes formas de procesar la información. Además, existen otros métodos para solucionar el problema de la comunicación, como es el caso de la transliteración que juega un papel muy importante en varios de estos sistemas.

Aunque anteriores trabajos realizados (Diekema, 2003) han identificado cuatro principales fuentes de traducción aplicadas a CLIR –ontologías, diccionarios bilingües, traductores automáticos y corpus (Figura 12), se ha comprobado que esta tipología ha aumentado en los últimos años y ciertos recursos han comenzado a ser bastante utilizados. Por ello, los resultados se han dividido en dos grandes grupos que recogen los principales recursos y herramientas utilizadas por los actuales SBR multilingües.

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

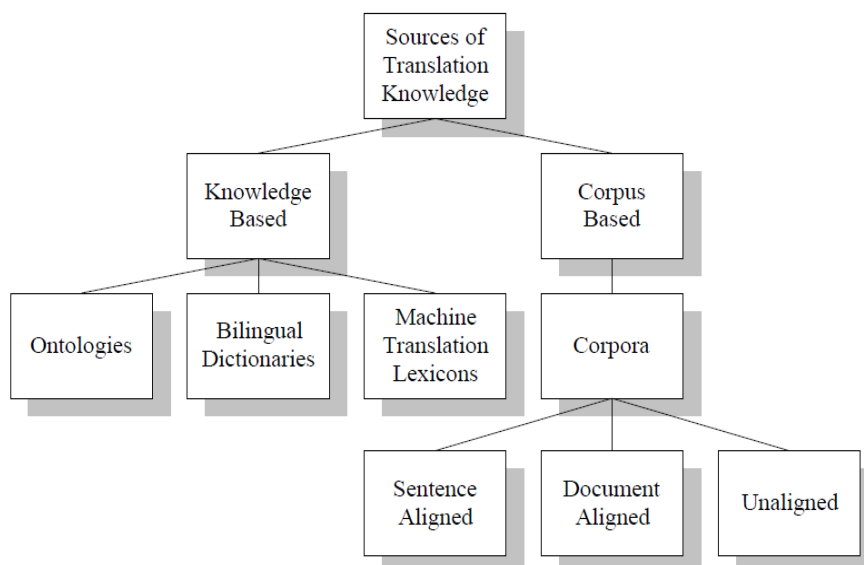


Figura 12. Recursos utilizados por los SBR multilingües para solucionar el problema de la traducción (Diekema, 2003)

En el apartado 1.1.3.1.1., se describen los principales recursos lingüísticos disponibles y sus tipos, distinguiendo en cada uno de ellos las características, usos, ventajas y desventajas que ofrecen en CLIR. Los recursos utilizados son: *a)* las bases de datos donde se explicará la indexación conceptual; *b)* los corpus donde se distinguirán las diferencias entre los sistemas que utilizan corpus comparables, aquellos que usan corpus paralelos y los que incorporan corpus monolingües combinados, así como aquellos sistemas que utilizan los corpus bilingües y multilingües como forma de traducción; *c)* los diccionarios electrónicos que son uno de los recursos más utilizados junto con los corpus y los traductores automáticos; *d)* las enciclopedias, en concreto se destaca la implementación de dos famosas enciclopedias *online* que están teniendo mucho auge en los SBR, como son Wikipedia y Wiktionary; *e)* EuroWordNet, recurso que a pesar de sus múltiples ventajas en el ámbito multilingüístico no ha sido utilizado demasiado por estos sistemas; *f)* las ontologías, cuyo interés e incorporación está aumentando vertiginosamente en los últimos años; *g)* la incorporación de las páginas web monolingües y multilingües como método para realizar la traducción; y por último, *h)* los tesauros que aunque en anteriores clasificaciones no han sido incorporados, su versatilidad y popularidad no puede pasar desapercibida en esta clasificación.

En el apartado 1.1.3.1.2., se explica el uso de las gramáticas computacionales y los traductores automáticos en los SBR multilingües. Se ha denominado a este apartado herramientas lingüísticas puesto que, aunque algunos investigadores anteriores los han considerado como recursos lingüísticos, estas herramientas no se pueden enmarcar como tales ya que son colecciones de datos que no contienen información sobre lenguas, sino que procesan tal información. Dentro del apartado de los traductores automáticos, se realizará un estudio minucioso de sus ventajas e inconvenientes ya que se trata de la herramienta lingüística más popular para los creadores de los SBR multilingües. En las últimas campañas de CLEF, casi una

treintena de participantes incorporaron un software de traducción automática para solucionar el problema de la traducción de las consultas, palabras claves o documentos (Bos y Nissim, 2006).

1.3.2. RECURSOS LINGÜÍSTICOS

La necesidad de superar las barreras lingüísticas y de documentarse correctamente son dos de las principales características de los profesionales de la Traducción. Dentro de las competencias traductoras está la denominada competencia lingüística (Sánchez Ramos, 2004) que supone la adquisición de ciertas destrezas lingüísticas que les permita adquirir una alternancia lingüística o capacidad de pasar de una lengua a otra. Aunque es necesario tener un buen conocimiento y maestría de las lenguas de trabajo, no siempre es suficiente cuando tratamos con los lenguajes de especialidad. Por ello, los avances y trabajos coordinados de otros especialistas –como los terminólogos, documentalistas, lexicógrafos, desarrolladores web, entre otros–, facilitan enormemente su trabajo y desarrollo de capacidades, así como la adquisición de mecanismos de alternancia lingüística y control de los mecanismos de interferencias.

Todo ellos han fomentado la evolución radical que han sufrido durante los últimos años los especialistas de la traducción. Esta evolución se ha plasmado en la creación de recursos que son unos de los activos lingüísticos más importantes para la traducción. Entendemos por recursos lingüísticos cualquier colección de datos en forma textual, ya sea hablada o escrita, o con información sobre lenguas, en donde el objetivo de la tecnología sea el procesamiento del material lingüístico (Bel y Marimon, 2008). La evolución en la ciencia computacional ha supuesto una remodelación en la consulta de estos recursos, lo que permite una consulta más cómoda y rápida en grandes conjuntos de datos y materiales.

1.1.3.1.1.1. BASES DE DATOS

Una base de datos es un sistema de información complejo formado por un conjunto de datos en soporte automatizado pertenecientes a un mismo campo y almacenados sistemáticamente para su posterior uso. Estos recursos lingüísticos se componen, por lo menos, de los siguientes elementos: un conjunto de registros, cada uno de los cuales describe una entidad mediante una combinación de campos y sistemas de metadatos, un sistema de indexación, un sistema de consulta y un sistema de ordenación y presentación de resultados (Codina Bonilla, 2011).

Los nuevos avances tecnológicos han impulsado la creación de diferentes tipos que se adecuan a distintas realidades (bases de datos bibliográficas, jerárquicas, especializadas, multidisciplinarias, entre otras). Los SBR multilingües hacen uso especialmente de dos de ellas: las bases de datos relacionales y las terminológicas.

Las bases de datos relacionales son obras diseñadas para generar un conjunto de esquemas de relaciones que permitan almacenar la información con un mínimo de redundancia, pero que a la vez faciliten la RI. Estas bases de datos son muy útiles para estos sistemas porque establecen patrones relacionales entre las distintas lenguas, sin embargo, necesitan de la ayuda de otro recurso o de la traducción de los documentos para ser eficientes. Aunque son utilizadas por

algunos SBR multilingües, y en concreto, por los translingües, este recurso no ha tenido mucha aceptación en el mundo académico y solamente se han constatado una pocas incorporaciones.

Se puede definir las bases de datos terminológicas como aquellas obras que contienen referencias bibliográficas y acceso a las mismas (como las bases de datos bibliográficas), pero contienen además diversos contextos de definición, ejemplos de uso y también contextos de los equivalentes en lengua extranjera, así como otra información complementaria a la entrada terminológica (Sager, 1993). Son de gran utilidad para estos sistemas, y especialmente para aquellos de dominio específico, puesto que establecen patrones de equivalencia entre los términos de varias lenguas.

De acuerdo con Pérez Hernández (2000), las bases de datos terminológicas se adaptan a las siete categorías de datos establecidas por Sager (1993), ya que se caracterizan por tener un determinado corpus especializado, términos de entrada que actúan como una etiqueta para los conceptos, especificación conceptual y semántica de los términos, especificación lingüística del término, especificación pragmática del término, información de mantenimiento o administrativa, y equivalente en otra lengua extranjera. Todas estas características convierten a las bases de datos en un recurso adecuado a las necesidades de los SBR multilingües, ya que trabajan con dos corpus especializados y equivalentes que permiten el acceso y recuperación de la información mediante las preguntas del lenguaje natural (Jung y Lee, 2003). El procesamiento de la información mediante este recurso permite que no se sacrifique la solidez y flexibilidad del SBR.

Sin embargo, el uso de este recurso tal cual se restringe a un número reducido de SBR, ya que presenta varios problemas: *a)* discordancia entre el vocabulario atributo y el procesamiento lingüístico, *b)* necesidad de un conocimiento amplio del recurso por parte de los usuarios, y *c)* uso de otros recursos o medios para ofrecer un amplia cobertura en la RI, como la TA y la indexación textual.

La TA es uno de los recursos más combinados en estos sistemas (véase la sección 1.1.3.1.2.2.), ya que ofrece una serie de características que la posicionan en un buen lugar. La otra opción más usada en combinación con las bases de datos es la indexación conceptual que consiste en traducir los documentos a una representación conceptual independiente del idioma (Peinado Herencia, 2004), es decir, crear un índice común y automatizado para todas las lenguas de modo que no se fusionen los resultados de las búsquedas y se pueda recuperar información en todos los idiomas del sistema. Esencialmente automatizan el proceso por el cual los lectores pueden tratar de adivinar el significado de un término desconocido basándose en las similitudes, en el deletreo o en la pronunciación.

Unas de las principales ventajas de este recurso son el número de lenguas disponibles que permite incorporar en los SBR multilingües, así como la posibilidad de solucionar problemas de equivalencia tan importantes como la sinonimia, hiponimia, meronimia y hipernomimia mediante la desambiguación semántica automática. No obstante, hasta ahora este recurso no se encuentra totalmente depurado, por ello los SBR multilingües que lo incorporan tienen problemas en la indexación y representación del conocimiento.

1.3.2.2. CORPUS

La lingüística de corpus se define como el estudio de las lenguas a través de ejemplos de uso de las mismas (Troyano Jiménez, 2007). Desde hace unos años el auge de la informática y la expansión en el uso de los ordenadores han modificado las técnicas y metodología de trabajo de los corpus.

Se denomina corpus al recurso lingüístico formado por una colección de fragmentos de documentos escritos en una determinada lengua que ha sido seleccionada y ordenada siguiendo unos criterios lingüísticos específicos para ser usada como muestra de un idioma o un lenguaje de especialidad (Sinclair, 1996). Aunque existen corpus en papel, las nuevas tecnologías han facilitado el trabajo y permitido la recopilación del material de forma electrónica, ahorrando material y facilitando la gestión y procesamiento a los desarrolladores, gestores y, sobre todo, a los usuarios.

Basado en las tipologías expuestas por Hunston (2002) y Marcos Marín (1994), proponemos la clasificación general de corpus desarrollada por Alcántara Plá (2007):

- Corpus especializados: compuestos por textos con unas características específicas, más o menos restringidas. Suelen desarrollarse para objetivos muy concretos.
- Corpus generales: al contrario que los anteriores no se encuentran restringidos por ningún tipo de especificación, pudiendo recoger cualquier tipo de texto.
- Corpus comparables: Compuesto por dos o más subcorpus que comparten textos de las mismas características pero redactados en diferentes lenguas.
- Corpus paralelos: Compuesto por dos o más subcorpus que comparten los mismos textos pero redactados en distintas lenguas. La principal diferencia con los comparables es que aquí nos encontramos con textos que, o bien han sido redactados en diferentes lenguas, o bien son traducciones de un texto origen.
- Corpus históricos: Son textos de distintos periodos históricos que se emplean para realizar estudios diacrónicos.

No obstante, para poder considerar cualquier corpus como una representación semántica de calidad es necesario que posean los siguientes requisitos básicos (Jurafsky y Martin, 2000): verificabilidad, desambiguación en la representación y eliminación de las imprecisiones, forma canónica de las palabras, capacidad de inferencia y, por último, expresividad del sistema de representación.

La clasificación que se presenta a continuación atiende al uso de este recurso por los SBR, por ello los dividimos en dos grandes grupos que, a su vez, se complementan: corpus clasificados según idioma y corpus clasificados según estructura.

1.1.3.1.1.2.1. CORPUS CLASIFICADOS SEGÚN IDIOMA

Se denominan corpus lingüísticos a aquellas colecciones de información lingüística (ya sea escrita, hablada o una combinación de ambas), que han sido creadas para ejercer un fin en particular, generalmente para caracterizar una variedad particular de una o más lenguas

(Sperberg-McQueen y Burnard, 2004). Dentro de los corpus clasificados según idioma, hay que distinguir las diferencias entre sus tres grandes grupos: los corpus monolingües, los corpus bilingües y aquellos multilingües.

Los corpus monolingües son aquellos que se recopilan con el objetivo de dar cuenta de una lengua o variedad lingüística. Su incorporación en los SBR multilingües se realiza por medio de dos métodos: mediante corpus monolingües combinados o mediante la combinación de corpus monolingües y diccionarios electrónicos. La primera opción consiste en combinar dos o más corpus monolingües en diferentes lenguas y entre los que no hay una relación total (Diekema, 2003). Este tipo de método solamente ha sido utilizado por aquellos sistemas de ámbito abierto y no ha tenido una repercusión muy amplia ya que los corpus no pueden ser alineados. La segunda alternativa ha sido acogida por un mayor número de sistemas ya que soluciona el problema del escaso número de corpus paralelos existentes. Esta consiste en utilizar una colección de documentos monolingües sobre una determinada área de conocimiento y ofrecer las traducciones mediante un diccionario bilingüe general o especializado dependiendo del ámbito del sistema (Chen y Gey, 2001). Las ventajas de este recurso estriban en que no es necesario alinear los documentos para su uso y en la facilidad para la recuperación de textos, sea la lengua que sea. La única desventaja que ofrece este método es que el uso combinado de dos recursos lingüísticos aumenta el gasto en recursos, gestión y tiempo.

Los corpus bilingües son una fuente inagotable de recursos lingüísticos útiles para el desarrollo de aplicaciones como la lexicografía, la terminografía, la TA, la enseñanza de segundas lenguas, la edición plurilingüe y la búsqueda translingüe de información. El avance de la tecnología en el tratamiento de corpus de lengua escrita ha coincidido con el desarrollo de la tecnología *web* en Internet. Esta coincidencia ha propiciado una confluencia de estrategias en los dos campos, lo que ha fomentado un interés común hacia los mecanismos de codificación y anotación, así como a la explotación de los contenidos (Abaitua, 2002). El uso de los corpus bilingües en los SBR multilingües es bastante habitual, convirtiéndose en uno de los recursos lingüísticos más utilizados. La mayoría de los corpus bilingües utilizados suelen ser *ad hoc*, es decir, son creados para satisfacer una necesidad inmediata de recuperación de cierta información, especializada o no, y escrita en dos idiomas distintos. Suelen ser corpus de tamaño reducido, ya que sus peculiaridades son muy determinadas.

La creación de corpus multilingües aumenta cada día debido a la abundancia de información y a las posibilidades de aprovechamiento que aportan. A los niveles típicos de procesamiento monolingüe (análisis morfológico, lematización, desambiguación, análisis sintáctico) se añade en los corpus multilingües un tipo de tratamiento particular mediante el que se establecen las equivalencias entre las unidades de textuales. Esta fase recibe distintos nombres: emparejamiento, correspondencia, alineación (el más utilizado). La alineación es el proceso que mayor valor añadido aporta a un corpus multilingüe (Abaitua, 2002). Este recurso lingüístico es bastante utilizado por SBR al igual que los corpus bilingües. Las ventajas que aporta en la implementación son numerosas pero el hecho de que no existan numerosos disponibles en la Web y de que los existentes tengan unas características muy específicas, hace que este recurso se convierta en un problema para los desarrolladores de los SBR. La mayoría de los corpus multilingües que se implementan en los SBR multilingües han sido creados *ad hoc* por los responsables del sistema con la intención de solventar el problema de la RI, pero no para ser utilizado con otra intención más.

1.1.3.1.1.2.2. CORPUS CLASIFICADOS SEGÚN ESTRUCTURA

La necesidad de diseñar sistemas cada vez más competitivos con la prontitud que el mercado exige en la sociedad de la información se ha traducido durante las dos últimas décadas en la creación de numerosos recursos lingüísticos con peculiaridades de lo más diversas. Los corpus también se pueden estructurar en función de la estructura o formato que presenten.

Se denomina corpus paralelos a la colección de textos formada por los originales y sus respectivas traducciones a diferentes lenguas (Sinclair, 1996). Normalmente, los corpus paralelos constan de dos lenguas, la original y aquella a la que se traducen todos los textos, sin embargo, es posible la incorporación de todas las lenguas que se deseen. Este recurso plantea numerosas ventajas en los SBR multilingües, ya que mediante la alineación de los mismos nos encontramos con dos corpus completamente iguales que ofrecen información coherente. Para facilitar su explotación, los textos están dispuestos unos al lado de los otros en párrafos o frases, de modo que sea más fácil la extracción de equivalencias. Este proceso de alineación en ocasiones se realiza de modo automático o semiautomático, lo que acelera enormemente la gestión del recurso (Peinado Herencia, 2004; Diekema, 2003). Algunos sistemas se ayudan de sistemas de TA o diccionarios electrónicos por si no hubiera una equivalencia completa en alguno de los segmentos alineados.

No obstante, este recurso presenta algunas desventajas tanto en el funcionamiento como en el material presentado. En primer lugar, resulta complicado encontrar disponible un número abundante de este tipo de corpus, sobre todo para algunos pares de lenguas, y por tanto, no siempre pueden ser incorporados. Algunos grupos de investigación han creado sus propios corpus y han traducido los textos correspondientes a las lenguas del sistema, pero como hasta ahora el problema de la comunicación correcta no ha sido un tema a tratar, la mayoría utilizaban traductores automáticos gratuitos disponibles en la red que normalmente ofrecían unos textos incomprensibles para el usuario. Este recurso lingüístico sería bastante recomendable si los textos fueran traducidos por profesionales de la traducción, de modo que los textos meta y originales tuvieran una calidad comparable.

Se denominan corpus comparables a aquellos que se componen de textos similares en más de una lengua (Sinclair, 1996). Estos recursos son utilizados por los SBR multilingües y translingües porque permiten superar el problema de la comunicación sin tener que utilizar otro recurso adicional, como un traductor automático, y esto evita la inevitable distorsión producida en las traducciones de corpus paralelos. Además, al estar los dos textos escritos por nativos no presentan los problemas de coherencia sintáctica, gramatical y léxica que aparecen en los sistemas que usan otros recursos.

Los SBR que incorporan este recurso se caracterizan por disponer de varios corpus comparables en los idiomas que trabaja el sistema. Para ello, es necesaria una fase previa de creación de los corpus en la que se recuperen textos que traten sobre el mismo tema y sean similares en especialización. Tras ello, es necesario alinear los documentos de los corpus para que se establezca la relación entre las lenguas (Nunzio y Ferro, 2005). Una de las principales desventajas de este recurso es que no siempre es posible encontrar documentos disponibles en todos los pares de lenguas, sobre todo si utilizamos lenguas no indoeuropeas. Además, es necesario realizar un correcto trabajo previo de recuperación y alineación si se desea un buen funcionamiento.

Desde hace relativamente poco tiempo, se están introduciendo los corpus multilingües tanto en los corpus comparables como en los paralelos. Se trata de conjuntos de dos o más corpus monolingües en diferentes idiomas, lo que permite establecer más fácilmente patrones entre dos o más lenguas (Baker, 1995). La ventaja de estos corpus es que permiten ahorrar espacio y tiempo en la gestión de la información utilizada.

1.3.2.3. DICCIONARIOS

Tradicionalmente se ha llamado diccionario a la recopilación de palabras, locuciones, giros y sintagmas de una lengua o, dentro de ella, los términos de una ciencia, técnica, arte, especialidad, etcétera, generalmente dispuestos en orden alfabético (Martínez de Sousa, 2009). Los avances tecnológicos y lexicográficos han dado lugar a nuevos tipos de formato y de organización de este recurso y, hoy en día, no solamente nos encontramos con diccionarios en formato papel sino que la mayoría de los creadores prefieren tener una versión electrónica del mismo.

Según la definición de Redondo Madrigal (1995), desde el punto de vista del idioma, entendemos los diccionarios como obras de referencia lexicográficas imprescindibles para toda persona interesada en las cuestiones de idioma, ya que ofrece un repertorio de palabras, modismos y expresiones para conocer y mejorar el uso de las lenguas; por añadidura, la revisión constante de que suelen ser objeto genera sucesivas ediciones que nos actualizan en el uso de las mismas. El interés mostrado por los investigadores en este recurso lingüístico ha influido en la profusión del mismo. Hoy en día, existen diccionarios de muy diversos tipos que se adecuan de mejor o peor manera a los SRI actuales.

Los diccionarios electrónicos son unos de los recursos más utilizados en los SBR disponibles, ya sean solos o acompañando a otro recurso o herramienta. Las ventajas que ofrecen son muchas y muy atractivas para los desarrolladores ya que hay muchos y de muy diversos tipos, lo que permite adaptarse muy bien a las características particulares de cada sistema; trabajan con pares de lenguas que son difíciles de trabajar con otros recursos; y hay muchos gratuitos en la red, lo que reduce substancialmente el coste económico. Además, aquellos investigadores que lo han incorporado a sus sistemas (Sorg y Cimiano, 2008) afirman que su método de trabajo es muy simple puesto que el sistema solamente tiene que seleccionar los términos que deben ser traducidos, el recurso provee la traducción de los mismos y se incorporan los términos al motor de búsqueda del sistema para que pueda recuperar la respuesta más adecuada a la consulta.

Sin embargo, este recurso no siempre puede satisfacer gratamente las necesidades de los SBR, ya que la cobertura de los diccionarios electrónicos no siempre es tan completa y ciertos términos especializados no se recogen. Además, los diccionarios no suelen presentar todos los derivados y compuestos de las palabras, ni nombres propios, lo que limita totalmente su producción (Peinado Herencia, 2004).

No obstante, el principal problema que atañe este recurso es sin duda la calidad en la traducción. La traducción palabra por palabra normalmente reproduce textos ininteligibles que afecta a la información mostrada al usuario, ya sea porque se le muestre ese mismo texto o porque el sistema recupere la consulta a través de él. Para solventar este problema, algunos investigadores han intentado combinar varios recursos (dos diccionarios o un diccionario y un

corpus), de modo que las traducciones no estuvieran tan limitadas y se solucionara el problema de la cobertura (Zhou *et al.*, 2007).

Los SBR translingües siempre intentan incorporar diccionarios automatizados (en inglés, *machine-readable dictionaries*), que permiten una incorporación sencilla al sistema y ofrecen una alta calidad (Hiemstra *et al.*, 2001). Este recurso se caracteriza por ofrecer cobertura para las principales lenguas europeas, lo que ha fomentado su uso entre aquellos sistemas que incorporan la traducción de las consultas. Dentro de este tipo de diccionarios, los diccionarios bilingües son los más comunes, aunque también se pueden encontrar casos de diccionarios multilingües *ad hoc*.

Los diccionarios bilingües se caracterizan por presentar dos equivalencias de un determinado término en dos idiomas determinados. Entre las ventajas de los bilingües se encuentra la aportación de ejemplos, dependiendo de las características del diccionario en sí.

Por último, cabe mencionar los diccionarios multilingües. La principal diferencia de estos diccionarios respecto a los anteriores es el número de lenguas, superior a tres idiomas. Las posibilidades que ofrecen estos diccionarios estriban en que pueden poseer varios diccionarios en uno, tanto en un área general como en una específica. Sobre todo en el ámbito especializado, son de gran utilidad en los SBR. Sin embargo, no existen muchos diccionarios multilingües disponibles en la Web y, por tanto, la mayoría son *ad hoc*. Aunque valoramos positivamente la creación de nuevos recursos lingüísticos aunque sean para situaciones muy específicas, hay que resaltar que estos diccionarios no suelen ser muy útiles para los usuarios, ya que presentan unas características muy específicas que le permite adecuarse al SBR y no suelen ser alimentados ni actualizados por los gestores del sistema.

1.3.2.4. ENCICLOPEDIAS

Las enciclopedias son compendios del conocimiento humano. La enciclopedia reúne y divulga datos especializados o dispersos que no podrían ser hallados con facilidad y que presentan un importante servicio a la cultura moderna. Las enciclopedias abarcan todas las ramas del saber de forma selectiva, escogiendo los temas más importantes de cada una de ellas, mientras que otras, más especializadas, se centran en una determinada materia estudiándola en profundidad.

El uso de las enciclopedias en los SBR no es muy frecuente, pero ha sido utilizado por algunos desarrolladores. Las ventajas que ofrecen son muy similares a otros recursos, sin embargo, el hecho de que no existan muchas disponibles en la Web hace que su uso sea menos habitual.

No obstante, un caso aparte es el uso de Wikipedia y Wiktionary. El uso de estas dos enciclopedias en línea, sobre todo en el caso de la primera, es totalmente diferente al resto de las enciclopedias. Su amplio compendio gratuito la convierte en uno de los recursos más útiles en los sistemas de RI, en general, y de los SBR, en particular.

1.1.3.1.1.4.1. WIKIPEDIA Y WIKTIONARY

La Wikipedia⁴⁵ es una enciclopedia libre multilingüe *online* y colaborativa con entradas para alrededor de 300 lenguas, de las cuales, el inglés es la más representativa con casi 3 millones de artículos (Otero y López, 2010). El rápido crecimiento de la Wikipedia en los últimos años ha propiciado la realización de numerosos trabajos que explotan este recurso para diferentes objetivos multilingües: extracción de diccionarios bilingües (Yu y Tsujii 2009; Tyers y Pienaar 2008), alineamiento y TA (Adafre y Rijke 2006), CLIR (Potthast *et al.* 2008; Olvera-Lobo y Gutiérrez-Artacho, 2011b), creación de corpus comparables (Filatova 2009; Otero y López, 2010).

Según su propia definición, Wikipedia es: “una enciclopedia, entendida como soporte para la recopilación, almacenamiento y transmisión de información estructurada. Una wiki que puede ser editada por cualquiera y de contenido libre”. Es un proyecto de contenido enciclopédico, creado por la comunidad *online* en todos los idiomas posibles. Cuando hablamos de enciclopedia libre, nos referimos a que su contenido no tiene restricciones legales de uso o distribución y que puede ser modificado (Olvera-Lobo *et al.*, en prensa). La auténtica Wikipedia está formada, no por una única Wiki, sino por gran número de ellas, diferenciadas e independientes según el idioma, formando un sistema multilingüe. Esto significa que, los contenidos, la gestión, los usuarios, etc., de cada Wikipedia son diferentes e independientes entre ellos.

Las ventajas de este recurso son muchas en CLIR, ya que el SBR disminuye sus costes tanto en la incorporación del recurso –es gratuita–, y en la gestión del mismo –no es necesario que se lleve a cabo por parte de los desarrolladores o gestores del sistema–. Wikipedia está siendo continuamente actualizada en todos los idiomas, los cuales se encuentran enlazados entre sí mediante hipervínculos que referencian a aspectos relacionados en el mismo idioma y al mismo artículo pero escrito en otros idiomas.

Sin embargo, uno de los principales problemas de Wikipedia es la calidad del contenido, la libre autoría de su contenido no da certeza de que la información sea veraz y concurrente. Además, no todas las lenguas tienen la misma información puesto que están escritas por personas diferentes dentro de culturas muy distintas. En muchas ocasiones, los textos son traducidos por los usuarios del inglés a otras lenguas de destino pero no es siempre así.

Los SBR translingües que incorporan este recurso se caracterizan por seguir unos patrones parecidos a los corpus comparables, por ello normalmente necesitan de otros recursos o herramientas para dar una cobertura amplia (TA, corpus, diccionarios).

Dentro de Wikipedia, se debe mencionar Wiktionary⁴⁶ que es un diccionario multilingüe con formato y objetivos muy parecidos a los de Wikipedia. Se puede definir como su proyecto hermano. Las principales diferencias entre estos dos recursos es que Wiktionary se centra en el léxico en lugar del contenido enciclopédico.

Aunque los contenidos de Wiktionary son más escasos que Wikipedia, éste cuenta con más de 500.000 entradas por lengua y está disponible en más de 300 idiomas. Al igual que Wikipedia, todos los contenidos se encuentran referenciados entre sí por medio de hiperenlaces a los contenidos del mismo idioma y a aquellos artículos escritos sobre el mismo tema en otros idiomas.

⁴⁵ <http://es.wikipedia.org/wiki/Wikipedia:Portada>

⁴⁶ <http://es.wiktionary.org/wiki/Wikcionario:Portada>

Las ventajas de Wiktionary son muy parecidas a las del recurso anterior, aunque presenta más limitaciones en cuanto a contenido y gestión. El uso de este recurso no se encuentra tan extendido como Wikipedia y sólo algunos investigadores han comenzado a experimentar con él. Los inconvenientes del recurso son que, al igual que Wikipedia, la calidad de su contenido no es del todo fiable puesto que cualquiera puede escribir y cambiar la información. Además, su efectividad no está del todo probada, solamente se ha incorporado en algunos sistemas y necesita de la ayuda de otras herramientas, como los traductores automáticos, para proporcionar una cobertura de información amplia.

1.3.2.5. EUROWORDNET

EuroWordNet, en adelante EWN, es un proyecto desarrollado por la Unión Europea, basado en la base de datos monolingüe WordNet. Este recurso lingüístico se creó en 1996 con la intención de desarrollar una versión multilingüe de WordNet con ciertos idiomas de la Unión Europea: italiano, inglés, holandés y español, y en 1997, se incluyeron el alemán, francés, checo y estonio (Vossen, 1998).

La base de EWN es WordNet, uno de los primeros intentos en el desarrollo de un lexicón multipropósito en forma de aplicación informática. La importancia de WordNet en los SBR no radica en el método de traducción, sino en la forma de gestionar la información. WordNet es un sistema electrónico de referencia léxica, desarrollado en forma de base de datos léxica por George A. Miller en 1985 (Miller, 1986). WordNet constituye un intento de reflejar el modelo de memoria léxica basado en redes semánticas en un modelo lexicográfico de organización léxica (Moreno Ortiz, 2000).

Una de las ventajas en el desarrollo de esta base de datos léxica fue el apoyo recibido por numerosas instituciones públicas y privadas norteamericanas. En el caso de EWN, se intentó desarrollar una base de datos similar multilingüe con apoyo institucional europeo, sin embargo, el apoyo recibido por las instituciones públicas y privadas de los diferentes países europeos no fue tan amplio. Los objetivos primordiales de WordNet son la validación de la organización léxica y su previsible utilización en diversas aplicaciones que requieran acceso a información léxica. WordNet divide el lexicón en cinco categorías: nombres, verbos, adjetivos, adverbios y elementos funcionales, que facilitan el análisis de las diferentes organizaciones semánticas. La fundamentación teórica del sistema tiene su origen en la idea de la “matriz de vocabulario” (*vocabulary matrix*) (Miller, 1986; Miller *et al.*, 1990), en el término forma léxica para referirse a la expresión física que se escribe o se pronuncia, y en el significado léxico para referirse al concepto lexicalizado que se expresa por medio de una forma léxica.

WordNet está organizado en base a relaciones semánticas, es decir, relaciones de significados, y los significados están representados por medio de *synsets*, conjunto de sinónimos o significados, proporcionando definiciones cortas y generales, y almacenando las relaciones semánticas entre estos conjuntos de sinónimos. El propósito del proyecto es doble: por un lado producir una combinación de diccionario y tesoro cuyo uso es más intuitivo, y ayudar al análisis automático de textos y a las aplicaciones de inteligencia artificial.

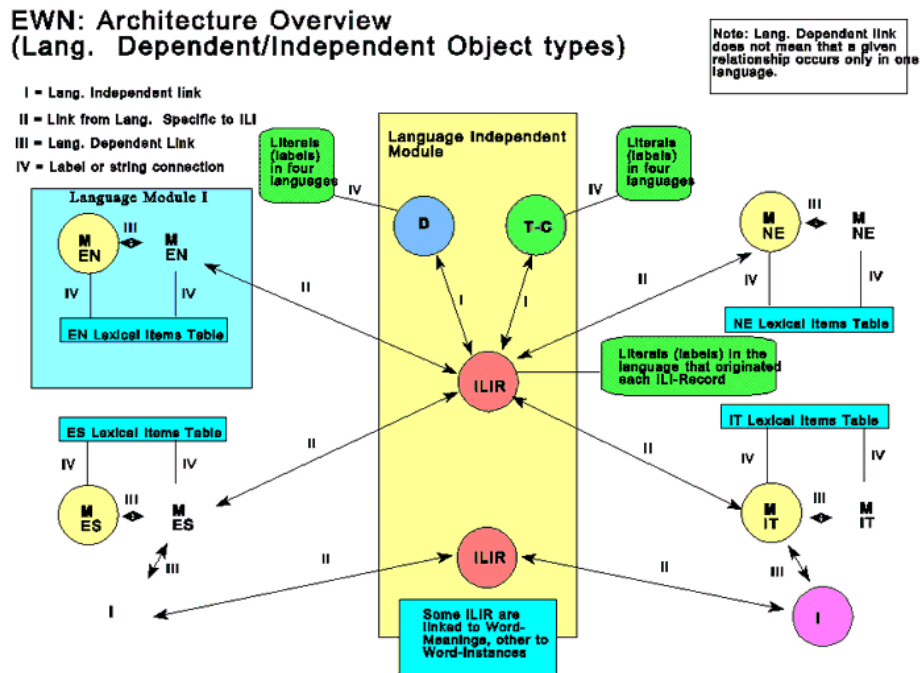


Figura 13. Arquitectura de EuroWordNet (Díez Orzas, 1999)

La principal ventaja de EWN es que sigue la arquitectura creada en WordNet, es decir, una serie de relaciones entre *synset* (*synonym sets*) dentro de la misma lengua, y se relacionan entre sí por medio de un índice interlingüístico, denominado *Inter-Lingual Index* (ILI) (Vossen 1998). La incorporación de EWN a los SBR multilingües ha sido posible gracias a esta arquitectura flexible, que permite un funcionamiento correcto con otras ontologías o bases de conocimiento léxico. Además, ILI disminuye el trabajo de los desarrolladores en los procesos de indexación del material.

Aunque no es gratuito, varios autores han incorporado EWN en sus prototipos de SBR. Las primeras aproximaciones con este recurso no han sido del todo satisfactorias, ya que aunque se consiguen mejores resultados que con otros recursos, se necesita un amplio trabajo de fondo en la recopilación e indexación del material y no se ha solucionado completamente el problema de la ambigüedad.

1.1.3.1.1.6. ONTOLOGÍAS

Las ontologías son especificaciones formales y explícitas de conceptualización compartida (Gruber, 1993). *Conceptualización* hace referencia a un modelo abstracto de algún fenómeno en el mundo del que se han identificado conceptos relevantes o una abstracción de las entidades asumidas dentro de un área y las relaciones que se establecen entre esas entidades. *Explícito* se refiere a la necesidad de especificar explícitamente los tipos de conceptos y las limitaciones en su uso. *Formal* hace referencia al hecho de que una ontología debería ser representada en un lenguaje formalizado que pueda ser comprensible por todos los sistemas. Finalmente, *compartida*

refleja la idea de que una ontología debe representar y dar cuenta del conocimiento aceptado (como mínimo, por el grupo de personas que deben usarla).

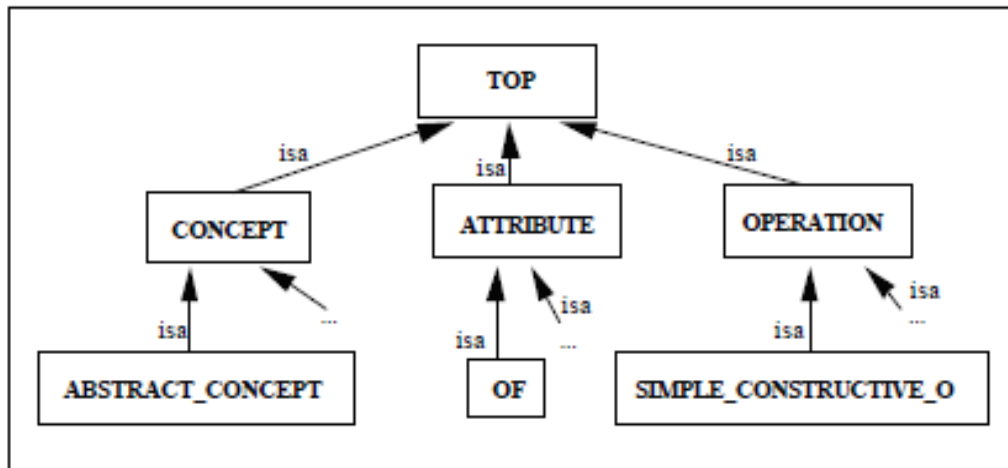


Figura 14. Representación de una ontología conceptual (Gatius, 2000)

Una de las principales ventajas de las ontologías es que codifican el conocimiento general o de un dominio determinado mediante las relaciones específicas de sus conceptos, de modo que su incorporación en los SBR multilingües permite desambiguar la representación del conocimiento interno de la misma, y establecer un mismo lenguaje entre las preguntas y el conocimiento del dominio (Diekema, 2003). Los investigadores que han usado este recurso confirman que la ventaja de organizar el conocimiento conceptual de esta manera radica en que todos los objetos, conceptos, relaciones y entidades que aparecen en el dominio se encuentran totalmente definidos, lo que fomenta la claridad, consistencia, accesibilidad, extensibilidad y reusabilidad del mismo (Gatius, 2000).

Estos recursos, principalmente las ontologías formadas por dos o más lenguas, son incorporadas a los SBR multilingües mediante un proceso de localización que consiste en adaptar la ontología a las particularidades del sistema.

1.1.3.1.1.7. PÁGINAS WEB

El uso de páginas web o sitios web en los SBR es muy frecuente, sobre todo en aquellos sistemas especializados en un ámbito. Se denomina página web al documento o información electrónica adaptada para la *World Wide Web* que generalmente forma parte de un sitio web. Este recurso se caracteriza por estar compuesto por información (en texto o módulos multimedia) así como de hiperenlaces o hipervínculos que nos asocian los datos relacionados.

Las características de las páginas web se adaptan perfectamente a aquellos de los SBR, ya que ambos están escritos en un lenguaje de marcado, que incluso pueden compartir, y tienen una estructura normalizada que facilita y simplifica la visualización e incorporación del contenido por parte de los SBR. Además, una de las tendencias actuales en la creación de páginas web no es

solamente que sean atractivas y fáciles de manejar para los usuarios, sino que también sean optimizadas para buscadores u otros sistemas informáticos a través de su código fuente.

La incorporación de este recurso lingüístico es bastante común en aquellos SBR multilingües especializados en un área determinada y cuyas lenguas de funcionamiento no sean muy comunes en la red. Normalmente se suelen incorporar junto con otro recurso lingüístico, como los diccionarios, el cual se encarga de traducir principalmente las palabras de la consulta planteada por los usuarios.

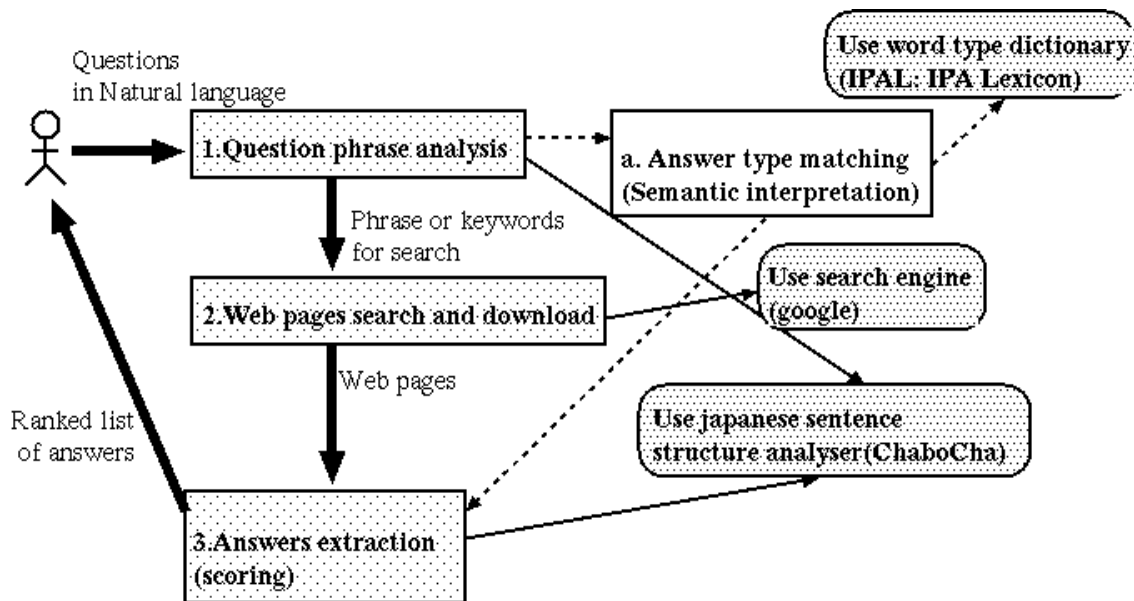


Figura 15. Ejemplo de un SBR cuya fuente se documenta en páginas web (Ramamonjisoa, 2003)

1.1.3.1.1.8. TESAUROS

Los tesauros se caracterizan por ser un tipo de lenguaje documental formado por un vocabulario controlado dotado de relaciones que son las que les distinguen de otros instrumentos terminológicos y que le proporcionan la categoría de lenguaje. Constituyen un instrumento terminológico que da coherencia al proceso de indexación y RI, ya que permite que la búsqueda coincida con la información ya indexada y almacenada. Este recurso junto con las ontologías suele estar diseñado específicamente para sustentar la RI (Diekema, 2003).

Este último recurso no tiene una popularidad muy extendida, sin embargo, está comenzando a ser bastante utilizado por varios SBR multilingües, ya que permite indexar la colección de documentos del sistema atendiendo a las diversas relaciones semánticas de los términos o expandiendo y enriqueciendo las consultas con términos relacionados (Soergel, 1997). Además, las relaciones establecidas por los tesauros, las cuales están muy bien definidas y suelen ser jerárquicas, ayudan a los usuarios a definir su consulta más detalladamente, de modo que pueden recuperar la información que se adecue mejor a sus necesidades.

Un tipo de tesauros que incluyen estos sistemas son los tesauros de similitud. La gran particularidad de estos tesauros es que se construyen automáticamente a partir de las relaciones de proximidad temática establecidas entre los términos de la colección. Se puede decir que son parecidos al funcionamiento de los corpus comparables pero indexados. Las ventajas que ofrecen a los SBR son varias, ya que por un lado, su funcionamiento ayuda a mejorar las búsquedas en el sistema, y al ser automáticos facilita el trabajo a los desarrolladores.

Los tesauros multilingües son el segundo tipo. Se caracterizan por ser un vocabulario controlado y flexible en donde se acepta la incorporación de nuevos idiomas. Los diferentes idiomas se encuentran relacionados entre sí gracias a la indexación del sistema. El uso de este recurso por parte de SBR translingües facilita CLIR sin necesitar la combinación con otro recurso o herramientas y con la información ordenada y preparada para su recuperación. Sin embargo, no existen muchos tesauros multilingües en el mercado que puedan ser incorporados a esos sistemas, excepto algunos de ámbitos muy especializados como la medicina. No obstante, algunos sistemas están haciendo uso del tesoro multilingüe del Parlamento Europeo EUROVOC, el cual fue creado a partir de la traducción de un tesoro monolingüe existente. Este tesoro está disponible en las principales lenguas de la Unión Europea, pero no se encuentra lo suficientemente desarrollado como para ofrecer una cobertura amplia a los usuarios. Lo que sí que está teniendo mucha aceptación son las técnicas de vocabulario controlado basado en tesauros multilingües, que consisten en representar los términos en varios idiomas de modo que cualquiera de estos idiomas pueda ser usado para indexar los documentos o las consultas formuladas (Diekema, 2003).

1.1.3.1.2. HERRAMIENTAS LINGÜÍSTICAS

Además de los recursos lingüísticos, los SBR multilingües hacen uso de ciertas herramientas lingüísticas que les permite resolver el problema de la comunicación translingüe. Se denominan herramientas lingüísticas al conjunto de software o programas que permiten la manipulación o gestión de recursos lingüísticos. La gran diferencia entre las herramientas y los recursos lingüísticos es que los primeros no cuentan con material léxico disponible sino que son meros gestores de información.

Las dos herramientas principales que usan estos sistemas son las gramáticas computacionales (apartado 1.1.3.1.2.1.) y los traductores automáticos (apartado 1.1.3.1.2.2.).

1.1.3.1.2.1. GRAMÁTICAS COMPUTACIONALES

La primera herramienta utilizada por los SBR multilingües son las gramáticas computacionales. Aunque no es tan utilizada como los traductores automáticos, su popularidad está creciendo cada vez más. Las gramáticas computacionales son una descripción formalizada del conocimiento lingüístico, es decir, son herramientas que contienen cualquier tipo de información lingüística y que se puede recuperar con ciertos formalismos (Llisterra, 2008). El autor distingue entre tres tipos de gramáticas: gramáticas de unificación, gramáticas de restricciones y sintaxis léxicas.

Las principales ventajas de las gramáticas computacionales son la fácil incorporación de esta herramienta y la buena combinación con otros recursos lingüísticos, lo que permite dar una buena cobertura lingüística sin mucho coste computacional. Además, estas gramáticas facilitan la incorporación de muchos idiomas gracias a una arquitectura muy flexible. No obstante, y aunque los investigadores confirman su eficacia y buen funcionamiento, los resultados actuales no son muy concluyentes ya que no se trata de una herramienta muy extendida y existen muy pocos prototipos.

Muchos de los SBR multilingües que han incorporado esta herramienta suelen introducir el sistema LKB (*Linguistic Knowledge Building*), cuyo formato y aplicación se adecua mejor a las necesidades de estos sistemas. Sin embargo, la escasa cobertura de vocabulario específico y de nombres propios provoca problemas de ambigüedad.

1.1.3.1.2.2. TRADUCTORES AUTOMÁTICOS

El crecimiento del mercado de la traducción en todo el mundo ha llevado a algunos países desarrollados a investigar e implementar programas de TA como una posible solución al aumento de la demanda de traducciones.

En CLIR, la herramienta estrella son los traductores automáticos. A pesar de sus limitaciones, la fácil y rápida solución que ofrecen a los desarrolladores les convierte en la herramienta más demandada, ya sea como primera opción o mediante la combinación con otros recursos.

Los traductores automáticos son sistemas informáticos que pueden llevar a cabo traducciones de una lengua a otra con o sin la ayuda de la intervención humana (López, 2002). Esta herramienta funciona introduciendo el texto deseado en la interfaz del programa, indicando la lengua de origen y seleccionando la lengua a la que queremos traducir el correspondiente texto. Tras ello, el programa nos facilitará la traducción del texto en el idioma elegido.

Hoy en día, existe un amplio abanico de traductores automáticos con diferentes características, algunos de ellos son gratuitos y otros comerciales, unos solamente trabajan con un par de lenguas y otros son plurilingües, algunos ofrecen muy buenas traducciones y otros muestran textos sin sentido.

Hay dos subtipos importantes en los métodos utilizados por la TA: la traducción basada en reglas y la traducción basada en analogías. En la TA basada en reglas, el programa analiza el léxico, la morfología, la sintaxis y otros aspectos lingüísticos del texto original, crea una representación abstracta de su sentido y genera a partir de ella el texto traducido. Uno de los ejemplos más reconocidos es *Systran*⁴⁷.

La proliferación de colecciones de textos en formato electrónico y su disponibilidad favorecida por Internet han favorecido probar métodos probabilísticos y conexionistas que tan buenos resultados habían dado ya en el tratamiento de corpora orales. Todo ello ha fomentado que el paradigma de la traducción por reglas comienza a perder adeptos en beneficio de la traducción por analogías (Abaitua, 2002). En la TA basada en analogías, el programa fragmenta el texto original y traduce los fragmentos por analogía con traducciones previas realizadas por humanos.

⁴⁷ Disponible en: <http://www.systran.es/>

El traductor de Google⁴⁸ es un claro ejemplo de un sistema basado en analogías. Desde hace aproximadamente diez años, Google, fiel a su misión de organizar la información del mundo, ha puesto la TA en manos de cualquier persona con acceso a Internet, en una gama de idiomas casi inimaginable.

La evolución de los traductores automáticos en los últimos años, así como la creciente demanda de este producto a nivel mundial, han influido en el desarrollo y creación de una serie de sofisticados TA disponibles en Internet (Hutchins, 2002). Hoy en día podemos encontrar todo tipo de TA en la red que facilitan la traducción de páginas web y otros documentos en línea. Los sistemas de traducción en línea que se pueden consultar en numerosos sitios web pueden prestar un buen servicio cuando el usuario quiere entender la idea general de una página web que está escrita en otro idioma (Diéguez y Lazo, 2004).

La aparición y la correspondiente evolución de estas herramientas lingüísticas han sido tan importante desde los años 90 que se ha visto reflejado en la creación de un *think tank* específico en esta área, denominado TAUS⁴⁹. Esta asociación compuesta por investigadores y empresas dedicadas a este sector trata de aumentar el tamaño e importancia de la industria de la traducción con el objetivo de mejorar las comunicaciones internacionales. Desde 2005 TAUS comienza a reunir a los líderes en la industria de TA con vistas a la promoción de un mayor conocimiento de la tecnología en el campo de la traducción, una industria que históricamente ha brillado por la ausencia de la mecanización en el trabajo. Para ello, esta asociación desarrolla investigaciones que se publican en su semanario virtual y en los múltiples congresos que desarrollan. Los objetivos de TAUS se ven reflejados en el siguiente gráfico (Figura 16), creado por la propia asociación.

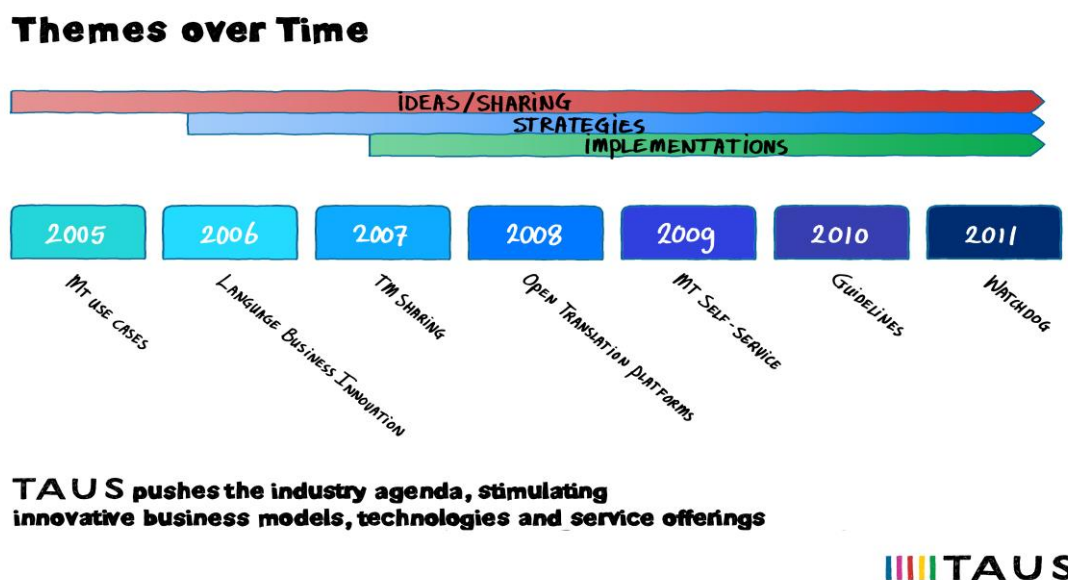


Figura 16. Objetivos de TAUS

⁴⁸ Disponible en: <https://translate.google.es/>

⁴⁹ Disponible en <http://www.translationautomation.com/>

Además, dentro del desarrollo y evolución de los TA, también han jugado un papel muy importante tanto las empresas dedicadas exclusivamente al sector de la traducción, como aquellas que se dedican a otros sectores relacionados. Numerosas empresas dedicadas al sector de la traducción han creado sus propios sistemas de TA, que junto con sus servicios de traducción tradicional, satisfacen las necesidades de sus clientes. Una de estas empresas es SDL⁵⁰, una de las empresas líderes en gestión de la información multilingüe. SDL se caracteriza por ser una empresa a la vanguardia de las nuevas herramientas tecnológicas disponibles para la traducción, de hecho, se trata de la empresa que comercializa la herramienta de traducción asistida más popular en este sector, SDL TRADOS⁵¹. La importancia de los traductores automáticos ha favorecido la creación por parte de SDL, junto con Click2translate⁵², de su herramienta Freetranslation⁵³, que se trata de un sitio web de TA desarrollado por una empresa del sector. Esta herramienta tiene sus ventajas y desventajas, ya que al ver sido creada por una empresa de la industria de la traducción la fiabilidad para los usuarios es bastante alta, sin embargo, no podemos destacarla como una de las mejores ya que empresas dedicadas a otros sectores invierten más dinero en el desarrollo de sus TA. Una de esas empresas es IBM, la cual para mejorar la comunicación con los clientes que hablan diferentes idiomas y también entre sus empleados y socios, ha iniciado el proyecto “n.Fluent” dedicado a la búsqueda de soluciones lingüísticas profesionales. Para ello colaboran activamente con otra importante empresa dedicada a la industria de la traducción, Lionbridge⁵⁴.

La TA, en general, ofrece muchas ventajas ya que resuelve el problema de CLIR a bajo coste y mediante una solución rápida y fácil. Hoy en día, nos encontramos en Internet con un número alto de traductores automáticos gratuitos o a muy bajo coste, que permiten su incorporación a los SBR multilingües sin la necesidad de mucha intervención de los desarrolladores. En las últimas ediciones de los foros más relevantes en CLIR, más del 50% de los investigadores se decantan por el uso de estas herramientas. Aunque la calidad obtenida en las traducciones ofrecidas suele ser bastante deficiente en la mayoría de ellos, no es el inconveniente que más preocupa a los investigadores puesto que les solucionan en gran medida el problema de la comunicación translingüe. Sin embargo, el uso de esta herramienta requiere la combinación de otras fuentes léxicas para poder ofrecer una cobertura amplia, como son los diccionarios automatizados, las ontologías, etc.

⁵⁰ <http://www.sdl.com/es/>

⁵¹ <http://www.trados.com/en/>

⁵² <http://www.click2translate.com/>

⁵³ <http://traducciongratis.freetranslation.com/>

⁵⁴ <http://es-es.lionbridge.com/Default.aspx?LangType=1034>

1.2. JUSTIFICACIÓN

Actualmente el nivel de desarrollo y de eficacia en el funcionamiento de los SBR puede calificarse de bastante desigual. Mientras que algunos sistemas despliegan un comportamiento más que aceptable, otros se desempeñan con dificultad, presentan un deficiente funcionamiento o, incluso, han dejado de estar operativos. Por esta razón, se hace necesario someterlos a diferentes pruebas que permitan el testeo y la evaluación rigurosa de los mismos considerando diferentes aspectos. Además, el auge de sistemas CLIR y, por ende, la incorporación de técnicas de traducción en los SBR exige un estudio riguroso de los recursos y herramientas lingüísticos aplicados. Solo así se puede determinar objetivamente cuáles son sus puntos fuertes y sus flaquezas y, por tanto, establecer dónde necesitan introducir mejoras.

Como se ha indicado, los SBR se diferencian de otros SRI por ofrecer respuestas cortas, palabras o números a las consultas planteadas. Esta peculiaridad obliga a evaluar a estos sistemas con nuevas medidas que se adapten de forma más adecuada a su arquitectura y a las expectativas de los usuarios.

El objetivo primordial de este trabajo ha consistido en diseñar y desarrollar una propuesta metodológica válida para analizar y evaluar los SBR monolingües, multilingües y translingües con el fin de identificar los recursos y herramientas que implementan. Este objetivo se encamina a la constatación de que algunas de las técnicas, medidas y herramientas de traducción y evaluación usadas tradicionalmente en CLIR, puesto que están, conceptualmente, bien fundamentadas, son adaptables y aplicables en el contexto de los SBR, aunque su arquitectura y tipología documental sea muy diferente a la de los SRI clásicos. Por otra parte, sus características también exigen que nuevas perspectivas y propuestas metodológicas se tengan en cuenta con el fin de llevar a cabo evaluaciones rigurosas y que respondan a las expectativas de los usuarios. Por tanto, debido a que, en Internet, tanto los SRI disponibles como la propia información de la que se nutren, son muy dinámicos, la finalidad principal de estos trabajos no ha sido la de determinar qué recurso lingüístico es mejor que otro como técnica de traducción, o cuál es el mejor SBR actualmente presente en la red, puesto que las conclusiones a las que se pueden llegar son poco perdurables, sino desarrollar un método que pueda aplicarse al análisis y la evaluación de los resultados por estos ofrecidos –y a la medición de la satisfacción de sus usuarios– permitiendo así la realización de estudios periódicos que determinen y analicen la evolución de estas herramientas de búsqueda tan apreciadas por los usuarios.

En la primera parte de este capítulo introductorio se ha presentado el estado de la cuestión respecto a los SBR monolingües y multilingües, y se han descrito en detalle nuestra clasificación de todos los recursos lingüísticos y herramientas de los que se valen estos sistemas. En esta sección, dedicada a la justificación del trabajo, se describe en primer lugar la motivación, hipótesis y objetivos que han fundamentado esta tesis doctoral, así como la metodología adoptada para realizar los ocho estudios presentados.

1.2.1. MOTIVACIÓN, HIPÓTESIS Y OBJETIVOS

La evolución de la RI en estos últimos años ha sido vertiginosa, desarrollándose una variedad de sistemas que abarcan todos los problemas actuales en recuperación. Uno de los problemas que trasciende a casi todos los ámbitos es la recuperación multilingüe, ya que como se ha mencionado una audiencia cada vez más heterogénea y un usuario más exigente fuerzan la creación de SRI que, no solo satisfagan las necesidades de búsqueda, sino que interactúen con ellos en su propio idioma.

Ante esta situación surge nuestra motivación para realizar el presente trabajo. Por una parte, se ha querido identificar qué recursos o herramientas lingüísticos son los más útiles para llevar a cabo la recuperación multilingüe y translingüe centrandó nuestro foco en aquellos recursos susceptibles de ser incorporados a los SBR. Como ya se ha mencionado, los investigadores en RI subrayan la importancia de la traducción en sus trabajos, pero obvian que los problemas ocasionados por no utilizar una herramienta de traducción adecuada, por ende, relegan el problema de CLIR. Los estudios realizados analizan y evalúan unos determinados SRI, los SBR monolingües y multilingües, desde diferentes perspectivas con la intención de proporcionar una visión más próxima de la relación entre la traducción y CLIR.

Por ello, al inicio de este trabajo se han formulado las siguientes hipótesis:

1. El auge en la creación de nuevos prototipos que superen las barreras lingüísticas nos muestra la importancia creciente de los SBR multilingües en la RI, y la preocupación de los desarrolladores, investigadores y empresas por llevar a cabo herramientas que tengan buena acogida en el mercado. Cualquier avance que se realice en la resolución de problemas de la comunicación multilingüe podrá ser incorporado a los sistemas de recuperación existentes.
2. La evaluación real del funcionamiento de los SBR monolingües y multilingües desde diferentes perspectivas y disciplinas proporciona un *feedback* de la situación real.
3. El uso de los SBR multilingües disminuye el tiempo de obtención de la respuesta y diversifica las posibilidades de que el usuario obtenga la respuesta adecuada.
4. En el novedoso ámbito de los SBR multilingües no todas las herramientas lingüísticas afectan igualmente a la eficacia de su funcionamiento.
5. Los investigadores, con nuestros estudios, podemos ofrecer una visión más o menos clara de la situación de estos sistemas, pero serán los usuarios finales con sus valoraciones los que expresen sus necesidades reales y los fallos del sistema.

Esta investigación se plantea como objetivo general proporcionar una perspectiva global y real del problema derivado del uso de herramientas y recursos lingüísticos susceptibles de ser utilizados en un sistema para el tratamiento y traducción de los textos en CLIR, centrandó nuestras investigaciones en el funcionamiento de los SBR monolingües y multilingües.

Para su consecución se han planteado cuatro objetivos específicos:

1. Analizar y evaluar los recursos y herramientas lingüísticos utilizados en estos sistemas. Para ello, se han identificado los principales recursos usados en los procesos de CLIR, y se comprobará el grado de utilización real que hacen los SBR de cada uno de los recursos y herramientas analizados.

2. Identificar todas las medidas de evaluación utilizadas durante los quince años de vida de los SBR para comprender mejor las características de estos sistemas, y comprobar qué medidas son más aplicadas en los procesos de evaluación.
3. Analizar y evaluar el funcionamiento real de los SBR disponibles en la Web. Para ello, se ha establecido dos categorías diferenciadas: la especialización o no del sistema (o sea, si es de dominio general o dominio especializado), y el idioma o idiomas disponibles en los que se puede recuperar la información, es decir, si es monolingüe, multilingüe o translingüe.
4. Analizar la satisfacción de los usuarios de estos sistemas como herramienta terminológica. La opinión y valoración de los usuarios finales sobre los SBR multilingües es tan importante como la evaluación de las respuestas recuperadas. Los estudios de usuarios permiten saber si los sistemas no solo cumplen con lo establecido, sino si también satisfacen sus necesidades.

1.2.2. METODOLOGÍA

Para refutar las hipótesis planteadas y cumplir con los objetivos propuestos, se han planteado una serie de estudios que se recogen siete artículos de investigación publicados en revistas internacionales con revisión por partes y un capítulo de un libro. En la siguiente tabla, se resume los aspectos tratados en los análisis realizados.

A1: Análisis orientado a los recursos y herramientas lingüísticos en los SBR multilingües		
Publicación 1: Análisis de los recursos lingüísticos utilizados en CLEF	Publicación 2: análisis general de recursos lingüísticos usados en los SBR	
A2: Evaluación orientada al sistema		
Publicación 3: identificación y análisis de las medidas de evaluación usadas para el testeo de los SBR		
Publicación 4: evaluación de los SBR como eficientes fuentes de información terminológica	Publicación 5: Evaluación de los SBR de dominio general frente a los de dominio especializado en el ámbito biomédico	Publicación 6: evaluación de la eficacia del funcionamiento de los SBR de dominio general
Publicación 7: evaluación final. La evaluación objetiva y subjetiva de la búsqueda multilingüe de respuestas		
A3: Evaluación orientada al usuario: satisfacción de usuarios en SBR multilingües (publicación 8)		

Tabla 5: Tabla resumen de los contenidos de las publicaciones realizadas

El primer análisis consta de dos estudios centrados en el análisis y evaluación de los recursos y herramientas lingüísticos utilizados por los SBR multilingües para solucionar el problema de la traducción. El primer estudio (*publicación 1*) se centra exclusivamente en las publicaciones presentadas en *Conference and Labs of Evaluation Forum*, ya que constituye la conferencia que más ha apostado por este tipo de sistemas y cada año celebra un *track* dedicado exclusivamente a este ámbito. El segundo estudio (*publicación 2*) amplía el número de contribuciones realizadas en este campo analizando todas las conferencias, congresos y foros existentes en el terreno de la RI.

El segundo análisis, denominado “Evaluación orientada al sistema”, consta de tres objetivos que están relacionados entre sí. En primer lugar, se han identificado todas las medidas utilizadas en las *QA tracks* de las tres principales campañas de evaluación de los SBR: *Text Retrieval Conference (TREC)*, *Conference and Labs of Evaluation Forum (CLEF)* y *NTCIR (publicación 3) Conference*. Se ha identificado los diferentes *tasks* o *labs* específicos creados en cada *QA track*, el tipo de preguntas de evaluación utilizadas, así como las medidas de evaluación utilizadas en las diferentes competiciones analizadas. Esta primera parte del análisis ha sido imprescindible para detectar las medidas de evaluación más recurrentes.

La segunda parte de esta evaluación consta de tres test (*publicación 4*, *publicación 5* y *publicación 6*) donde se analizan, evalúan y comparan los SBR monolingües y multilingües, generales y especializados, disponibles en la Web; así como de un cuarto análisis, donde se ha realizado una evaluación final (*publicación 7*) que aúna todos los objetivos anteriores. Este consta de una evaluación objetiva, ya que analiza y evalúa las respuestas ofrecidas por los SBR multilingües, y una evaluación subjetiva, ya que incorpora los resultados de los anteriores estudios y analiza las fuentes de información de las que se extraen las respuestas.

Finalmente, se ha realizado un cuarto análisis (*publicación 8*) que se centra en la evaluación orientada a las necesidades de los usuarios. La opinión y valoración de los usuarios finales sobre los SBR multilingües es tan importante como la evaluación de las respuestas recuperadas. Los estudios de usuarios permiten saber si los sistemas no solo cumplen con lo establecido, sino si también satisfacen sus necesidades.

Esta serie de pruebas, con una perspectiva tanto cuantitativa como cualitativa, ha permitido evaluar el funcionamiento y la calidad de los SBR monolingües y multilingües disponibles en la Web así como aquellos que, aun habiendo sido desarrollados, no se han puesto todavía a disposición de los usuarios finales.

Para llevar a cabo el análisis y la evaluación de los procesos de búsqueda de respuesta nos hemos servido tanto de los diferentes SBR disponibles en la Web, como de aquellos SBR presentados por los desarrolladores en los diferentes congresos y foros internacionales dedicados a este tema. Como paso previo, para poder analizar estos sistemas, ha sido necesario construir dos colecciones de preguntas de evaluación (véase Tabla 6) que permitieran analizar el funcionamiento real de los mismos.

C1	Publicación 4, 5 y 7
C2	Publicación 6

Tabla 6: Colecciones utilizadas en los test realizados

La primera colección de preguntas de evaluación es de dominio especializado. Se creó una muestra de preguntas médicas que permitieran evaluar el funcionamiento y las fuentes de información utilizadas por los SBR analizados. Las preguntas se obtuvieron del sitio web WebMD⁵⁵, un portal estadounidense creado por especialistas médicos para dar respuesta a las incertidumbres de los pacientes y en el que se ofrece información, en forma de preguntas y respuestas, sobre una amplia lista de temas médicos de diferente grado de especialización, de los que se explican brevemente, en su caso, las características de esa enfermedad, tratamiento o síndrome, entre otros. La colección de preguntas se ha obtenido tras plantear la expresión “*What*

⁵⁵ Disponible en: <http://www.webmd.com/>

is” en el motor interno de búsqueda del sitio web, ante la cual WebMD proporcionó más de 6000 preguntas.

Esta colección ha sido utilizada en las publicaciones 4, 5 y 7. La elección de preguntas de ámbito biomédico se debe al hecho de que han proliferado los SBR especializados en esta área del saber. Por tanto, en las evaluaciones se ha querido comprobar la eficacia de los sistemas especializados en ese dominio temático.

La colección 2, que incluye preguntas factuales y de definición en lengua inglesa, se creó a partir de las colecciones de preguntas de evaluación generadas por las principales conferencias sobre recuperación de información (*Text Retrieval Conference* y *Cross-Language Evaluation Forum*). Estas colecciones de preguntas se generan para que los participantes en estos foros las utilicen al llevar a cabo la evaluación de sus sistemas y los resultados obtenidos puedan compararse con los de los demás. A partir de las listas de preguntas disponibles en estos dos foros durante los años 2000 a 2004, se obtuvo una serie (véase Anexo 1) de casi 2000 preguntas de definición, factuales y de lista, y que versaban sobre diferentes especialidades (véanse tablas 7 a 9).

Lugar de extracción			
	CLEF	TREC	Total
Nº	597	1383	1980

Tabla 7. Origen de la colección de preguntas de evaluación

Año						
	2000	2001	2002	2003	2004	Total
Nº	730	475	100	475	200	1980

Tabla 8. Preguntas extraídas por años

Tema	
Animales	23
Arte y Literatura	193
Biología	44
Ciencia	184
Cine	41
Demografía	43
Deportes	91
Economía	156
General	324
Geografía	160
Historia	255
Medicina	86
Música	35
Orografía	18
Personajes	219
Política	108
Total	1980

Tabla 9. Preguntas según la especialidad temática

La metodología seguida en esta investigación comprende diversas perspectivas para la evaluación de los SBR monolingües y multilingües. Hasta hace relativamente poco, los únicos trabajos disponibles sobre los SBR, ya fueran monolingües o multilingües, eran excesivamente “técnicos” y las evaluaciones realizadas se debían meramente a intenciones particulares, es decir, un investigador analizaba, evaluaba y/o comparaba sistemas para mostrar las ventajas o desventajas que proporcionaba la elección de las características del suyo propio.

1.2.2.1. ANÁLISIS ORIENTADO A LOS RECURSOS Y HERRAMIENTAS LINGÜÍSTICOS EN LOS SBR MULTILINGÜES

El primer análisis que hemos realizado cumple con el objetivo general y el primer objetivo específico planteados en esta tesis doctoral. Se ha llevado a cabo una investigación basada en la observación documental con el fin de identificar y analizar la situación actual de los SBR multilingües, y de evaluar la relación existente entre los problemas de recuperación y la implementación de ciertos recursos lingüísticos para solventar el problema de la traducción. Para ello, se ha adoptado una metodología que permitió recopilar y estudiar la información ofrecida acerca de cada una de las diferentes herramientas y recursos lingüísticos utilizados por estos sistemas, así como sobre su uso e implementación.

Los siete niveles lingüísticos establecidos para el PLN (Liddy 1998; 2003) –el fonético, el morfológico, el léxico, el sintáctico, el semántico, el discursivo y el pragmático– resultan totalmente válidos y aplicables a los SBR (tanto para el tratamiento de las preguntas que constituyen la entrada, como para los documentos de los que se extrae la respuesta). Este ha sido el punto de partida para la recogida y estudio de la información sobre las herramientas y recursos lingüísticos utilizados por los SBR, si bien las características individuales y específicas de estos sistemas obligan a incidir especialmente en unos niveles de tratamiento lingüístico más que en otros.

A sabiendas de que la mayor parte de las investigaciones sobre estas cuestiones se estaban realizando en el contexto de determinados foros y congresos internacionales, se restringió la observación documental a esas aportaciones. Conscientes de que no se trata de un estudio exhaustivo al respecto, no es menos cierto que es absolutamente representativo e ilustrativo de la evolución experimentada por este tema en los últimos años. Como muestra, decir que, a finales de 2011, si se hacía una búsqueda en la base de datos *Web of Science* con el *topic “question-answering system”* únicamente se obtenían 414 resultados (con solo 149 artículos, parte de los cuales, a su vez también eran contribuciones a congresos), distribuidos como indica la Tabla 10.

<input type="checkbox"/> 2009 (68)	<input type="checkbox"/> 2010 (25)	<input type="checkbox"/> 1973 (2)	<input type="checkbox"/> 1971 (1)
<input type="checkbox"/> 2007 (61)	<input type="checkbox"/> 2002 (21)	<input type="checkbox"/> 1982 (2)	<input type="checkbox"/> 1976 (1)
<input type="checkbox"/> 2008 (61)	<input type="checkbox"/> 2001 (7)	<input type="checkbox"/> 1992 (2)	<input type="checkbox"/> 1980 (1)
<input type="checkbox"/> 2006 (42)	<input type="checkbox"/> 2000 (5)	<input type="checkbox"/> 1995 (2)	<input type="checkbox"/> 1991 (1)
<input type="checkbox"/> 2005 (37)	<input type="checkbox"/> 2011 (4)	<input type="checkbox"/> 1998 (2)	<input type="checkbox"/> 1993 (1)
<input type="checkbox"/> 2003 (30)	<input type="checkbox"/> 1978 (3)	<input type="checkbox"/> 1964 (1)	<input type="checkbox"/> 1994 (1)
<input type="checkbox"/> 2004 (28)	<input type="checkbox"/> 1979 (3)	<input type="checkbox"/> 1966 (1)	<input type="checkbox"/> 1997 (1)

PROCEEDINGS PAPER (337) ARTICLE (149) MEETING ABSTRACT (2)

Tabla 10. Publicaciones sobre SBR en Web of Science

Si a esto se le incluía la condición *AND Topic=(cross-language OR cross-lingual OR multilingual)* los resultados se limitaban a 37, 32 de los cuales provenían de actas de congresos.

Por tanto, el primer paso fue la identificación de los principales congresos, conferencias y foros que han tratado y tratan los SBR multilingües, con la intención de identificar, analizar y comparar los diferentes tipos de recursos lingüísticos utilizados. Aunque cada año se celebra un número mayor de congresos centrados en la RI, no todos ellos incluyen una sección dedicada exclusivamente a la investigación sobre los SBR, y aún son menos los que abordan el problema multilingüe. Aún así, existen varios congresos o foros donde la investigación de los SBR multilingües ocupa un lugar primordial.

El primer artículo se ha centrado exclusivamente en los recursos utilizados por los SBR multilingües presentados en CLEF, que centra sus cometidos en MLIR y CLIR. Este foro comenzó a celebrarse a partir en el año 2000, cuando la evaluación multilingüe empieza a tener auge dentro del ámbito de la RI. Es el foro europeo más destacado y está financiado con fondos de la Unión Europea. Sus dos objetivos más importantes son promover el desarrollo de infraestructuras para la puesta en marcha y evaluación de los SRI que trabajan con lenguas europeas en contextos monolingües y multilingües; y crear prototipos con información que pueda ser útil y usada por los desarrolladores de sistemas. El análisis exclusivamente de las publicaciones presentadas en las múltiples ediciones de CLEF se ha debido a los siguientes motivos:

- a) Es uno de los pocos congresos actuales que fomenta la investigación y el debate en MLIR y CLIR, de ahí que en cada edición se presente un número elevado de publicaciones que tratan en exclusiva sobre los SBR multilingües y translingües.
- b) El auge de ciertos recursos y herramientas lingüísticos ha propiciado la creación de áreas temáticas centradas meramente en la investigación de los mismos, lo cual ha facilitado el presente trabajo de investigación.

En total, se han analizado 947 artículos publicados en CLEF desde el año 2000 hasta la fecha de publicación del artículo. La metodología aplicada ha consistido en analizar la forma en la que cada uno de los investigadores ha descrito las características de su sistema, incluyendo los recursos y herramientas utilizados. Aunque muchas de las contribuciones presentadas a CLEF abordan los aspectos lingüísticos de los SRI, solamente en algunas de ellas tratan como tema principal los SBR multilingües o translingües y sus recursos.

Tras esta primera etapa del estudio se determinaron los trabajos de los que finalmente se extraería la información relativa a las herramientas y a los recursos lingüísticos utilizados en los

SBR. Estos constituyen un total de 165 publicaciones de las 947 contribuciones presentadas en CLEF durante el periodo de estudio.

A continuación, se procedió a identificar los recursos lingüísticos utilizados en los sistemas evaluados en CLEF y a analizar sus características a partir de la descripción recogida en las publicaciones. Esta etapa consistió en una exhaustiva fase de observación documental cuyo resultado permitió determinar la evolución y los progresos conseguidos en este ámbito de investigación.

Tras haber analizado exclusivamente las publicaciones presentadas en *Cross-Language Evaluation Forum*, el segundo paso fue ampliar la muestra de la investigación al resto de congresos, conferencias y foros que han tratado y tratan los SBR multilingües. En el segundo artículo se estudiaron 315 publicaciones presentadas entre los años 2000 y 2010 en las conferencias ya mencionadas y en otras de interés. Al igual que en el anterior estudio, se ha abordado el tema que trataban (aunque todas versaban sobre los aspectos lingüísticos de los SBR multilingües, solamente un porcentaje lo trataba como tema principal) así como los recursos lingüísticos o herramientas que han aplicado.

Por otro lado, se han identificado en ambos estudios los idiomas implementados en los distintos sistemas (véase Figura 17). El idioma más utilizado es el inglés ya que la mayoría de los sistemas trabajan con un par de lenguas y uno casi siempre es este idioma. Como ha sido CLEF la conferencia de la que más artículos se han analizado, explica el hecho que la mayoría de los idiomas usados sean occidentales, aunque el chino y el japonés tienen también un papel importante puesto que la mayor parte de los artículos presentados en NTCIR los utilizan.

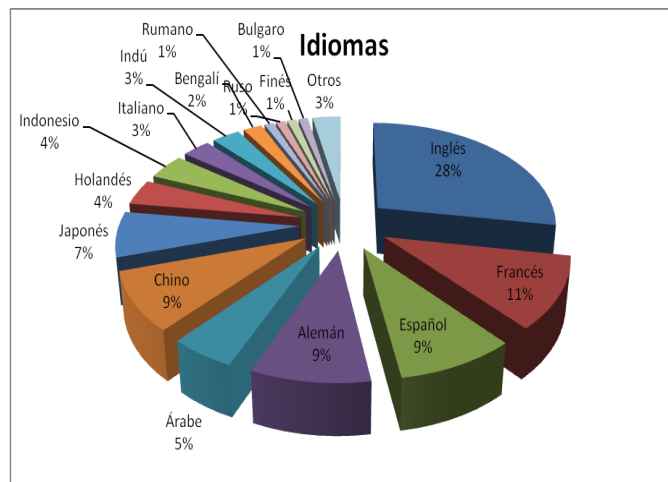


Figura 19. Idiomas utilizados en los SBR multilingües de las publicaciones analizadas

1.2.2.2. EVALUACIÓN ORIENTADA AL SISTEMA

El objetivo principal de la RI es localizar documentos que satisfagan las necesidades de información de los usuarios. Esta actividad implica una serie de procedimientos tales como la representación, el almacenamiento, la organización y el acceso a los elementos de información (normalmente documentos) (Baeza-Yates y Ribeiro-Neto, 2011).

Un modelo adecuado para la evaluación de la RI, como cualquier otro cuerpo riguroso de conocimientos, ha de cumplir con el *principium rationis sufficientis* de Leibnitz. En otras palabras, debe ser lo suficientemente poderoso como para aprehender todos los hechos relevantes del fenómeno estudiado. La premisa, pues, de todo estudio evaluativo ha de tener en cuenta la combinación de los tres elementos fundamentales que intervienen en la RI: el sistema, el usuario y la información (Olvera-Lobo, 1999). El desarrollo de medidas y métodos de evaluación que se adapten a los nuevos tipos de SRI y a las crecientes expectativas de los usuarios es fundamental para garantizar el progreso adecuado de este ámbito.

La RI propicia su necesidad crítica de evaluación al igual que cualquier otra actividad que aspire a ser clasificada como ámbito científico (Blair, 1990). Por ello y paralelamente al desarrollo de su tecnología, surge un amplio campo de trabajo dedicado específicamente a la determinación de medidas que permitan valorar su efectividad (Martínez Méndez, 2004).

La publicación 3 que inicia esta segunda etapa de la investigación, ha identificado todas las medidas de evaluación usadas por los organizadores de los principales foros en RI para la evaluación de los SBR monolingües, multilingües y translingües en los últimos 15 años. Se ha comprobado que la exhaustividad y la precisión basadas en la relevancia, y consideradas como las medidas tradicionales de la RI (Harman, 1998; Cleverdon, 1997; Salton y McGill, 1983), siguen contando en la actualidad con gran aceptación y consenso en la comunidad investigadora. No obstante, las medidas tradicionales de evaluación se han ido enriqueciendo con otras que las completan y complementan.

Los siguientes cuatro artículos que completan este apartado han servido para analizar y evaluar el funcionamiento y eficacia real de los SBR disponibles en la Web.

Evaluación orientada al sistema		
Publicación 4: START y MedQA	Publicación 5: START, QuALiM, MedQA y HONqa	Publicación 6: START, QuALiM, SEMOTE y TrueKnowledge
Publicación 7: HONqa		

Tabla II. SBR utilizados en la investigación realizada

Aunque como ya se ha comentado, desafortunadamente, aún existe un número bastante limitado de SBR en la Web, sí que se encuentran algunos ejemplos de SBR generales y específicos, monolingües y multilingües, etc. Por ello, esta fase de la investigación consistió en la evaluación orientada al sistema mediante la aplicación de diferentes medidas de evaluación. La Tabla 12 recoge el objetivo principal de cada test, los SBR implicados y las medidas utilizadas.

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

Evaluación orientada al sistema		
<p>Publicación 4: evaluación de los SBR como eficientes fuentes de información terminológica SBR: START y MedQA Medidas: MRR, TRR, FHS y precisión</p>	<p>Publicación 5: evaluación de los SBR de dominio general frente a los de dominio especializado en el ámbito biomédico SBR: START, QuALiM, MedQA y HONqa Medidas: MRR, TRR, FHS y precisión</p>	<p>Publicación 6: evaluación de la eficacia del funcionamiento de los SBR de dominio general SBR: START, QuALiM, SEMOTE y TrueKnowledge Medidas: MRR, TRR, FHS, precisión, AveP y MAP</p>
<p>Publicación: evaluación final. La evaluación orientada al sistema y al usuario en la búsqueda multilingüe de respuestas SBR: START, QuALiM, MedQA y HONqa Medidas: MRR, TRR, FHS, precisión, exhaustividad, MAP y AveP</p>		

Tabla 12. Características de cada test para la evaluación orientada al sistema

Con todos los test diseñados, se ha analizado, desde diferentes perspectivas y aplicando numerosas medidas de evaluación, la eficacia de funcionamiento de seis SBR, monolingües y multilingües, generales y especializados. A continuación, se detallan las características de cada uno de los estudios llevados a cabo.

La publicación 4 se centra en la búsqueda de información especializada. En este caso, el estudio se llevó a cabo en el SBR de dominio general START y en MedQA, de ámbito biomédico. Aunque este último ha desaparecido recientemente, han ido surgiendo, de forma progresiva, otros sistemas en la Web que ofrecen información sobre este tema, como EAGLi o HONqa. El dominio biomédico es el área de especialización en la que, gradualmente, se observa una mayor expansión de SBR disponibles para el usuario final. Además, se trata de un dominio temático muy popular, puesto que la mayor parte de usuarios se encuentran en mayor o menor medida familiarizados con ese ámbito por lo que, en un momento u otro, les resultará potencialmente útil para satisfacer algunas de sus necesidades de información más frecuentes. Por otro lado, es una de las áreas donde se desarrolla más investigación tanto dentro como fuera del ámbito de la RI –como muestra, indicar que, en la actualidad, existen un número bastante elevado de revistas científicas dedicadas exclusiva o parcialmente a algún aspecto del dominio biomédico–. Concretamente, en relación con la RI, y particularmente centrados en los estudios dedicados al análisis y evaluación de los SBR, van apareciendo numerosos ejemplos de estudios centrados en esta área (Athenikos y Han, 2010; Cruchet *et al.*, 2009; Fahmi, 2009; Jacquemart y Zweigenbaum, 2003; Lee *et al.*, 2006; Olvera-Lobo y Gutiérrez-Artacho, 2010; Olvera-Lobo y Gutiérrez-Artacho, 2011c; Olvera-Lobo y Gutiérrez-Artacho, 2013b; Terol *et al.*, 2007; Yu *et al.*, 2007; Zweigenbaum, 2005, 2009).

Para llevar a cabo el estudio (Olvera-Lobo y Gutiérrez-Artacho, 2010), se partió de la colección de preguntas *CI*, donde se seleccionaron las casi 250 preguntas que versaban sobre diferentes temas médicos y presentaban la estructura sintáctica aceptada por los SBR. Finalmente, y puesto que lo que se pretendía evaluar no era la cobertura de las bases de datos de los sistemas analizados, START y MedQA, sino la eficacia de su funcionamiento y las fuentes de información utilizadas para ofrecer las respuestas, el corpus de preguntas se redujo a 200, que son las que obtuvieron respuesta en los dos sistemas analizados. Tras plantear las preguntas en START y MedQA, se analizaron y evaluaron las respuestas. La puntuación obtenida por cada respuesta a partir de la valoración realizada sirvió de base para la aplicación de las siguientes medidas de evaluación del funcionamiento de los sistemas (Radev *et al.*, 2001): *Mean Reciprocal*

Rank (MRR), *Total Reciprocal Rank* (TRR), *First Hit Success* (FHS) y precisión (Salton y McGill, 1983).

La publicación 5 amplía la evaluación llevada a cabo en el anterior estudio realizando un estudio comparativo de la búsqueda de respuestas especializadas en los SBR de dominio general (START y QuALiM) frente a los de dominio restringido, específicamente los de ámbito biomédico (MedQA y HONqa). Al igual que el anterior estudio, se partió de la colección de preguntas de evaluación *C1*. A partir de la colección, se seleccionaron las 150 preguntas de definición sobre diversos temas médicos que obtuvieron respuesta en los cuatro SBR analizados –START, QuALiM, MedQA y HONqa–, sistemas elegidos para este estudio debido a que sus bases de datos presentaban una extensa cobertura y estaban actualizadas. El conjunto de preguntas utilizadas superó el test de validez interna con un alfa de Cronbach de 0,997.

Tras plantear las preguntas en estos cuatro sistemas, varios profesionales médicos valoraron las respuestas como incorrectas, inexactas o correctas. La valoración realizada para cada respuesta sirvió de base para la aplicación de las medidas de evaluación del funcionamiento de los sistemas que, al igual que para la publicación 4, fueron MRR, TRR, FHS y precisión.

En la publicación 6 se evalúa la eficacia en el funcionamiento y la calidad de las respuestas de los principales SBR de dominio general disponibles en la Web (START, QuALiM, SEMOTE y TrueKnowledge) ante preguntas diversas en tipo (de definición, factuales y de lista) y en temas (Arte y Literatura, Biología, Personajes, Historia, Economía o Deportes, entre otros). Las preguntas utilizadas para analizar el funcionamiento de los sistemas fueron las *test questions* que constituyen la colección 2. De las 1980 preguntas de *C2*, solamente se obtuvieron respuestas por parte de cada uno de los cuatro sistemas evaluados en 500 de ellas, por lo que, finalmente fueron esas las utilizadas en este test.

Las respuestas ofrecidas por cada sistema fueron valoradas por un grupo de estudiantes de diferentes titulaciones de la Universidad de Granada. A partir de la valoración de las respuestas obtenidas, se aplicaron las medidas de evaluación MRR, TRR, FHS y precisión, así como AveP – promedio de la precisión– y *Mean Average Precision* (MAP).

El diseño y desarrollo de herramientas multilingües, o en otros idiomas diferentes al inglés, han fomentado la investigación y la creación de nuevos instrumentos para la búsqueda y recuperación de información multilingüe (Ferrández Escámez, 2008). No obstante, el interés suscitado por esta área ha tenido simplemente un tímido reflejo en los SBR disponibles en la Web. Si bien, como se muestra en los artículos 1 y 2, el volumen de propuestas y prototipos desarrollados por los investigadores ha sido bastante numeroso en los últimos diez años, no se observa la misma evolución en la comercialización de este tipo de sistemas ni en su disponibilidad para el usuario final.

Actualmente solo se encuentran disponibles un número bastante limitado de SBR multilingües que, además, mayoritariamente, están desactualizados por parte de sus desarrolladores y no ofrecen buena cobertura. Únicamente destacaba de forma satisfactoria un SBR multilingüe en la Web, HONqa, aunque actualmente no se encuentra en la misma situación. Se trata de un sistema que recupera información en inglés, francés e italiano (Cruchet *et al.*, 2009) desarrollado por la *Health On the Net Foundation*⁵⁶, una organización suiza sin ánimo de lucro fundada en 1995. Este sistema se caracteriza por tener una interfaz muy sencilla donde plantear la pregunta deseada, si bien ofrece la posibilidad de una búsqueda avanzada que permite especificar un idioma determinado, la fuente de información de donde se extraerán las

⁵⁶ Health On The Net Foundation, <http://www.hon.ch/>

respuestas así como el número máximo de respuestas que se desea obtener. Una de las características principales de este sistema es que permite al usuario evaluar positiva o negativamente la respuesta recuperada, ofreciendo esa información a los desarrolladores.

Por todo ello, se planteó un estudio final (publicación 7), en el cual se trata de aunar los objetivos planteados en cada uno de los *test* anteriores con el fin de ofrecer una visión general de la situación. Para ello se utiliza el SBR multilingüe HONqa. Este artículo permite ofrecer una visión global de la situación, puesto que abarca el punto de vista del análisis y la evaluación cuantitativa –al analizar las respuestas proporcionadas por el SBR en tres idiomas (inglés, francés e italiano)– y también proporciona una evaluación cualitativa, que respalda los resultados obtenidos en la evaluación orientada al usuario y en el análisis de los recursos –mediante una evaluación comparativa de los recursos utilizados para la obtención de las respuestas en cada uno de los idiomas–.

Para llevar a cabo este test se partió de la colección de preguntas de evaluación *C7*, las cuáles fueron traducidas por un equipo de traductores profesionales al francés e italiano. De este conjunto inicial se seleccionaron las 120 preguntas de definición sobre diversos temas médicos, que obtuvieron respuesta en los tres idiomas disponibles en el SBR HonQA. El conjunto de preguntas utilizadas superó el test de validez interna con un alfa de Cronbach de 0,936. Las respuestas ofrecidas por el sistema fueron valoradas por un grupo de expertos de diferentes ámbitos médicos. A partir de la valoración de las respuestas obtenidas, se aplicaron las medidas de evaluación MRR, TRR, FHS, precisión, exhaustividad, MAP y AveP.

1.2.2.3. EVALUACIÓN ORIENTADA AL USUARIO: SATISFACCIÓN DE USUARIOS EN SBR MULTILINGÜES

Con el rápido crecimiento de Internet y el desarrollo de las nuevas tecnologías en los últimos años, los SBR han convertido en importantes aplicaciones. Aunque muchos son los que se han puesto en práctica, pocos son los trabajos que se han desarrollado en una evaluación centrada en el usuario. El éxito de la evaluación de los sistemas de información y la satisfacción de los usuarios son temas importantes en el campo de la gestión de la información, especialmente en los SBR. El objetivo de nuestro último estudio lo consideramos de especial interés, ya que actualmente la mayoría de los estudios se han centrado exclusivamente en la evaluación de los sistemas.

La publicación 8 parte del único trabajo hasta hoy realizado para medir la satisfacción de los usuarios (Ong *et al*, 2009). Nuestro objetivo ha sido conocer el grado de satisfacción del usuario para el SBR HONqa utilizando la herramienta diseñada por Ong y colegas que se basa en un examen de modelos y teorías. El modelo incorpora la satisfacción de los usuarios y la aceptación de la tecnología. Nuestro propósito es responder a cómo los usuarios evalúan el SBR multilingüe HONqa, y qué grado de satisfacción muestran los usuarios en el SBR multilingüe HONqa como fuente de información terminológica.

Con este último artículo, se ha querido completar la evaluación cuantitativa y cualitativa de los SBR desde diferentes perspectivas y métodos propuestos.

CAPÍTULO II. ESTUDIOS REALIZADOS

CAPÍTULO 2

2.1. LANGUAGE RESOURCES USED IN MULTI-LINGUAL QUESTION ANSWERING SYSTEMS (PUBLICACIÓN 1)

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011) "Language resources used in multi-lingual Question Answering Systems". *Online Information Preview*. 35 (4): 543 – 557

ABSTRACT

Purpose of this paper: In the field of information retrieval, some multi-lingual tools are being created to help the users to overcome the language barriers.

Nevertheless, these tools are not developed completely and it is necessary to investigate more for their improvement and application. One of their main problems is the choice of the linguistic resources to offer better coverage and to solve the translation problems in the context of the multi-lingual information retrieval.

Designs: This research is focused in the analysis of resources used by the multilingual question answering systems, which respond to users' queries with short answers, rather than just offering a list of documents related to the search. We have carried out an analysis of the main publications about the multi-lingual QA systems with the aim of identifying the typology, the vantages and advantages, and the real use and trend of each one of linguistic resources and tools used in this new kind of system.

Findings: We identified and studies five of the most used resources used in the cross-languages QA systems: databases, dictionaries, corpora, ontologies and thesauri. The three most popular traditional resources (automatic translators, dictionaries, and corpora) are gradually leaving a widening gap for others – such as ontologies and the free encyclopaedia Wikipedia.

What is value of paper: The perspective offered by the translation discipline can improve the effectiveness of QA systems.

2.1.1. INTRODUCTION

In the field of information retrieval (hereafter IR) monolingual and multi-lingual tools are being created that can greatly assist specialists in their work; as well as helping other users find a wide variety of information. Multi-lingual tools are evolving but several years of study and research are still needed to improve implementations. One of the main difficulties facing these tools is the task of translating queries made by users and the documentary sources found in response (Diekema, 2003). Given the current expansion in research, development, and the creation of multi-lingual IR systems, it was considered worthwhile analysing and evaluating the resources used by one type of these systems: multi-lingual question answering systems (hereafter QA systems).

Although research in this area began just over a decade ago, QA systems remain largely unknown outside the field of IR. In this context, a study from the perspective of translation may offer a different focus on the problem of translation and resources. Researchers currently working in the field of multi-lingual QA systems are searching for new methods to improve the efficiency of IR without using too many resources for language problems. These multi-lingual tools are being created to help the users to overcome the language barriers. Nevertheless, these tools are not developed completely and it is necessary to investigate more for their improvement and application. One of their main problems is the choice of the linguistic resources to offer better coverage and to solve the translation problems in the context of the multi-lingual information retrieval.

A multi-lingual QA system cannot easily retrieve relevant information for the user without an optimal solution for translation resources. For this reason, translation is crucial in this environment and enables problems to be analysed from a fresh point of view. Any progress made in solving problems of multi-lingual communication can be added to existing information retrieval systems.

The basic premises that have guided this study were: a) the use of QA systems enables research in multiple languages, provides faster responses, and increases the likelihood of the user obtaining the right answer; b) the multi-lingual QA systems try to overcome the language barrier, which is one of the most maxim in the IR field; c) the linguistic tools that most affect efficiency in the field of multi-lingual QA systems must be identified.

This paper is primarily intended as a general purpose analysis and aims to encompass translation in the study of multi-lingual QA systems. The second general aim is to identify and analyse the linguistic resources and tools found in these systems. Specific objectives include

identifying the main types of language resources and tools useful in the multi-lingual IR processes associated with multi-lingual QA systems, and establishing how much use is made of these tools by multi-lingual QA systems.

2.1.2. STATE OF THE ART ABOUT THE QUESTION ANSWERING SYSTEMS

Recent advances in IR and Web globalization mean that multi-lingual search systems have been developed in which translation and language resources are as important as the documentary and computer tools. This type of system has opened a new research field that examines the most effective methods for IR, as well as studying which resources are required for a correct translation.

Information overload is felt more strongly on the Web than elsewhere. All too often a query made with a web search tool (search engine, meta-search engine) results in the retrieval of too many pages – many of which are useless or irrelevant to the user. Therefore, professionals from various areas are beginning to recognize the usefulness of other types of systems, such as QA systems, for quickly and effectively finding specialist information (Crouch *et al.*, 2005, Lee *et al.*, 2006).

Traditionally, IR is understood as a fully automatic process that responds to a user query by examining a collection of documents and returning a sorted document list that should be relevant to the user requirements as expressed in the query. An optimal IR system recovers all the relevant documents (implying an exhaustive search, i.e. a high recall) and only the relevant documents (implying perfect accuracy, that is to say, a high precision). This traditional model involves many implied restrictions: a) the assumption that users want full-text documents, rather than answers, and that the query will be satisfied with these documents; b) that the process is direct and unidirectional rather than interactive; c) and finally, that the query and document share the same language.

Multi-lingual IR or CLIR (cross-lingual information retrieval) involves at least two languages in this process. Traditionally, CLIR is described like the problem to offer documents for users which they can read (Oard and Gonzalo, 2001). However, it is not only. In a multi-lingual environment such as the Web, most IR systems (search engines) are limited to finding documents in the language of the query; or alternatively, include machine translation systems, which are only useful once the documents are located and do not effectively cross the language

barrier. QA systems are an evolutionary improvement in IR systems. As an alternative to traditional IR systems they give correct and understandable answers to factual questions – rather than just offering a list of documents related to the search (Jackson and Schilder, 2005). The benefit is that users do not have to read whole documents to find the desired information. QA systems have attracted major attention since the TREC-8 (Text Retrieval) conference on information retrieval held in Voorhees in 1999. TREC conferences have been the major forum for sharing and encouraging international research in information retrieval since 1992.

QA systems are based on short-answer models (Blair-Goldensohn, 2004) that divide the question by assigning to the keyword a label that indicates the type of questions that can be answered. The system replaces this label with the right words to give users a selection of texts that respond correctly to the query (Perez-Coutiño *et al.*, 2004b). The main advantage is that the user does not have to read the documents to find the required information because the system provides the correct answer in the form of a number, a noun, a short phrase, or a brief piece of text.

Although there are various templates for making queries in QA systems, most of these systems understand questions expressed with interrogative particles (who, what, where, why, when, and how); while some understand the imperative form (tell me). When a query is entered into the interface, the system proceeds to analyze the question by separating the word or keywords. The system then locates and extracts one or several answers from different sources of information, depending on the specialist area of the question (Olvera-Lobo and Gutierrez-Artacho, 2010). Subsequently, the system evaluates and eliminates redundant information, or information that does not respond correctly to the question, and submits one or more prepared responses to the user (Cui *et al.*, 2004; Tsur, 2003).

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

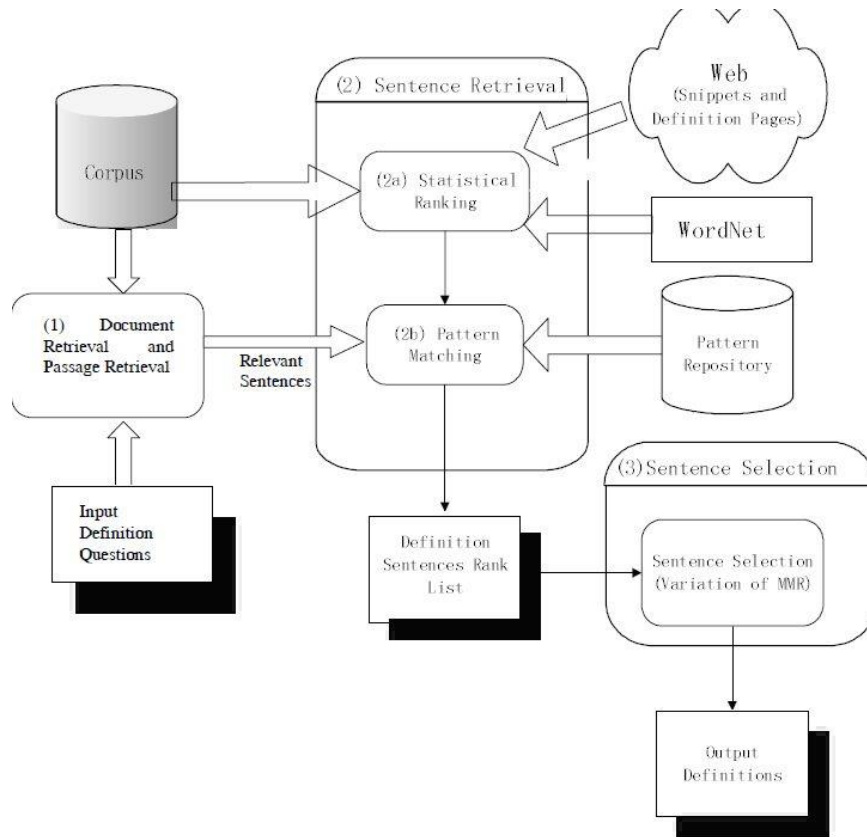


Figure 18. An example of a basic architecture of a monolingual QA system (Cui *et al.*, 2004)

These systems usually have a simple interface where users can enter their queries, while some offer a list of recent queries to help users understand how the system works. QA systems handle these queries by applying algorithms and methods of linguistic analysis; as well as using natural language processing to identify the components and determine the expected response (Zweigenbaum, 2005). This analysis usually uses a variety of standard questions in which certain words are replaced by labels accepted by the system.

QA systems may be general domain and so answer questions from diverse fields in the same way START or NSIR (see figure 19). Alternatively, they may be domain-specific and focus on a specialized area, in the same way as MedQA (Frank *et al.*, 2006). Domain-specific systems use specific linguistic resources that enable more precise answers to be given (see figure 20).

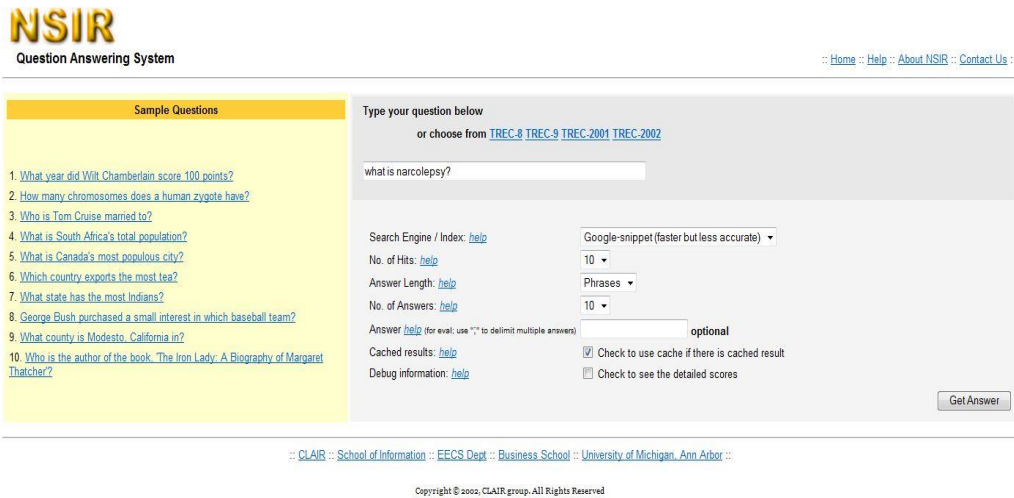


Figure 19. NSIR answer search interface

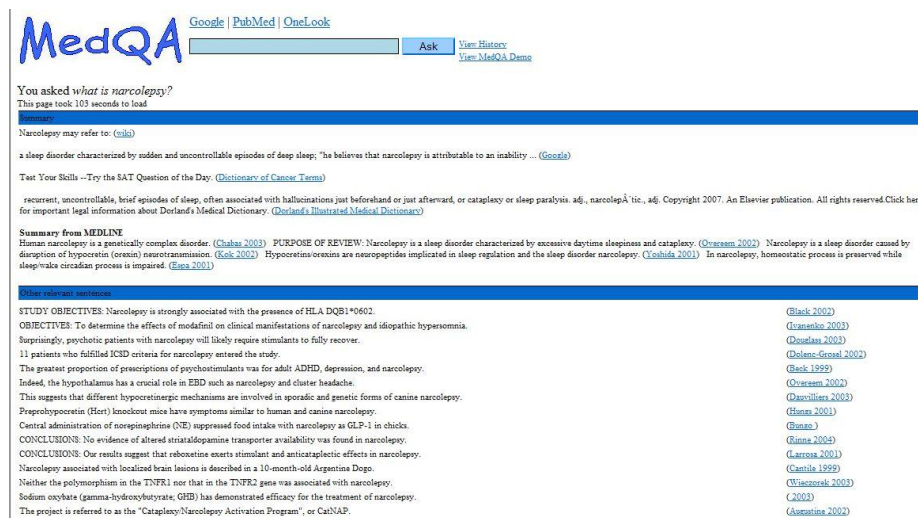


Figure 20. Answers provided by MedQA to the query: What is narcolepsy?

Another key aspect of these systems is that the system-user relationship is two-way. Establishing an interaction helps QA systems find better answers, and in turn, the QA system helps users find answers more quickly. However, it remains necessary to deepen the interactive design of these systems and enable true feedback between questions and answers, so that users communicate with the system in a conversational manner.

While the development of QA systems represents progress, the systems nevertheless suffer restrictions. Many were only developed as prototypes, or demonstration versions, and few were marketed. Some researchers have designed and created systems that were presented and discussed at various forums and conferences. However, because the usefulness of the systems was limited to very specific contexts, or because of problems of implementation, only a few of these systems were later developed for end users.

These circumstances have fuelled academic interest in cross-lingual IR, or CLIR, and the techniques of natural language processing. This interest led to many conferences dealing exclusively, or partially, with CLIR – such as TREC, the Cross-Language Evaluation Forum (CLEF), the NII Text Collection for IR Systems (NTSIR), the Language Resources and Evaluation Conference (LREC), among others. However, research on CLIR, which mostly began in 1996, has not led to commercial success and so dissemination was limited.

Given a particular query, CLIR systems run on a collection of multi-lingual documents and retrieve relevant information regardless of the language used in the query (Grefenstette, 1998). Within the area of multi-lingual IR, the object of our study is multi-lingual QA systems and these systems are opening a new field of research that is becoming increasingly important within CLIR. In multi-lingual QA systems, the language of the question may differ from the language of the retrieved document. However, QA systems differ from other CLIR systems because they do not retrieve whole documents and instead respond to queries with a short answer. According to Aceves Pérez (2008), QA systems are a set of coordinated monolingual systems in which each extracts responses from a collection of separate monolingual documents. Normally, multi-lingual QA systems are similar to monolingual QA systems, the main difference being the incorporation of a translation module and/or linguistic tool for cross-lingual recovery (figure 21). Most systems tackle the cross-lingual problem by translating the question or query posed in the source language in the target language, and then use a QA system developed for the target language for retrieving an answer (Bos and Nissim, 2006).

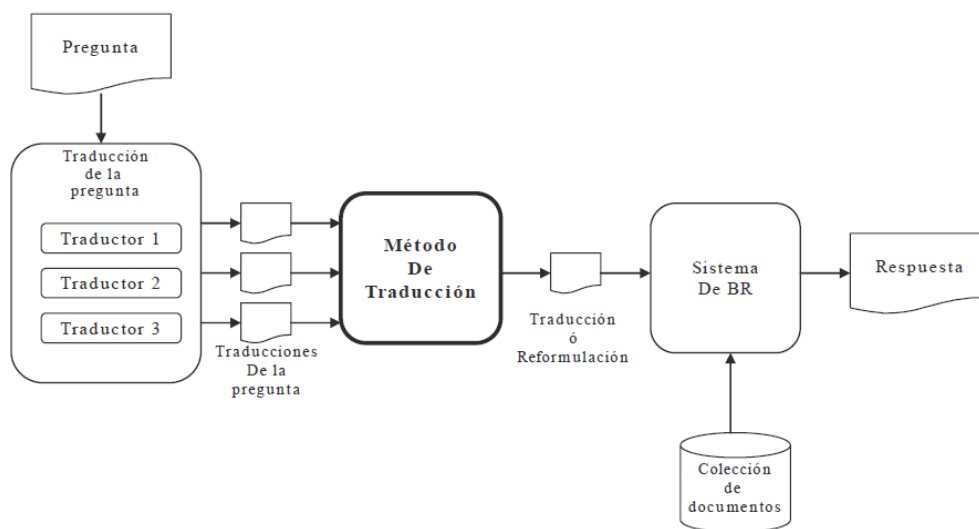


Figure 21. An example of a basic architecture of a multi-lingual QA system (Aceves Pérez, 2008)

Multi-lingual QA systems have emerged as a complementary research task, representing a promising direction for at least two reasons. First, it allowed users to interact with machines in their native languages, contributing to easier, faster, and more equal information access. Second, multi-lingual capabilities enabled QA systems to access information stored only in language-specific text collections (Forner *et al.*, 2010).

QA System	Web page
ACQUILEX	http://www.cl.cam.ac.uk/research/nl/acquilex/
AnswerBus	http://www.answerbus.com/index.shtml
DFKI - MULTI LINGUAL WEB QUESTION ANSWERING SYSTEM	http://experimental-quetal.dfki.de/
HonQA	http://services.hon.ch/cgi-bin/QA10/qa.pl
Inferret	http://asked.jp/edw/pc/

Table 13. List of the most important multi-lingual QA systems

Translation is crucial in CLIR because queries and documents do not always share the same language. The main translation problems identified are: lexical ambiguity, lack of translation coverage, multi-modal lexemes, and errors in lexical resources (Diekema, 2003). However, translation aspects have been relatively neglected during the development of these systems. Most of the work in this field has been carried out by artificial intelligence and computer specialists, so that the above problems have not been given priority.

2.1.3. METHOD SECTION

A semi-empirical methodology was adopted for this study and the collection of data about the tools and linguistic resources employed by these systems; as well as their use and implementation.

The first stage of the study focused on identifying the major conferences, meetings, and forums that address multi-lingual QA systems. The aim was to find, analyze, and compare the different types of linguistic resources. Although a growing number of IR conference are held each year, not all include a section devoted exclusively to QA systems, and even fewer tackle multi-lingual aspects. However, we identified several conferences and forums that mainly focus on research into multi-lingual QA systems. The most important is CLEF (see figure 22). This European conference was first held in 2000, when interest in the multi-lingual aspects of IR began to grow.

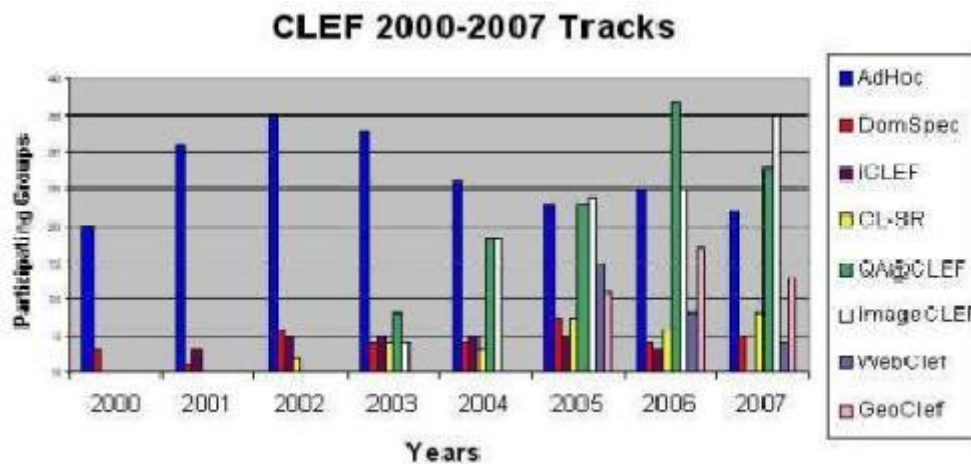


Figure 22. Groups participating in each of the subject areas of CLEF

Papers presented at TREC were also monitored to carry out this analysis. Although TREC is centred on retrieval techniques in English, it also covers other languages and each conference features an interesting area devoted to multi-lingual retrieval. TREC was the first conference to address the issue of information retrieval and has been held annually since 1992. NTCIR was the third conference we reviewed and is dedicated exclusively to IR. This conference is mainly focused on Asian languages (Japanese, Chinese, and Korean), while accepting multi-lingual research on other language pairs. Our study also included other papers dealing with language resources and describing work presented at conferences that have either only been active for a few years, or are not dedicated exclusively to IR – such are LREC. Some important studies presented in other journals were also analyzed; together with the few PhD dissertations addressing linguistic tools and resources in multi-lingual QA systems.

In total, some 165 papers published between 2000 and 2008 at the above, and other, conferences were reviewed. No papers from 2009 were included because some of the conferences that were held are yet to publish their proceedings. Over 75% of the published articles were presented at CLEF. NTCIR had the second highest number of analysed articles, and TREC was in fourth place (see Figure 23).

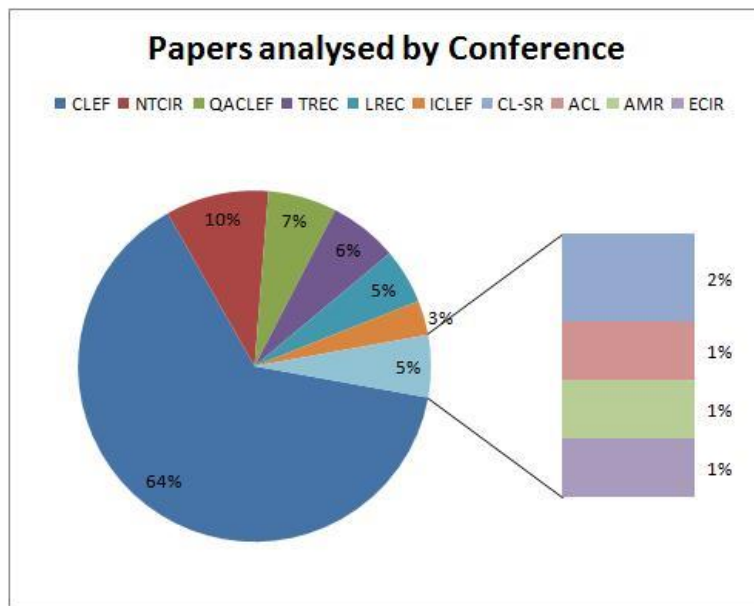


Figure 23. Papers analysed by Conference

For the studied period, the years with the largest number of papers published on multi-lingual QA systems were 2005 and 2008. A growing level of interest peaked in 2008; and from 2007 interest began shifting to other types of QA systems such as image, voice, and expertise domains (see Figure 24).

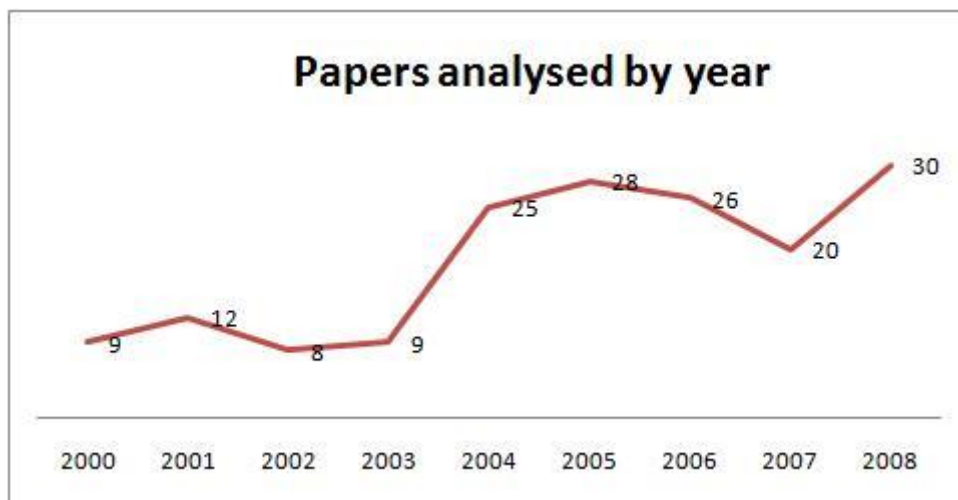


Figure 24. Papers analysed by year

We studied the subject discussed in each paper – including the language resources and tools used. Although all the papers discussed the linguistic aspects of multi-lingual QA systems, only some tackled this as the main theme.

In a second phase we explored the resources used by existing multi-lingual QA systems. For some systems, it was relatively easy to obtain information because the linguistic resources were freely accessible and developers provided all the relevant literature. However, these were the exceptions. Most of the systems were partially developed prototypes and access was not available. For this reason, the documentary observation phase of our study was so important because it enabled us to monitor the progress made by these developers.

2.1.4. RESULTS AND DISCUSSION

Five main types of linguistic resources used in multi-lingual QA systems were identified following an analysis of the literature. The main resource types were databases, corpora, dictionaries, ontologies, and thesauri.

There were also two types of linguistic tools used by these systems, namely, automatic translators and computational grammars. These resources and tools, along with their various types and subtypes, do not run in the same way and use differing methods of processing information. Other methods for solving the problems of cross lingual communication were also used – such as translation and transliteration – and these tools play an important role in several of the systems. Sometimes, a single resource was insufficient and several resources were used together to achieve better results.

Previous works (Diekema, 2003) identified four major sources of translation in CLIR – ontologies, bilingual dictionaries, automatic translators, and corpora (see Figure 25). This study shows that CLIR has grown in popularity in recent years and that some resources are often used.

Following an analysis of the literature and after identifying the resources and tools used by multi-lingual QA systems, a classification was made dividing these resources into two large groups: linguistic resources and linguistic tools.

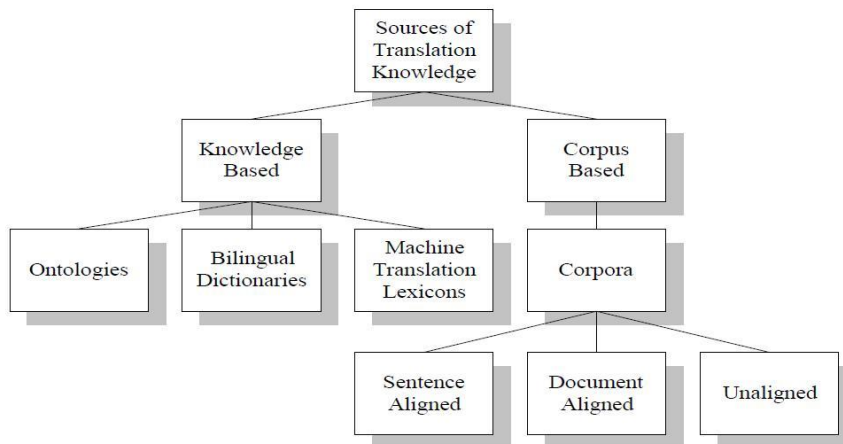


Figure 29. Resources used by multi-lingual QA systems for translation (Diekema, 2003)

Recent research and advances made in multi-lingual QA systems relate mainly to the more effective incorporation of new language resources, the creation of faster and more efficient systems, and the production of more transparent results. However, there remains an unsolved challenge: translation.

In analyzing the literature, it was found that the resource most used by multi-lingual QA systems was automatic translation, followed by corpora (mostly parallel), and dictionaries (see Figure 26). These results confirm the findings of Nguyen *et al.* (2009).

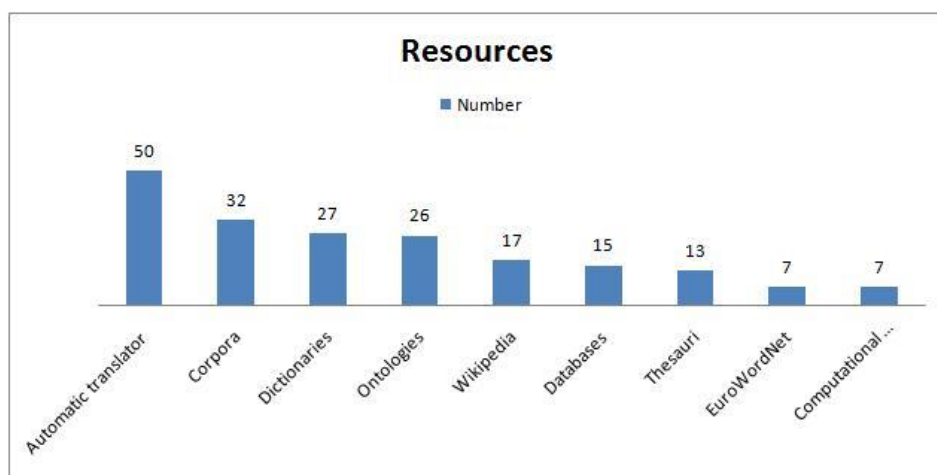


Figure 26. Resources used in the papers reviewed

Automatic translators were used in 50 of the 165 papers reviewed. This tool is often incorporated individually or in combination with other linguistic resources to offer better coverage. Although most authors confirm the problems of ambiguity and the poor quality of

texts, they continue to prefer this tool because it is one of the cheapest and easiest to incorporate into systems. Automatic translation usually gives better results in general domain QA systems than in specific domains. This is because automatic translators cannot identify and correctly translate certain specialized terms. Nor can this tool be recommended for systems that use non-Western languages, or more than two languages. In fact, few automatic translators are effective in these tasks.

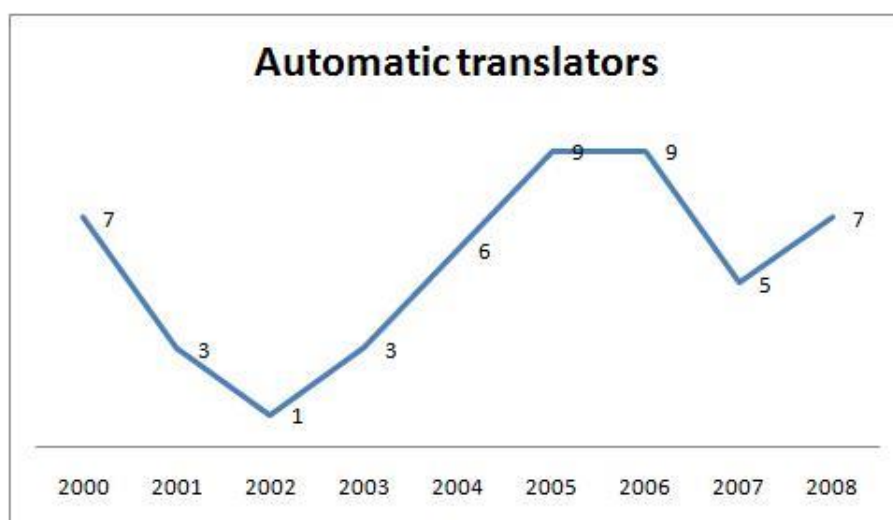


Figure 27. Use of automatic translators per year

However, the use of automatic translators has declined in recent years (see Figure 26). They were used in seven out of nine papers reviewed in the year shown in the figure. However, their presence declines substantially over the next three years (2001, 2002, and 2003). The number of multi-lingual QA systems using automatic translation rose again after 2006, yet not individually as in earlier years, but in combination or in support of other language resources. Automatic translators have continued to be used in the most recent years – but in a smaller number of systems.

The second most commonly used resource is the corpus with 32 occurrences. The apparent popularity of corpora is explained by the fact that many variants of corpora are included within the heading. The most surprising aspect of this resource was its nearly steady growth in recent years and the peak in 2008 – when corpora appeared in 11 of the 30 papers reviewed (Figure 28). We saw a significant decline in use in 2007, but this may be partially attributed to the fact that only 20 papers on multi-lingual QA systems were found for the year.

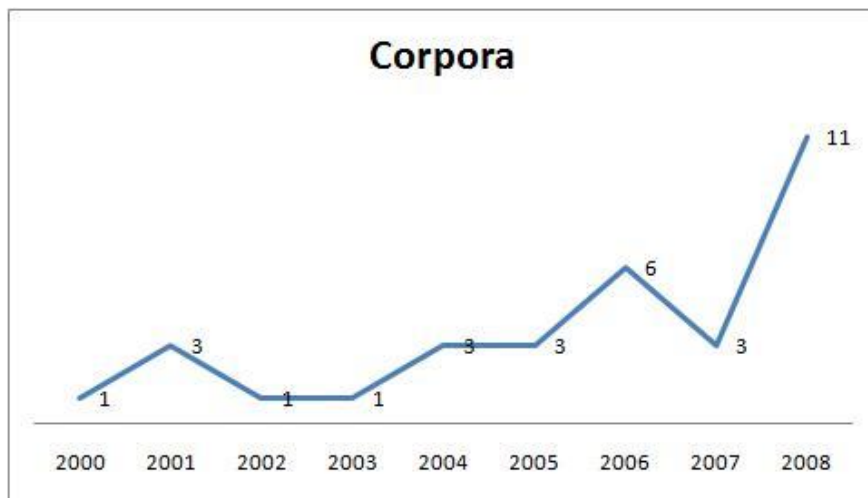


Figure 28. Use of corpora per year

Linguistic corpora are very useful resources for specialized domains. This is because the information received by users will be complete and correct when a translation is made or reviewed by professional translators. Existing corpora can be made available on the web in several languages, so solving two of the main problems raised earlier: computational cost and storage.

The third and fourth most commonly used resources are dictionaries and ontologies, with 27 and 26 appearances respectively. Dictionaries, together with automatic translators and corpora, are the resources traditionally used by these systems, and so a similar trend is found for all three resources. However, grammar and ambiguity problems have recently reduced their popularity, so that only 5 of the 79 systems studied over the past four years used this resource.

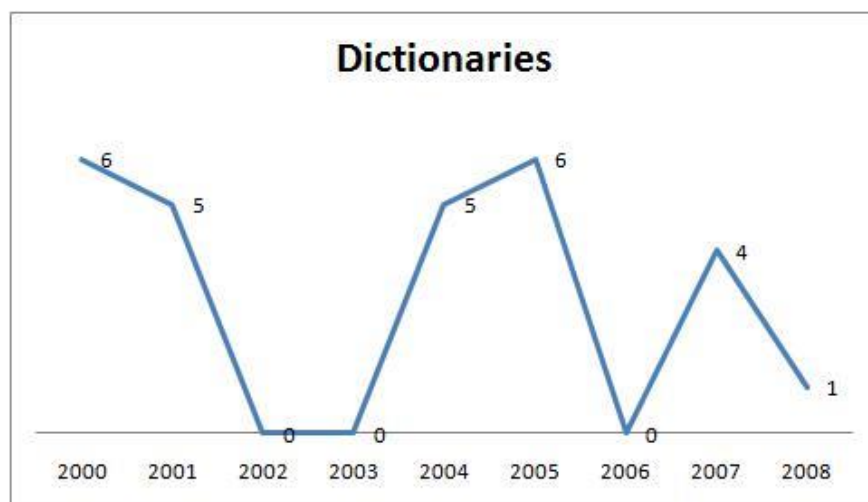


Figure 29. Use of dictionaries per year

Very different behaviour is seen with the fifth application – ontologies (see Figure 30). This resource was not used in the early years, but from the year 2004 has begun to slowly gain acceptance in multi-lingual QA systems. Ontologies offer many advantages and especially in specialized domain systems. Most systems are composed of texts that have been completely translated into various working languages, and so relationships are easily established. Another advantage is that there are many research teams working closely with multi-lingual ontologies and studying the various relationships that can be made between terms – and this existing body of work ensures a quality final product.

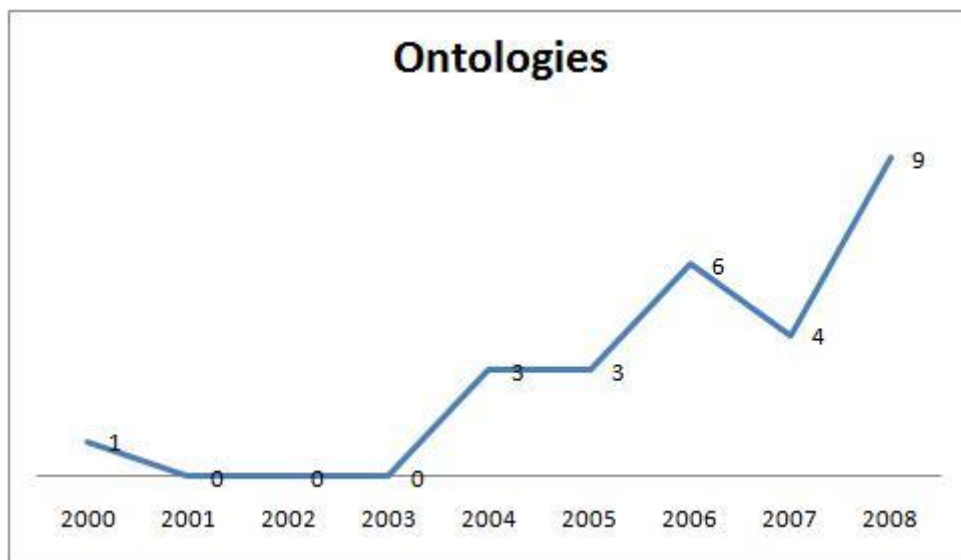


Figure 30. Use of ontologies per year

Wikipedia was used on 17 occasions. This is one of the most innovative resources and is growing rapidly in popularity. It was first incorporated in 2005, and its presence grew substantially the following year. The use of databases, which were used as often as Wikipedia, was very irregular – being entirely absent during some years. The results obtained by incorporating Wikipedia into such systems are unclear; some researchers claim it is one of the most useful resources, while others stress that there are many errors that are difficult to resolve.

Thesauri were used 13 times. The use of this resource was very limited and irregular – despite a peak in 2004 with six incorporations. This is surprising given that the architecture of thesauri makes them one of the most suitable resources for these systems; although when used alone they do not offer very good results. However, this may change when specialist domain multi-lingual QA systems are developed because many well established thesauri exist on a wide

range of topics. The final resource analysed was the EuroWordNet and computational grammars – and they were used a total of seven times.

The type of translation performed by each of the systems was also analyzed (see Figure 31). The most popular option was the translation of queries, and this is because this is often the most convenient for the system and therefore cheaper. However, the disadvantages and problems of ambiguity do not mean that it is the most appropriate option. The second most popular option was document translation – and this indicates a trend towards quality. We believe that the combination of these first two options is the most appropriate approach because the problems of ambiguity caused by nouns and homonyms are reduced. It is worth noting that QA translation has slightly grown in popularity – although it must be stressed that this is because QA is a relatively new option. However, we believe that the translation of keywords and answers causes the same problems as the translation of queries, and this involves a series of computational costs that could be invested in solving the language problem.

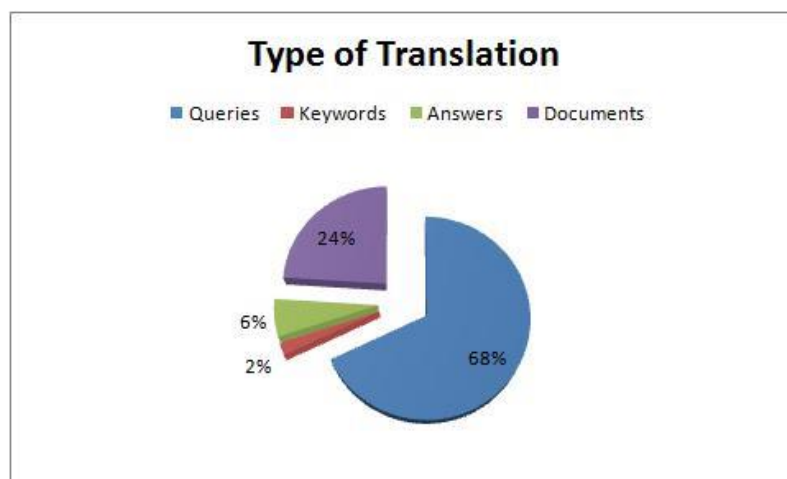


Figure 31. Type of translation used by the multi-lingual QA systems

2.1.5. CONCLUSION

This study has analyzed the main publications over the past nine years – from 2000 to 2008. Literature from 2009 was not included because it was insufficient and did not include real data regarding the situation. In total, 165 papers presented at major conferences (CLEF, TREC, NTSIR) were studied and as much data as possible was extracted for an overview of the situation.

Five of most used resources were identified and studied: databases, dictionaries, corpora, ontologies and thesauri. The final two resources were included in the same group as they referred to specific resources (Wikipedia and EuroWordNet) that were described by the researchers using them as either thesauri or ontologies. The second group in our study consisted of two linguistic tools: computational grammars and automatic translators. The inclusion of grammars in multi-lingual QA systems is relatively recent, and so the above classifications have not been taken into account. Finally, we studied the various other types of translation used by researchers and some new approaches such as transliteration and compensation translation.

After defining and describing these resources and tools, we considered each of the systems and techniques presented in the 165 papers. We found that automatic translators remain the most popular option, despite the fact that the authors of the papers recognise the resulting problems of ambiguity (see Figure 32). It is the low computational cost and ease of storage that accounts for the popularity of automatic translation among developers. In our opinion, this tool can be adequate for IR when combined with other resources. However, there have been some changes in the use and incorporation of these resources and tools. The three most popular traditional resources (automatic translators, dictionaries, and corpora) are gradually leaving a widening gap for others – such as ontologies and the free encyclopaedia Wikipedia. In addition, other approaches such as computational grammars are slowly attracting more researchers who are experienced in handling the results they produce. This data suggests that we may see unexpected changes in the future and this area deserves to be studied and evaluated in future research.

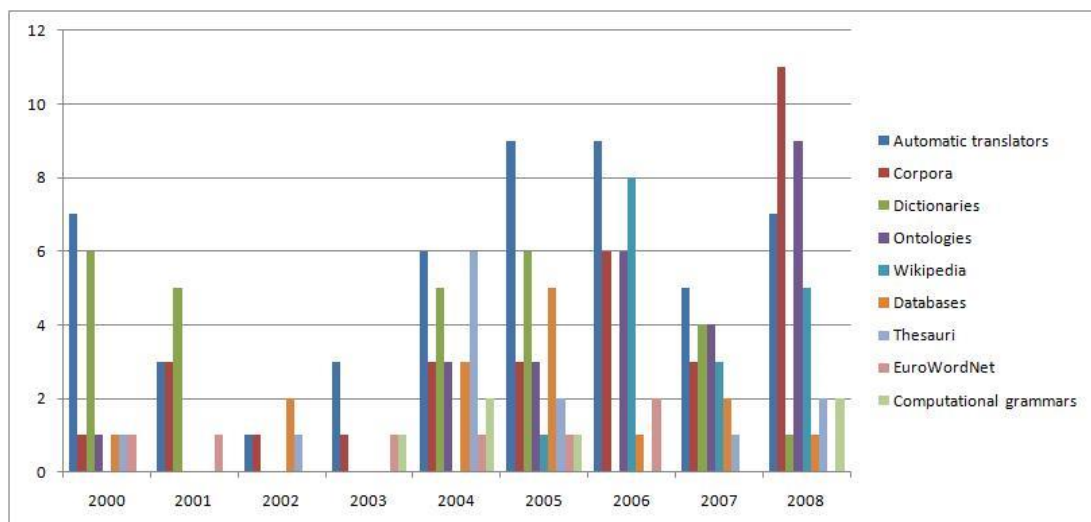


Figure 32. Resources and tools per year

A comparison of the evolution and use of different resources and tools shows that trends favour the traditionally more popular tools (automatic translators, dictionaries and corpora). However, ontologies and Wikipedia show trends that match, or nearly match, the traditional resources. The remaining tools are timidly growing in popularity and have promising futures. However, the trends for each combination of tools in multi-lingual QA systems were not studied exhaustively.

There is a growing trend toward the translation of documents, although the option to translate queries remains the most widely used by researchers. We believe the combination of both approaches is the most useful route and offers the best results – even handling more than two languages without difficulty.

2.2. LANGUAGE RESOURCES IN MULTI-LINGUAL QUESTION ANSWERING SYSTEMS (PUBLICACIÓN 2)

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2012) "Language Resources for Translation in Multi-lingual Question Answering Systems." *Translation Journal*, 16 (2)

ABSTRACT

In the field of Information Retrieval monolingual and multi-lingual tools are being created that can greatly assist specialists in their work; as well as helping other users find a wide variety of information. Multi-lingual tools are evolving but several years of study and research are still needed to improve implementations. One of the main difficulties facing these tools is the task of translating queries made by users and the documentary sources found in response (Diekema, 2003). Given the current expansion in research, development, and the creation of multi-lingual IR systems, it was considered worthwhile analyzing and evaluating the resources used by one type of these systems, the multi-lingual Question Answering systems.

Our article is primarily intended as a general purpose analysis and aims to encompass translation in the study of multi-lingual QA systems. The second general aim is to identify and analyse the linguistic resources and tools found in these systems. Specific objectives include identifying the main types of language resources and tools useful in the multi-lingual IR processes associated with multi-lingual QA systems, and establishing how much use is made of these tools by multi-lingual QA systems.

KEYWORDS: Information Retrieval, Question Answering Systems, Multi-lingual Question Answering Systems, Evaluation, Linguistic Resources, Translation

2.2.1. INTRODUCTION

In the field of information retrieval (hereafter IR) monolingual and multi-lingual tools are being created that can greatly assist specialists in their work; as well as helping other users find a wide variety of information. Multi-lingual tools are evolving but several years of study and research

are still needed to improve implementations. One of the main difficulties facing these tools is the task of translating queries made by users and the documentary sources found in response (Diekema, 2003). Given the current expansion in research, development, and the creation of multi-lingual IR systems, it was considered worthwhile analysing and evaluating the resources used by one type of these systems, the multi-lingual question answering systems (hereafter QA systems).

Although research in this area began just over a decade ago, QA systems remain largely unknown outside the field of IR. In this context, a study from the perspective of translation may offer a different approach focused on the problem of translation and linguistic resources. Researchers currently working in the field of multi-lingual QA systems are searching for new methods to optimize the efficiency of IR without using too many resources for language problems. However, a system cannot easily retrieve relevant information for the user without an optimal solution for translation resources. For this reason, translation is crucial in this environment and enables problems to be analysed from a fresh point of view. Any progress made in solving problems of multi-lingual communication can be added to existing information retrieval systems.

Recent advances in IR and Web globalization mean that multi-lingual search systems have been developed in which translation and language resources are as important as the documentary and computer tools. This type of system has opened a new research field that examines the most effective methods for IR, as well as studying which resources are required for a correct translation.

This article is primarily intended as a general purpose analysis and aims to encompass translation in the study of multi-lingual QA systems. The second general aim is to identify and analyse the linguistic resources and tools found in these systems. Specific objectives include identifying the main types of language resources and tools useful in the multi-lingual IR processes associated with multi-lingual QA systems, and establishing how much use is made of these tools by multi-lingual QA systems.

2.2.2. INFORMATION RETRIEVAL

Traditionally, IR is understood as a fully automatic process that responds to a user query by examining a collection of documents and returning a sorted document list that should be relevant to the user requirements as expressed in the query. An information retrieval system is a

system used to store items of information that need to be processed, searched, retrieved, and disseminated to users (Salton and Mc Gill, 1983).

When a need for information arises, a process called the “search strategy” is set in motion, which leads to the supply of documents by the system (Belkin and Croft, 1987). The process have to entail (Chowdhury, 1999): a) definition of the informational need; b) selection of the information sources to be used; c) translation of the user query expressed in natural language into the indexing language of the information source, if necessary; d) translation of the expression from the indexing language to the query language of each information system; e) implementation of expressions obtained from the query language; f) assessment of results by the user and the redefinition of the query expressions if the results are not relevant; and g) selecting and obtaining the documents that respond to the user’s needs.

An optimal IR system retrieves all the relevant documents (implying an exhaustive search, i.e. a high recall) and only the relevant documents (implying perfect accuracy, that is to say, a high precision) (Baeza-Yates and Ribeiro-Neto, 1999). This traditional model involves many implied restrictions: a) the assumption that users want full-text documents, rather than answers, and that the query will satisfied with these documents; b) that the process is direct and unidirectional rather than interactive; and finally, c) that the query and document share the same language.

Multi-lingual IR or CLIR (cross-lingual information retrieval) involves at least two languages in this process. In a multi-lingual environment such as the Web, most IR systems (search engines) are limited to finding documents in the language of the query; or alternatively, include machine translation systems, which are only useful once the documents are located and do not effectively cross the language barrier. Given a particular query, CLIR systems run on a collection of multi-lingual documents and retrieve relevant information regardless of the language used in the query (Grefenstette, 1998). Within the area of multi-lingual IR, the object of our study is multi-lingual QA systems and these systems are opening a new field of research that is becoming increasingly important within CLIR. Traditionally, CLIR is described like the problem to finding documents that the user cannot read (Oard, 2001).

However, it is not the only. One of the earlier works in CLIR was conducted by Salton (1970) and compared the effectiveness of English and German queries with that of queries obtained using a bilingual thesaurus for retrieving documents in both languages. Salton (1970) empirically showed that CLIR using a hand-crafted bilingual thesaurus is comparable with monolingual information retrieval in performance. Usually with CLIR a multilingual thesaurus of some sort is created to hold a list of descriptors for each document in a collection and the semantic relations

between them, and each term in the thesaurus must be translated for each language involved (Fluhr, 1996). The descriptors can be added to the thesaurus manually or automatically (if the system can learn from previous indexing which terms are likely to be important) (López Ostenero, 2002).

These circumstances have fuelled academic interest in multi-lingual IR, or CLIR, and the techniques of natural language processing. Although Salton (1970) is considered the “father” of the earliest research initiatives concerning CLIR, the first Workshop geared specifically to CLIR topics was celebrated in Zurich and it was organized by the Association for Computing Machinery (ACM) during the Special Interest Group on Information Retrieval, SIGIR-96 conference (Grefenstette, 1998). Nowadays, there are three important international forums about the evaluation of IR systems focusing on techniques and proceedings related with CLIR: Text Retrieval Conference (TREC), the Cross-Language Evaluation Forum (CLEF), the NII Text Collection for IR Systems (NTSIR) or the Language Resources and Evaluation Conference (LREC) (Olvera-Lobo, 2009).

The TREC, co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 (Vorhees, 1999) as part of the TIPSTER Text program. TREC was the first conference to address the issue of information retrieval and has been held annually. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. TREC has successfully met its dual goals of improving the state-of-the-art in information retrieval and of facilitating technology transfer. Retrieval system effectiveness approximately doubled in the first six years of TREC. Most of today’s commercial search engines include technology first developed in TREC.

One year later a new workshop is developed, called NII Test Collection for IR Systems (NTCIR). The NTCIR Workshop is a series of evaluation workshops designed to enhance research in Information Access (IA) technologies including information retrieval, question answering, text summarization, extraction, etc. The main aims of this conference are: a) to encourage research in IA technologies by providing large-scale test collections reusable for experiments and a common evaluation infrastructure allowing cross-system comparisons; b) to provide a forum for research groups interested in cross-system comparison and exchanging research ideas in an informal atmosphere; c) to investigate evaluation methods of IA techniques and methods for constructing a large-scale data set reusable for experiments.

In NTCIR it is presented works in Japanese, Chinese, Korean and English, although multi-lingual papers with other languages are accepted. Nine workshops have been celebrated since 1998 and they are divided in different areas or disciplines.

In 2000, it is created a new conference, CLEF, the most important European forum to the evaluation of multilingual and multimedia retrieval systems. CLEF is developed to promote R&D in multilingual information access by a) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and multi-lingual contexts, and b) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes. The final objective is to boost and encourage information retrieval technologies development in Europe in order to guarantee its competitiveness on a global sphere.

2.2.3. QUESTION ANSWERING SYSTEMS

Information overload is felt more strongly on the Web than elsewhere. All too often a query made with a web search tool (search engine or meta-search engine) results in the retrieval of too many pages – many of which are useless or irrelevant to the user. Therefore, professionals from various areas are beginning to recognize the usefulness of other types of systems, such as QA systems, for quickly and effectively finding specialist information (Crouch, *et al.* 2005; Lee, *et al.* 2006; Yu, *et al.* 2007).

QA systems are an evolutionary improvement in IR systems. As an alternative to traditional IR systems they give correct and understandable answers to factual questions – rather than just offering a list of documents related to the search (Jackson and Schilder, 2005). The benefit is that users do not have to read whole documents to find the desired information. QA systems have attracted major attention since the TREC-8 (Text Retrieval) conference. The QA systems were intended to stimulate human language behavior, therefore being able to answer natural language questions. As a result, the machine attempts to understand the question and thus responds by answering. In other words, understanding natural language by the machine is an essential process in the development of QA systems (Belkin and Vickery, 1985; Sultan, 2006).

Although there are various templates for making queries in QA systems, most of these systems understand questions expressed with interrogative particles (who, what, where, why, when, and how); while some understand the imperative form (tell me). When a query is entered into the interface, the system proceeds to analyse the question by separating the word or keywords. The system then locates and extracts one or several answers from different sources of information, depending on the specialist area of the question (Olvera-Lobo & Gutierrez-Artacho,

2010). Subsequently, the system evaluates and eliminates redundant information, or information that does not respond correctly to the question, and submits one or more prepared responses to the user (Cui, *et al.* 2004; Tsur, 2003).

These systems usually have a simple interface where users can enter their queries, while some offer a list of recent queries to help users understand how the system works. QA systems handle these queries by applying algorithms and methods of linguistic analysis; as well as using natural language processing to identify the components and determine the expected response (Zweigenbaum, 2005). This analysis usually uses a variety of standard questions in which certain words are replaced by labels accepted by the system.

QA systems may be general domain and so answer questions from diverse fields like START Natural Language Question Answering System, or NSIR Question Answering System. Alternatively, they may be domain-specific and focus on a specialized area, in the same way as HONQA (Health On the Net Foundation). Domain-specific systems use specific linguistic resources that enable more precise answers to be given.

Another key aspect of these systems is that the system-user relationship is two-way. Establishing an interaction helps QA systems find better answers, and in turn, the QA system helps users find answers more quickly. However, it remains necessary to deepen the interactive design of these systems and enable true feedback between questions and answers, so that users communicate with the system in a conversational manner.

While the development of QA systems represents progress, the systems nevertheless suffer restrictions. Many were only developed as prototypes, or demonstration versions, and few were marketed. Some researchers have designed and created systems that were presented and discussed at various forums and conferences. However, because the usefulness of the systems was limited to very specific contexts, or because of problems of implementation, only a few of these systems were later developed for end users (Olvera-Lobo and Robinson-García, 2009).

The aim of QA systems is to find exact and correct answer for user's questions. Analysis of question, search and choosing answer are three important items in a QA system. In addition to user interaction, various QA systems contain at least three following parts: question processing, document processing and answer processing (Kangavari *et al.*, 2008). Question reformulation or processing module tries to identify various ways of expressing an answer given a natural language question. This reformulation is often used in Question Answering system to retrieve answers in a large document collection (Verona & Motta, 2004). The objective of this step is to obtain features from the question that could be helpful in the following steps. All the information obtained by this module is given to the following steps of the system.

Document processing module depend the architecture of the QA systems. It is used to perform a first selection of paragraphs that are considered relevant to the input question.

And finally, answer processing module consists of two main components: answer extraction and answer validation. First, candidate answers extract from documents which are retrieve by search engine in answer extraction module. After that it validates answers with filtering and ranking candidate answers and final system suggests answers with user voting (Kangavari *et al.*, 2008). The mission of this step is to eliminate possible incorrect paragraphs contained in the list returned by the QA system. Thus, there are more possibilities of giving at the end of the process a correct answer (Rodrigo, *et al.* 2010).

QA systems are classified into two types according to the source of the question answer, that is, whether it is from a structured source or free text. The former systems return answers drawn from a database, and there are considered to be the earliest QA systems. The next generation of QA systems evolved to extract answers from the plain machine-readable text that is found in a collection of documents or on the Web.

2.2.4. MULTI-LINGUAL QA SYSTEM

The emergence of multi-language QA systems is a relatively recent development, and many of the existing systems are still in an experimental stage. In multi-lingual QA systems, the language of the question may differ from the language of the retrieved document. However, QA systems differ from other CLIR systems because they do not retrieve whole documents and instead respond to queries with a short answer. According to Aceves Pérez (2008), QA systems are a set of coordinated monolingual systems in which each extracts responses from a collection of separate monolingual documents. Normally, multi-lingual QA systems are similar to monolingual QA systems, the main difference being the incorporation of a translation module and/or linguistic tool for multi-lingual retrieval. The most basic method of multi-lingual QA systems contains two steps. In the first step, the original question is translated to the target language by machine translation. In the second step, the translated question is processed by a QA system and the system returns an answer in the target language. More advanced methods can incorporate natural language processing tools and text mining.

Usually QA systems that deal with multiple languages rely on a translation module. The user enters his specific query, generally including some interrogative adverb (How? When? Where?)

in a given natural source language. This question is translated by an automatic translator. In the stage of query analysis, the QA system examines the user's question and determines what type of information is being demanded. The classification of the questions is a key for the system, as this information will be utilized in the search stage, and in the selection and extraction of the potential responses (García Cumbereras *et al.*, 2005). The resulting search expression will be, then, the input, or the formulation of the query to be used by the search engine of the system for comparing and matching it with the documents in the database. Once the documents that are relevant to the query are located, the system breaks them up into sections, selects the excerpts that include the candidate responses, and selects a final response. This response, along with its location in the corresponding document, is finally delivered to the user (Olvera-Lobo and García-Santiago, 2010).

Besides the user's interface there is translation software in the multi-lingual QA systems architecture. Nowadays, machine translation shows a number of different facets and views. All of them have in common that this translation must be carried out by software in a more or less automatic way. The rate and quality of this translation can vary. But even the most sophisticated IR systems cannot yet produce translations on a large scale that do not need absolutely any revision by a person. The IR systems also have restrictions about the nature of the texts that they can translate better (Olvera-Lobo and García-Santiago, 2010; García-Santiago and Olvera-Lobo, 2010).

Most of the work in the field of multi-lingual QA systems has been carried out by artificial intelligence and computer specialists, so that the translation problems have not been given priority. However, translation is crucial in CLIR because queries and documents do not always share the same language. The main translation problems identified are: lexical ambiguity, lack of translation coverage, multi-modal lexemes, and errors in lexical resources (Diekema, 2003). There are some researches about the different linguistic resources and tools used by the multi-lingual QA systems (Olvera-Lobo and Gutierrez-Artacho, 2011b). These previous works have showed that the automatic translators remain the most popular option, despite the fact that the authors of the papers recognise the resulting problems of ambiguity.

2.2.5. METHOD SECTION

An empirical methodology was adopted for this study and the collection of data about the tools and linguistic resources employed by these systems; as well as their use and implementation.

The first stage of the study focused on identifying the major conferences, meetings, and forums that address multi-lingual QA systems. The aim was to find, analyze, and compare the different types of linguistic resources. Although a growing number of IR conference are held each year, not all include a section devoted exclusively to QA systems, and even fewer tackle multi-lingual aspects. However, we identified several conferences and forums that mainly focus on research into multi-lingual QA systems.

In total, some 315 papers published between 2000 and 2011 at the above, and other, conferences were reviewed. No papers from 2011 were included because some of the conferences that were held are yet to publish their proceedings. Over 75% of the published articles were presented at CLEF. NTCIR had the second highest number of analysed articles, and TREC was in fourth place (see Figure 33).

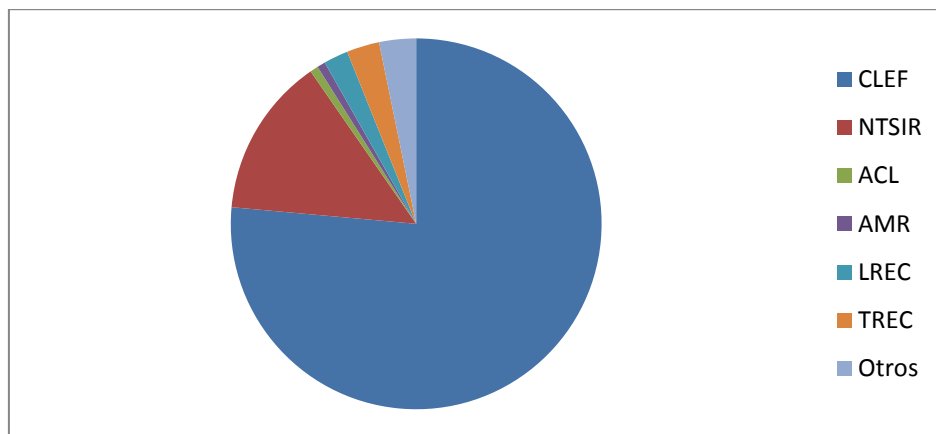


Figure 33. Papers analysed by Conference

For the studied period, the years with the largest number of papers published on multi-lingual QA systems were 2005 and 2008. A growing level of interest peaked in 2008; and from 2007 interest began shifting to other types of QA systems such as image, voice, and expertise domains.

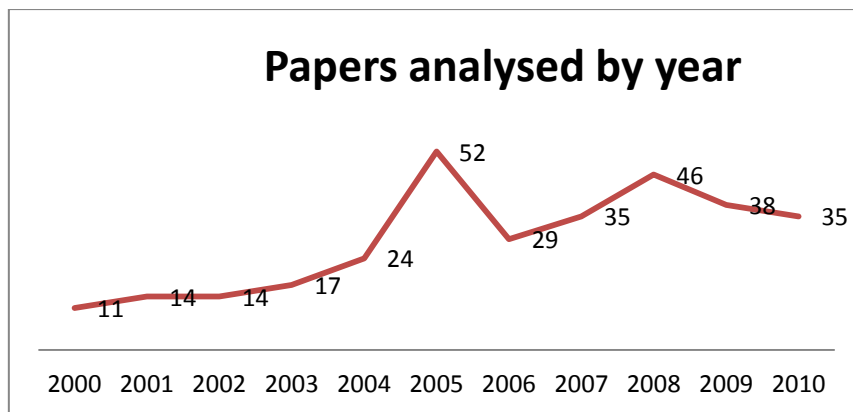


Figure 34. Papers analysed by year

We explored the resources used by multi-lingual QA systems showed in different conferences. We did an important documentary observation phase, because of analyzing and evaluating how many different linguistic resources and tools are used by these systems, so, it enabled us to monitor the progress made by these developers.

We studied the subject discussed in each paper – including the language resources and tools used. Although all the papers discussed the linguistic aspects of multi-lingual QA systems, only some tackled this as the main theme.

In a second phase we explored the resources used by existing multi-lingual QA systems. For some systems, it was relatively easy to obtain information because the linguistic resources were freely accessible and developers provided all the relevant literature. However, these were the exceptions. Most of the systems were partially developed prototypes and access was not available. For this reason, the documentary observation phase of our study was so important because it enabled us to monitor the progress made by these developers.

2.2.6. RESULTS AND DISCUSSION

Five main types of linguistic resources used in multi-lingual QA systems were identified following an analysis of the literature. The main resource types were databases, corpora, dictionaries, ontologies, and thesauri.

There were also two types of linguistic tools used by these systems, namely, automatic translators and computational grammars. These resources and tools, along with their various types and subtypes, do not run in the same way and use differing methods of processing information. Other methods for solving the problems of multi-lingual communication were also

used – such as translation and transliteration – and these tools play an important role in several of the systems. Sometimes, a single resource was insufficient and several resources were used together to achieve better results.

Previous works (Diekema, 2003) identified four major sources of translation in CLIR – ontologies, bilingual dictionaries, automatic translators, and corpora. This study shows that CLIR has grown in popularity in recent years and that some resources are often used. Following an analysis of the literature and after identifying the resources and tools used by multi-lingual QA systems, a classification was made dividing these resources into two large groups: linguistic resources and linguistic tools.

Recent research and advances made in multi-lingual QA systems relate mainly to the more effective incorporation of new language resources, the creation of faster and more efficient systems, and the production of more transparent results. However, there remains an unsolved challenge: translation.

In analyzing the literature, it was found that the resource most used by multi-lingual QA systems was corpora (mostly parallel), followed by machine translation, and Wikipedia (see Figure 35). We can realize that the traditional trilogy of resources in multi-lingua QA systems (dictionaries, machine translation and corpora) has changed (Nguyen, *et al.* 2009).

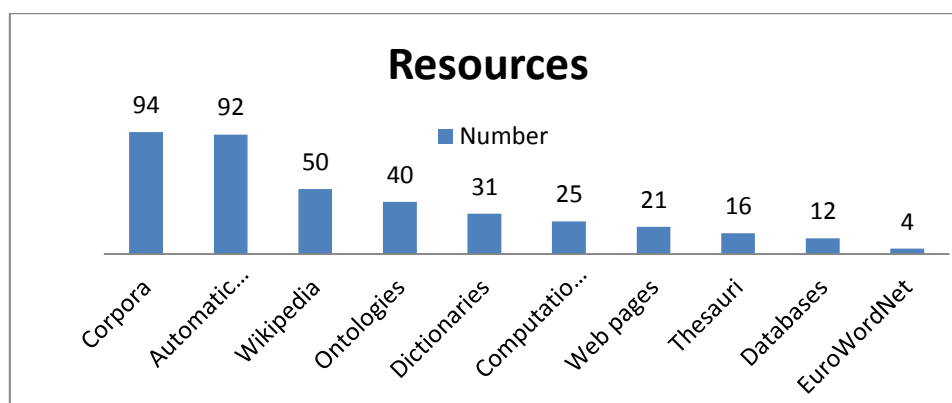


Figure 35. Resources used in the papers reviewed

Corpora were used in 94 of the 315 papers reviewed. The apparent popularity of corpora is explained by the fact that many variants of corpora are included within the heading. The most surprising aspect of this resource was its nearly steady growth in recent years and the peak in 2008. We saw a significant decline in use in 2006, but this may be partially attributed to the fact that only 29 papers on multi-lingual QA systems were found for the year. Linguistic corpora are very useful resources for specialized domains. This is because the information received by

users will be complete and correct when a translation is made or reviewed by professional translators. Existing corpora can be made available on the web in several languages, so solving two of the main problems raised earlier, computational cost and storage.

Automatic translators were used in 92 of the 315 papers reviewed. This tool is often incorporated individually or in combination with other linguistic resources to offer better coverage. Although most authors confirm the problems of ambiguity and the poor quality of texts, they continue to prefer this tool because it is one of the cheapest and easiest to incorporate into systems. Automatic translation usually gives better results in general domain QA systems than in specific domains. This is because automatic translators cannot identify and correctly translate certain specialized terms. Nor can this tool be recommended for systems that use non-Western languages, or more than two languages. In fact, few automatic translators are effective in these tasks.

However, the use of automatic translators has declined in recent years. However, their presence declines substantially over the next three years (2001, 2002, and 2003). The number of multi-lingual QA systems using automatic translation rose again after 2006, yet not individually as in earlier years, but in combination or in support of other language resources. Automatic translators have continued to be used in the most recent years – but in a smaller number of systems.

The third most commonly used resource was Wikipedia. Wikipedia was used on 50 occasions. This is one of the most innovative resources and is growing rapidly in popularity. It was first incorporated in 2005, and its presence grew substantially the following year. The use of databases, which were used as often as Wikipedia, was very irregular – being entirely absent during some years. The results obtained by incorporating Wikipedia into such systems are unclear; some researchers claim it is one of the most useful resources, while others stress that there are many errors that are difficult to resolve.

The fourth and fifth most commonly used resources were ontologies and dictionaries, with 40 and 31 appearances respectively. Dictionaries, together with automatic translators and corpora, are the resources traditionally used by these systems, and so a similar trend is found for all three resources. However, grammar and ambiguity problems have recently reduced their popularity, so that only 5 of the 79 systems studied over the past four years used this resource.

Very different situation is seen with the fifth application – ontologies. This resource was not used in the early years, but from the year 2004 has begun to slowly gain acceptance in multi-lingual QA systems. Ontologies offer many advantages and especially in specialized domain systems. Most systems are composed of texts that have been completely translated into various working languages, and so relationships are easily established. Another advantage is that there

are many research teams working closely with multi-lingual ontologies and studying the various relationships that can be made between terms – and this existing body of work ensures a quality final product.

Other five linguistic resources were used to solve the translation problem: computational grammars, web pages, thesauri, databases and EuroWordNet.

2.2.7. CONCLUSION

This study has analyzed the main publications over the past nine years –from 2000 to 2010. Literature from 2011 was not included because some of the conferences that were held are yet to publish their proceedings and did not include real data regarding the situation. In total, 315 papers presented at major conferences (CLEF, TREC, NTSIR, among others) were studied and as much data as possible was extracted for an overview of the situation.

Seven of most used resources were identified and studied: databases, dictionaries, corpora, ontologies, thesauri, Wikipedia and EuroWordNet. The second group in our study consisted of two linguistic tools: computational grammars and automatic translators. The inclusion of grammars in multi-lingual QA systems is relatively recent, and so the above classifications have not been taken into account.

After defining and describing these resources and tools, we considered each of the systems and techniques presented in the 315 papers. We found that corpora, exactly parallel, remain the most popular option. The second one is machine translation, despite the fact that the authors of the papers accept the resulting problems of ambiguity. The low computational cost and ease of storage are two of the main advantages of these two resources. In our opinion, these resources can be adequate for IR when combined with others. However, there have been some changes in the use and incorporation of these resources and tools. The three most popular traditional resources (machine translation, dictionaries, and corpora) are gradually leaving a widening gap for others –such as ontologies and the free encyclopaedia Wikipedia. In addition, other approaches such as computational grammars are slowly attracting more researchers who are experienced in handling the results they produce.

A comparison of the evolution and use of different resources and tools shows that trends favour the traditionally more popular tools (automatic translators, dictionaries and corpora). However, ontologies and Wikipedia show trends that match, or nearly match, the traditional

resources. The remaining tools are timidly growing in popularity and have promising futures. However, the trends for each combination of tools in multi-lingual QA systems were not studied exhaustively.

There is a growing trend toward the translation of documents, although the option to translate queries remains the most widely used by researchers. We believe the combination of both approaches is the most useful route and offers the best results – even handling more than two languages without difficulty. Language Resources for Translation in Multi-lingual Question Answering Systems

2.3. QUESTION ANSWERING TRACK EVALUATION IN TREC, CLEF AND NTCIR (PUBLICACIÓN 3)

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2015) "Question Answering Track Evaluation in TREC, CLEF and NTCIR". In: *New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing*, 353 (1): 13-22, Berlín: Springer International Publishing.

ABSTRACT:

Question Answering (QA) Systems are put forward as a real alternative to Information Retrieval systems as they provide the user with a fast and comprehensible answer to his or her information need. It has been 15 years since TREC introduced the first QA track. The principal campaigns in the evaluation of Information Retrieval have been specific tracks focusing on the development and evaluation of this type of system. This study is a brief review of the TREC, CLEF and NTCIR Conferences from the QA perspective. We present a historical overview of 15 years of QA evaluation tracks using the method of systematic review. We have examined identified the different tasks or specific labs created in each QA track, the types of evaluation question used, as well as the evaluation measures used in the different competitions analyzed. Of the conferences, it is CLEF that has applied the greater variety of types of test question (factoid, definition, list, causal, yes/no, amongst others). NTCIR, held on 13 occasions, is the conference which has made use of a greater number of different evaluation measures. Accuracy, precision and recall have been the three most used evaluation measures in the three campaigns.

Keywords: Question Answering Systems, Evaluation metrics, QA Tracks, TREC, CLEF, NTCIR

2.3.1. INTRODUCTION

Frequently, a keyword query entered into a web search tool (search engine or meta-search engine) to satisfy a user's information need, provides too many result pages – many of which are useless or irrelevant to the user. In effect, modern Information Retrieval (IR) systems allow us to locate documents that might have the associated information, but the majority of them leave it to the user to extract the useful information from an ordered list (Dwivedi & Singh, 2013). In

contrast to the IR scenario, a Question Answering (QA) system processes questions formulated into Natural Language (NL) instead of keyword based queries, and retrieves answers instead of documents (Peñas *et al.*, 2012). Therefore, the usefulness of these types of systems for quickly and effectively finding specialized information has been widely recognized (Mollá & Vicedo, 2007; Diekema *et al.*, 2004).

The aim of QA systems is to find precise and correct answers for users' questions. QA systems perform three basic actions: question analysis, passage retrieval, and answer extraction. In the question analysis process, the system locates words that offer some clues about the type of answer we are looking for question words, nouns, adjectives or verbs. The system also seeks text elements and classifies them into categories such as the names of people, organizations, locations, expressions of time, quantities, monetary values, etc. Later, it processes the documents in order to retrieve those parts of text with the highest probability of containing the answer. Finally, it extracts the definitive answer.

The evaluation of QA systems is a major research area that needs attention, especially with the emergence of domain-oriented question answering systems based on natural language understanding and reasoning (Sing *et al.*, 2005). Although we find different analyses referring to the evaluation of QA systems, such as those focusing on the evaluation of QA systems on the Web (Olvera-Lobo & Gutiérrez-Artacho, 2010; Olvera-Lobo & Gutiérrez-Artacho, 2011a; Olvera-Lobo & Gutiérrez-Artacho, 2011b; Olvera-Lobo & Gutiérrez-Artacho, 2011c; Olvera-Lobo & Gutiérrez-Artacho, 2013; Radev *et al.*, 2002), or in some of the international evaluation forums (Peñas *et al.*, 2012; Forner *et al.*, 2010), up to now there has been no broad analysis of the use of evaluation measures of these systems in the main international forums.

Every year the organizers of the major conferences on this issue propose a particular objective for each specific QA track. Then, they choose an appropriate evaluation method which involves determining specific features of the collections, and selecting the measures for assessing the performance of participating systems (Peñas *et al.*, 2012). This study is primarily intended as a general-purpose review, and aims to identify and analyze briefly several aspects of the different QA tracks in TREC, CLEF and NTCIR. We have examined the reports and proceedings of those conferences from when they began up to but not including 2015.

2.3.2. QUESTION ANSWERING IN INFORMATION RETRIEVAL EVALUATION CAMPAIGNS

2.3.2.1. TEXT RETRIEVAL CONFERENCE (TREC)

TREC introduced the first QA track in TREC-8 (1999). During the following 8 conferences –the QA track ran out of in 2007–, the aim of the TREC QA campaigns was to assess the capability of systems to return exact answers to open-domain English questions. The QA track at TREC was the first attempt in IR to stress the importance on these systems (Peñas *et al.*, 2012). TREC workshops consisted of a set of areas of focus in which particular retrieval tasks are defined. The tasks in the track have evolved over the years to focus research on particular aspects of the problem deemed important to improving the state-of-the-art (Voorhees, 2003). Table 14 shows the different tasks, which respond to research concerns and developments in this field, performed between 1999 and 2007 focusing on QA systems.

Task	Characteristics	No. Times performed
Main task	Give the most up-to-date answer supported by the document collection	9
List task	Require systems to assemble an answer from information located in multiple documents	2
Context task	Test systems' ability to track discourse objects (context) through a short series of questions	1
Document ranking	Build infrastructure that would allow a closer examination of the role document retrieval techniques play in supporting QA technology	1
Relationship task	Give to systems topic statements that ended with a question asking for evidence for a particular relationship	1
CiQA (complex task)	Promote the development of interactive systems capable of addressing complex information needs	2
Passages	Find a small chunk of text that contains the exact-phrase answer of a given question from a large document collection	1

Table 14. Tasks in QA Track in TREC

2.3.2.2. CONFERENCE AND LABS EVALUATION FORUM (CLEF)

The first QA Track at CLEF took place in 2003. The systems were fed with a set of questions and were asked to return one or more exact answers per question –where exact means that neither more nor less than the information required is returned. The answer needed to be

supported by the identification of the document in which the exact answer was found, and depending on the year, also by portion(s) of text, which provided enough contexts to support the correctness of the exact answer. Table 24 summarizes all the novelties that have been introduced in the main task over the years of QA campaigns. Each year the tasks proposed were more and more challenging by addressing different types of questions and requiring different types of answer format as output (Forner *et al.*, 2010). The principal difference between the QA tracks of CLEF and TREC is the multilingual component of each of its tasks. The combination of the main languages spoken in Europe in the IR has allowed for the creation of multilingual and crosslingual QA systems.

Task	Characteristics	No. Times performed
Main task	Give the most up-to-date answer supported by the document collection	9
ICLEF (Interactive CLEF)	How best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language	1
QA4MRE (QA for Machine Reading Evaluation)	Focus on the reading of single documents and the identification of the answers to a set of questions about information that is stated or implied in the text	3
WIQA (QA using Wikipedia)	Given a source page from Wikipedia, identify snippets from other Wikipedia pages, possibly in languages different from the language of the source page, that add new and important information to the source page, and without repetition	2
AVE (Answer Validation Exercise)	Decide for given a question and an answer from a QA system, whether the answer is correct or not and it was defined as a problem of recognizing textual entailment in order to promote a deeper analysis in QA	4
QAST (Question Answering on Speech Transcriptions)	Evaluate the task of QA in Speech Transcripts. Accessing information in spoken documents provides additional challenges to those of text-based QA, needing to address the characteristics of spoken language, as well as errors in the case of automatic transcriptions of spontaneous speech.	3
WSD (Word Sense Disambiguation)	Explore the contribution of Word Sense Disambiguation to QA providing document collections and topics which have been automatically tagged with Word Senses from WordNet.	1
RespubliQA	Evaluation task over European legislation	2
Others		5

Table 15. Tasks in CLEF QA Track

2.3.2.3.NTCIR CONFERENCE

Since 2001, NTCIR workshop, co-sponsored by Japan Society for Promotion of Science, has performed specific tracks for the development and evaluation of QA systems. As with CLEF, NTCIR has permitted the evaluation of systems of QA multilingual and crosslingual systems, with the principal difference that the languages used were predominantly found in Asia. Table 3 shows the tasks performed from 2001 to 2010. Although in recent years it had been decided not to have a specific track for QA, in the last NTCIR workshop, NTCIR-II, which took place in December 2014, they resumed evaluation of QA systems with a pilot task called “QA Lab”, focusing on module-based platforms for system performance evaluations and comparisons for solving real-world university entrance exam questions.

Name of task	Characteristics	No. Times performed
Main task: 3 parts	Give the most up-to-date answer supported by the document collection	2
CLQA (Cross-lingual QA)	Promote research on cross-lingual QA technology mainly for East Asian languages	2
QAC (QA Challenge)	Obtain appropriate answers to given domain independent questions written in NL from a large corpus	4
IR4QA (Information Retrieval for QA)	Evaluate traditional ranked retrieval of documents using well-studied metrics	2
CCLQA (Complex Cross-Lingual QA)	Evaluate QAS on complex	2
QA-Lab Task	Provide a module-based platform for advanced QAS and comparative evaluation for solving real-world university entrance exam questions	1

Table 16. Tasks in NTCIR

The evolution of this type of QA track in each competition has been different, CLEF being the conference which has for longest performed specific QA tracks for the evaluation of QA systems. The different QA tracks lead to the opportunity to experiment how the systems and technologies of QA systems respond in different scenarios.

2.3.3. TEST QUESTIONS

With regard to the number of the test questions that have had to be used by participants in the three campaigns herein analyzed so as to evaluate their QA systems, significant differences are seen. While at TREC the number of evaluation questions has varied significantly from one session to another, at both CLEF and NTCIR a stable number has been maintained (200 in the former, and 100 or 300 in NTCIR). Furthermore, in TREC, possibly due to being the first to evaluate QA systems, it is the case that not all the questions proposed to the participants are finally accepted as valid. Indeed, at TREC, at its earliest conferences, questions were rejected for not offering answers with the necessary requirements or for not being within the proceedings. Given the multilingual and crosslingual tracks of CLEF and NTCIR, the test questions of these conferences have been translated into all the languages involved. Table 26 gives the average number of questions used on each occasion of each of these conferences for the evaluation of QA systems on all the occasions of each of these conferences.

Conference	Average of questions
TREC	481
CLEF	300
NTCIR	224

Table 17. Average of test questions per conference

The evaluation campaigns use diverse types of questions (factoid, definition, list, causal, yes/no, among others) which are best adapted to the peculiarities of the tasks of each occasion of the conference. Figure 36 shows the questions proposed by the organizers of the three conferences according to the name they use, ordered by frequency.

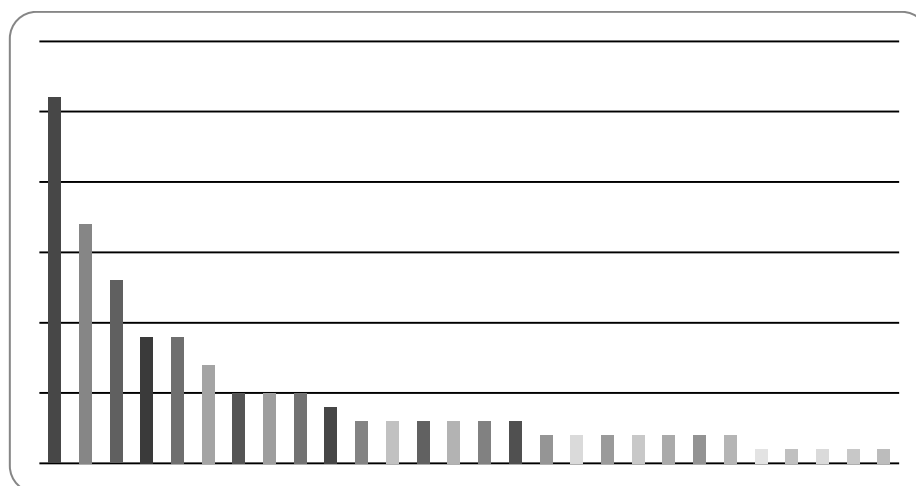


Figure 36. Frequency each type of test questions is been used.

As can be seen, the most used type of question is the factoid question (that is, fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc.), followed by definition questions and list questions, respectively. In addition, CLEF is the conference which has applied a greater variety of types of test question (see Table 18), although, as has been stated, it has also been the international forum which has organized the most tracks on the evaluation of QA systems.

Conference	Total of type of questions
CLEF	18
NTCIR	15
TREC	7

Table 18. Type of test questions per conference

2.3.4. ANSWER ASSESSMENT

Organizers of the various evaluation campaigns are sent by participants the results given by the latter's systems to the questions put. While at NTCIR the parameters established for the evaluation of the answers have been changed, at the other two conferences the same criteria for all the tasks have been maintained, save for a few exceptions. All answers were judged by native language human assessors, who assigned to each response a unique judgment following a schema pre-established.

In TREC the process was begun with a binary evaluation of the questions (0-correct and 1-incorrect) in order to apply thereafter the already traditional evaluation of correct/right, incorrect/wrong, not exact or not support. The latest QA tasks saw the introduction of the dichotomy "globally correct" or "locally correct" for the correct questions. CLEF has followed the same model as TREC, also used by NTCIR on a few occasions, with the following classification: right (R), wrong (W), un-support (U) or inexact (X), except in AVE tasks (Answer Validation Exercise) where the answers have been evaluated as validated, unknown or rejected.

2.3.5. EVALUATION MEASURES

Several evaluation measures have been used in the QA tasks. In each competition a main measure was selected to rank the results of the participating systems, while several additional measures were adopted in order to provide more information about the systems' performances (Peñas *et al.*, 2012).

After having identified the different evaluation measures proposed in the tasks to be used in the evaluation of the QA systems, it can be seen that around 50 evaluation measures of very different types have been used (see Table 19).

	CLEF	TREC	NTCIR	TOTAL (not repeated)
No. Measures	28	18	18	48

Table 19. Measures used by the conferences

As seen in Table 20, it is the traditional measures of IR (precision, recall, accuracy) which continue to be the most used. However, we find variances in the different tracks since the measures are applied so that they are suited best to the particularities of each task and the evaluation needs of each competition. Another of the commonly used measures is MRR (Mean Reciprocal Rank) proposed by TREC, which is very useful for the evaluation of QA systems as it makes it possible to take into account all the answers retrieved by the system and to assign it a reciprocal value in accordance with the ranking of the system. The traditional F-measure, which combines with recall and precision, is also prominent.

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

Measure	Frequency of Measure Uses	Frequency by Conferences		
		TREC	CLEF	NTSIR
Accuracy	26	7	17	2
Overall Accuracy	6	-	6	-
QA rejected accuracy	1	-	1	-
Random QA Accuracy	1	-	1	-
QA Accuracy Max	1	-	1	-
Precision	22	7	10	5
Overall Precision	2	-	2	-
Instance Precision	5	5	-	-
MAP (Mean Average Precision)	1	1	-	-
R-Precision	1	1	-	-
Average Precision	3	1	-	2
Recall	21	8	9	4
Overall Recall	1	-	1	-
Instance Recall	5	5	-	-
MRR (Mean Reciprocal Rank)	15	3	8	4
F	13	6	7	-
Overall F	1	-	1	-
Average F-measure (AFM)	2	-	-	2
Modified F measure(MF)	2	-	-	2
Mean F	6	-	-	6
Confidence Weighted Score (CWS)	9	4	5	-
Final Score	7	7	-	-
Nugget pyramid	7	4	-	3
C@1	5	-	5	-
Ave Length	5	5	-	-
K1	4	-	4	-
Exact match	3	-	-	3
Softmatch	3	-	-	3
aBinarized	3	-	-	3
Nugget recall	3	-	-	3
Q	2	-	-	2
Top 5	2	-	-	2
Ndcg (Normalized Discounted Cumulative Gain)	2	-	-	2
Error rate	2	2	-	-
BA-HIT@1	1	-	-	1
P@N	1	-	1	-
Yield	1	-	1	-
Px	1	-	-	1
CAS (Correct answer set)	1	-	-	1
UR /t	1	-	-	1
MRC (Mean Reciprocal Cost)	1	-	-	1
Px	1	-	-	1
FHS (First Hit Success)	1	1	-	-
Qrels	1	1	-	-
CR (Correct Test)	1	-	-	1
PMM (Population Marginal Mean)	1	1	-	-

Table 20. Frequency each evaluation measure used

As stated, each conference has its own objectives, tradition and history. This entails that not all the evaluation measures are equally applicable to the distinct tracks of the different campaigns analysed. However, space could have been found for some of those measures at several of those conferences. That is the case with Instance Precision, Mean Average Precision and R-Precision, which have only been applied in TREC. There is a similar situation with Error Rate, with the measure which determines the average length of the answers (Ave Length) and con FHS (First Hit Success), amongst others.

On the other hand, possibly due to the fact that CLEF is a campaign with more delimited objectives, that is to say, essentially oriented at the evaluation of systems types, the multilingual and the crosslingual, this conference introduces some specific measures albeit some of them with notably reduced frequency. For example, in relation to the measure Accuracy, others such as Overall Accuracy, QA rejected accuracy, Random QA Accuracy and QA Accuracy Max have been proposed.

However, NTSIR is the conference with the greater number of specific measures (exact match, softmatch, aBinarized, Normalized Discounted Cumulative Gain, Correct answer set, amongst others) implemented in the evaluations of its tracks, and while some are found in up to six campaigns, the majority have been applied infrequently.

2.3.6. CONCLUSIONS

Evaluation of Information retrieval evaluation is a constantly evolving field (Kelly & Sugimoto, 2013; Moghadasi & Ravana, 2013). In this study we have set out the QA tasks or specific labs organized in the three principal IR evaluation campaigns over the last 15 years. Likewise, the type of test question used and the main evaluation measures proposed have been identified.

The new scenario raised by the appearance of novel QA systems constitutes an excellent benchmarking to test the relevance of the traditional evaluation and to propose and to adapt new methods and measures appropriate for this environment.

2.4. QUESTION-ANSWERING SYSTEMS AS EFFICIENT SOURCE OF TERMINOLOGICAL INFORMATION: EVALUATION (PUBLICACIÓN 4)

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2013) "Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general". *Revista Española de Documentación Científica*, 36(2), abril-junio 2013, e009 ISSN-L:0210-0614

ABSTRACT

Question-answering systems (or QA Systems) stand as a new alternative for Information Retrieval Systems. Most users often need to retrieve specific information about a factual question in order to obtain a whole document. We conducted a study to evaluate the efficiency of QA systems as terminological sources for physicians, specialized translators, and users in general. To this end we analyzed the performance of one open-domain QA system, START, and one restricted-domain QA system, MedQA. The research entailed a collection of two hundred definitional questions (What is...?), either general or specialized, from WebMed. We studied the sources that QA systems used to retrieve the answers, and later applied a range of evaluation measures to mark the quality of answers. Both QA systems were determined to be appropriate for the retrieval of terminology, proving reliable as sources and supplying correct answers.

Keywords: Decision support techniques, evaluation studies as topic, information storage and retrieval, natural language processing, MedQA, Start.

2.4.1. INTRODUCTION

Question-answering systems (heretofore QA Systems) can be viewed as a new alternative to the more familiar Information Retrieval Systems. These systems try to offer detailed, understandable answers to factual questions, in order to retrieve a collection of documents related to a particular search (Jackson and Schilder, 2005). In recent years, the development of QA systems has been encouraged and furthered through the TREC meetings (Text Retrieval Conference) — mainly since TREC-8 (Voorhees, 1999). This Conference has proven to be an important

international forum, putting together and improving research efforts behind the different aspects of information retrieval.

We have carried out a research to evaluate the quality and efficiency of two question-answering systems as terminological sources for physicians and users in general. We have analysed an open-domain QA systems, START⁵⁷ which is developed at the Massachusetts Institute of Technology, and a restricted-domain QA systems, MedQA⁵⁸ which is specialized in the biomedical domain and developed by Columbia University. So, we have obtained a collection of two hundred definitional questions (What is...?), either general or specialized, from WebMed, and we have asked in the two QA systems (Methodology). We have studied all the sources used by the both QA systems and applied different evaluation measures to mark the quality of answers. This evaluation has allowed us to obtain informational data about the sources and their efficiency (Results). Finally, in Conclusions, we have realized START and MedQA are appropriate for retrieval terminological information and for supply correct answers.

2.4.2. BACKGROUND

The QA systems try to make retrieval easier through the short-answer question models (Voorhees, 1999; Blair-Goldensonh and Schlaikjer, 2004). Accordingly, users do not have to read the full text of documents such as a web page, an article of scientific journal, etc., in order to arrive at the information they need because the QA system shows the correct answer by means of a number, a noun, a short phrase or a concise extract of text.

The questions used in QA systems can be expressed using interrogative adverbs (who, what, which, how, when, where), or in imperative form (tell me, show, list...). Once the question is provided, the QA systems extract natural language answers, to be comprehended in a natural way for humans (Costa and Santos, 2007). QA systems follow these main steps: first of all, the systems retrieve the documents to obtain relevant sentences about the search term; dealing with the questions posed by the users, they identify their components parts and then determinate the kind of answer anticipated (Zweigenbaum, 2004); they retrieve and select the sentences; and finally, they choose non-redundant definition sentences from the overall results of sentence retrieval, to delimit the response (Cui *et al.*, 2004; Mollá and Vicedo, 2005). The objective pursued is for the systems to retrieve only the correct information to answer the users'

⁵⁷ Access to START (Natural Language Question Answering System). Available from: <http://start.csail.mit.edu/>

⁵⁸ Access to MedQA. Available from: <http://monkey.ims.uwm.edu:8080/MedQA/>

questions (Tsur, 2003). Evaluation is one of the most important dimensions in QA systems, as the process of assessing, comparing and ranking is key to monitoring progress in the field (Sing *et al.*, 2005; Fahmi, 2009). The main component of these systems consists of measuring modules, which analyze the tagged sentences in the documents selected, and compare them with the question in order to find the most similar sentence (Alfonseca *et al.*, 2002; Jacquemart & Zweigenbaum, 2003). Generally speaking, QA systems feature very simple and user-friendly interfaces, and rely on methods of linguistic analysis and natural language processing in the different phases of operation. The ones that allow users to query in different languages are known as multilingual QA systems.

All of these QA systems are based on prototypes; that is, they are available as demos, and only in a few cases have been marketed. A further problem can be found in the design of these systems: what is needed is a more interactive QA procedure that allows for real feedback between questions and answers, and user communication with the system on a conversational level.

While not many QA systems are available on the Internet, we do have some open-domain QA systems such as START, it is a very atypical which includes calls to OMNIBASE, a system that integrates heterogeneous data sources using an object-property-value model (Katz *et al.*, 2002); NSIR⁵⁹, developed by the University of Michigan; or Qualim⁶⁰, financed by Microsoft; in addition to some restricted-domain QA systems including MedQA. In the case of NSIR and Qualim, answers are constructed on the basis of information provided by Google⁶¹ and Wikipedia⁶², respectively. Although START also retrieves information from Wikipedia, it uses other specialized sources such as directories, databases, dictionaries, or encyclopaedias. Meanwhile, MedQA retrieves information from the medical database Medline, specialized dictionaries, Wikipedia and certain search engines like Google.

In the Web setting, the overload of information may be perceived more acutely than in other contexts. When users pose a given question by means of search engine tools (including directories or metasearchers), the systems tend to retrieve an excessive number of web pages, many of which are not relevant or useful in light of the users' needs. Professionals in different areas claim that QA systems constitute a good method to obtain specialized information in quick and efficient manner (Crouch *et al.*, 2005; Lee *et al.*, 2006; Yu *et al.*, 2007).

⁵⁹ Access to NSIR (Question Answering System). Available from: <http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi/>

⁶⁰ Access to QuaLiM (Question Answering Demo). Available from: <http://demos.inf.ed.ac.uk:8080/qualim/>

⁶¹ Access to Google. Available from: <http://www.google.com/>

⁶² Access to Wikipedia. Available from: <http://www.wikipedia.org/>

In a study by Ely (1999), participating physicians spent on the average less than two minutes looking for information to resolve clinical queries, although many of their questions remained unanswered. Regarding this point, some researchers have shown that the physicians trust QA systems as search methods for specialized information retrieval (Lee *et al.*, 2006; Yu *et al.*, 2007).

The general public increasingly consults knowledge resources like the Web as well: before or after seeing a doctor, for themselves or for relatives, to obtain information about the nature of a disease, the indications and contraindications of a treatment, etc. (Zweigenbaum, 2004)

While researchers have looked into various aspects of QA systems in recent years, one facet that is widely overlooked is the formal evaluation of this tool and the results it supplies. Indeed, no study to date has focused specifically on the information sources from which responses are derived. This is the main aim of our line of research. Ideally, QA systems should create coherent definitions in a dynamic way, and ones that contain and summarize the most descriptive information contained in a document collection, in view of the specific term or focus of the user query. (Zweigenbaum, 2004; Blair-Goldensohn *et al.*, 2003).

Our objective led us to use definition-type questions in order to evaluate two QA systems and determine the different sources behind the retrieval of medical information. In the sections below we describe the questions used, the QA systems analyzed and the measures of evaluation applied. Finally, we show our results and briefly expound some conclusions.

2.4.3. METHODOLOGY

We took a sample of two hundred definitional questions about different medical issues as the basis of this study. The questions were obtained from the webpage WebMD⁶³, a US health portal providing valuable health information and support, tools for managing health problems, and specialized background on a number of illnesses. It was created by health specialists who aspire to explain, briefly yet credibly, in-depth medical information, reference material, and online community programs.

The collection of two hundred questions was created using the expression “What is...? (i.e. what is irritable bowel syndrome?) in the internal search engine of the website; and in turn, WebMD provided a list of some 6000 responses in their characteristic question-answer format. We chose around 250 factual questions about different health issues, specified in Table 30. It was not our intention to evaluate the coverage of the databases sources of QA systems START

⁶³ Access to WebMD. Available from: <http://www.webmd.com/>

and MedQA, but merely to appraise how they work and what sources they retrieve data from. This led us to finally choose 200 questions to be answered by both systems. Our classification was based on the most generic questions of the taxonomy developing by Ely *et al.* (2003) about generic clinical questions (Table 21).

Question	Pain	Inflammation	Disease	Syndrome	Infection	Treatment	Others
Number	8	16	97	11	10	38	15

Table 21. Categories of reference of definitional questions

START, a QA system allowing users to pose questions about various health issues, can respond to even very specialized questions within the area of health care (Katz *et al.*, 2002). It has a dynamic yet easy interface, and responds quickly. Information is retrieved from a very wide list of sources, such as World Book, The World Factbook 2008, START KB, Internet Public Library, and many others.

Meanwhile, MedQA is a specialized QA system that analyses thousands of documents to arrive at a coherent response. Because it works specifically in the area of health care, its sources are more specialized (Yu *et al.*, 2007). It also has a user-friendly interface, but it is slower than START. It retrieves information from a wide array of sources, including Wikipedia, Medline or Medline Plus.

After presenting the questions to both QA systems, we analyzed and evaluated the answers obtained, and identified the source or sources used by the system. Answers were marked as: incorrect (0 points), inexact (1 point) or correct (2 points), according to the guidelines of CLEF. To be judged as correct, the answer had to respond accurately to the question asked, not use more than 100 words in its response, and not contain irrelevant information. All the questions that were answered correctly yet did not fulfil these criteria were considered inexact (Cao *et al.*, 2009). Likewise, we recorded the response time and the partial or total repetitions of information by the systems. The mark obtained by each question was the baseline for application of further evaluation measures, drawn from a 2001 study by Raved (2001).

Mean Reciprocal Rank (MRR) is a statistical tool for evaluating any process that produces a list of possible answers to a query. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer (for example, if a question gets the correct answer in the 1st place, it will receive a score of 1, it would be ½ if it is in the 2nd place, 1/3 in the 3rd place...). If the answer is not found, a score of 0 is assigned. MRR can be used with several correct answers, but it only takes into account the first correct answer found.

Total Reciprocal Rank (TRR) is useful when there is more than one correct answer to a question. In these cases, it is not sufficient to consider the first correct answer in evaluations; instead, TRR takes into consideration all the correct answers and assigns a weight to each according to its ranking in the list provided by the system. For example, if the QA system provides two correct answers (the first and the third ones), the TRR will be $1/1 + 1/3$.

First Hit Success (FHS) assigns 1 if the first answer returned by the system is correct, and 0 if it is not. This measure, then, only accepts the first questions in the list of results.

We used the measurement of “precision” in the evaluation of information retrieval. It is understood as the capacity of system to retrieve documents or answers (in the case of QA systems) relevant to the query and well ranked (in the case of systems ranking the results).

2.4.4. RESULTS

After posing 200 questions in our QA systems, we identified the sources used by them to obtain the answers. START provided answers to the medical questions from six sources, appearing in this order: Wikipedia, American Medical Association (the only specialized source used by START), The Internet Movie Data Bases, Webopedia.com, Yahoo and Merriam Webster Dictionary.

Very briefly, Wikipedia is a widely used online encyclopaedia able to offer information about different issues in several languages. The website American Medical Association (Cao *et al.*, 2009) offers useful information about health for patients and physicians. The Internet Movie Database (IMDb)⁶⁴ is an American movie site, available in some languages, with data about movies, series and actors from all over the world. Yahoo⁶⁵ is a directory that categorizes web pages under different subjects. Webopedia.com⁶⁶ is an online computer dictionary and internet search engine for internet terms and technical support. And finally, the Merriam-Webster Dictionary⁶⁷ is a free dictionary and thesaurus more strictly speaking, with definitions, etymology, pronunciation, etc. for each entry.

The source that offered more answers was Wikipedia, with a total of 182. Second was Merriam Webster Dictionary with 84 answers –although 31 of these answers repeated exactly the same information, for which reason we rejected them. American Medical Association, the

⁶⁴ Access to Internet Movie Database (IMDb). Available from: <http://www.imdb.com/>

⁶⁵ Access to Yahoo. Available from: <http://www.yahoo.com/>

⁶⁶ Access to Webopedia. Available from: <http://www.webopedia.com/>

⁶⁷ Access to Merriam-Webster. Available from: <http://www.merriam-webster.com/>

only specialized source, gave 36 answers. The sources providing the fewest answers were The Internet Movie Data Base (IMDB), Yahoo and Webopedia.com, with 5, 2 and 1 answers obtained, respectively.

Sources	Answers obtained
Wikipedia	182
Merriam-Webster Dictionary	84 (31 repetitions)
American Medical Association	36
IMDB	5
Yahoo	2
Webopedia.com	1
Total	310

Table 22. Sources used by START

In evaluating the quality of the results by the START sources (Table 22), Wikipedia was found to be the source giving more correct answers (104), with 42 answers that were inexact and 36 others that were incorrect. Some of the inexact answers pointed to an intermediating “window” of sorts with several options related with the query. The general dictionary Merriam-Webster Dictionary offered 45 correct answers, 7 inexact ones and only one incorrect answer. The American Medical Association supplied just one correct answer and 35 inexact answers. The only response obtained through Webopedia.com was considered correct, whereas all the answers of IMDB and Yahoo were incorrect.

START's reply

==> what is whiplash?

[Whiplash \(TV series\)](#)

Whiplash is a British/Australian [television](#) series made by the [Seven Network](#) and [ATV](#) and [ITC Entertainment](#). Filmed in 1959-60, the series was first broadcast September 1960 and had opening titles featuring the Australian locale and terrain and a dozen wild kangaroos as a Cobb & Co stage passed pulled by a team of five horses driven by Cobb himself.

I know about 13 more terms called "Whiplash": [Whiplash \(band\)](#), [Whiplash \(1948 film\)](#), [Whiplash \(album\)](#), [Whiplash \(comics\)](#), [Whiplash \(fetish magazine\)](#), [Whiplash \(Gladiators\)](#), [Whiplash \(Law & Order episode\)](#), [Whiplash \(Masters of the Universe\)](#), [Whiplash \(medicine\)](#), [Whiplash \(song\)](#), [Whiplash \(Stellar* song\)](#), [Whiplash \(video game\)](#), and [Whiplash](#)

Source: [Wikipedia](#)

Figure 37. The intermediating ‘window’ of an inexact answer

The number of answers retrieved by MedQA was higher than for START, and most sources were of a specialized nature. Medline answered all the questions. This bibliographic database created by the U.S. National Library of Medicine includes citations and specialized articles from approximately 5000 selected journals, from 1966 to the present.

Sources	Answers obtained
Medline	200
Dictionary of Cancer Terms	192
Wikipedia	191
Google	174 (34 repetitions)
Dorland's Illustrated Medical Dictionary	143
Medline Plus	105
Technical and Popular Medical Terms	29
National Immunization Program Glossary	3
Total	1037

Table 23. Sources used by MedQA

The Dictionary of Cancer Terms⁶⁸ (created by the U.S. National Institute of Cancer) and Wikipedia offered 192 and 191 answers, respectively. Google is appraised by previous authors as one of the best sources for answering definitional questions (Yu & Kaufman, 2007), this search engine offered 174 answers in our experience, though 34 were rejected as repetitions. Dorland's Illustrated Medical Dictionary⁶⁹, another non-free dictionary for health issues, gave 143 answers. Medline Plus⁷⁰, as a multilingual medical portal with information about medication, disease and other health issues, features a medical encyclopaedia, tutorials and videos for patients; it gave us 105 answers. The multilingual glossary of Technical and Popular Medical Terms⁷¹, set up by The European Commission and executed by Heymans Institute of Pharmacology and Mercator School, provided 29 results. The National Immunization Program Glossary⁷² of the U.S. Department of Health & Human Services supplied just 3 answers.

The two QA systems evaluated here gave similar figures for repeated answers (31 repetitions in START and 34 in MedQA). In START, all the repetitions were exactly identical, and came from the same sources (Merriam-Webster Dictionary). In MedQA, the repetitions offered more or less the same answer, but their sources were different (Wikipedia and Google). Although a question may harvest different yet equally valid answers at a given time, when the same answer

⁶⁸ Access to Dictionary of Cancer Terms. Available from: <http://www.cancer.gov/dictionary/>

⁶⁹ Access to Dorland's Illustrated Medical Dictionary. Available from: <http://www.dorlands.com/wsearch.jsp/>

⁷⁰ Access to MedlinePlus. Available from: <http://www.medlineplus.gov/>

⁷¹ Access to Glossary of Technical and Popular Medical Terms. Available from: <http://users.ugent.be/~rvdstich/eugloss/welcome.html/>

⁷² Access to National Immunization Program Glossary. Available from: <http://www.cdc.gov/vaccines/about/terms.htm/>

is repeated, users tend to feel confused, and the list of results increases unnecessarily. This is why we “penalized” the QA systems by not considering these answers as valid.

Source	Correct	Inexact	Incorrect
Google	122	26	26
Wikipedia	117	31	43
Medline Plus	95	1	9
Dictionary of Cancer Terms	51	0	140
Technical and Popular Medical Terms	21	3	5
Dorland’s Illustrated Medical Dictionary	14	94	35
Medline	12	61	127
National Immunization Program Glossary	2	0	1
Total	434	216	386

Table 24. Answers provided by MedQA

As we see in Table 24, there were five sources providing more correct answers than inexact or incorrect ones: these were Medline Plus, Wikipedia, Google, Technical and Popular Medical Terms and National Immunization Program Glossary. The only source supplying a majority of inexact answers was Dorland’s Illustrated Medical Dictionary, which remitted irrelevant information about Dorland’s itself (copyright, edition and other non-pertinent information) in most responses. Medline and the Dictionary of Cancer Terms gave more incorrect answers, and this dictionary sometimes offered irrelevant or incorrect information. Medline is a bibliographical database, and it rarely showed definitions about specific terms, but instead supplied extracts from studies (or abstracts) by health specialists or other researchers. Thus, we may infer that the questions were not expressed in the best possible terms. This is due to MedQA was specifically designed and evaluated on definitional question-answering.

Calculation of the time of response (time elapsing before appearance of results on the screen) for each question led us to some interesting findings. The values obtained were quite different for the two systems: the average response time for START was 2 to 4 seconds, while MedQA was considerably slower –with a minimum of 10 seconds and a maximum of 135 seconds. Overall, nearly 50% of the queries were solved in a period between 26 and 35 seconds (Figure 1). During the wait, MedQA tells users that operations are underway at that moment –first of all, the system looks over Google, then in Medline, and finally, it removes all the redundant answers to generate the coherent ones.

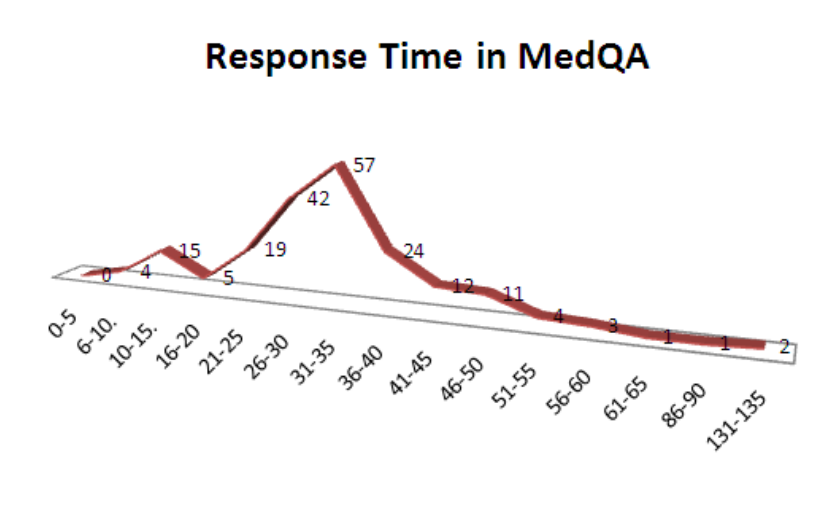


Figure 38. Analysis of frequencies according to the response time in MedQA

In identifying the sources used by the two systems, we applied specific measures for the evaluation of information retrieval. Table 25 indicates that the average number of answers retrieved for each question is considerably higher with MedQA (5.2) than with START (1.41). Moreover, MedQA gave, on the average, more correct responses per question, 2.17, as compared with the 0.94 of START. This finding comes to confirm that the more specialized system offers a more adequate coverage by subject for the sort of query collection used here; and aside from the greater yield of responses provided by MedQA, the average offerings of incorrect and inexact responses are also greater under this system (1.93 and 1.08, respectively) than with the general-domain system START (0.22 incorrect and 0.25 inexact ones).

	Average answers retrieved per question	Average correct answers per question	Average incorrect answers per question	Average inexact answers per question	MRR	FHS	TRR	Precision*	Precision†
MedQA	5.18	2.17	1.93	1.08	0.86	0.75	0.40	42%	63%
START	1.41	0.94	0.22	0.25	0.60	0.61	0.59	67%	84%

MRR, Mean Reciprocal Rank; TRR, Total Reciprocal Rank; FHS, First Hit Success

*Taking only correct responses into account.

†Taking both correct and inexact responses into account.

Table 25. Measures for evaluating the quality of answers

As we explained in the section on Methods, MRR calculates the inverse value of the first correct answer, whereas FHS simply evaluates if the first answer was correct or not. The two measures show us, in this case, that MedQA ranks their results more adequately, because the first correct answer tends to appear in the first place of the list (more frequently than with START). This proves very important, as no algorithm is involved in the ranking process. These systems, then,

maintain the ranking of answers as determined by the source they came from. In terms of user-friendliness, FHS might be a somewhat more realistic or convenient measure, because users usually focus on the first answer retrieved.

The measure TRR is lower in MedQA, however. This figure takes into account not just the first one but all the correct responses supplied by the system, and weights the value of the correct response in light of its placement within the list of results. Since MedQA provides a greater amount of results, the correct responses in the lower positions of the ranking receive less weight, and the TRR drops with respect to that of the START, which consistently yielded fewer responses.

Finally, we assessed the precision of the two systems. The value obtained for START precision was higher (67% relevant responses) than for MedQA (42%). The percentages increased if the inexact answers were also included as relevant (84% with START and 67% for MedQA). Therefore, we may affirm that the more specialized system produces a greater degree of documental noise –that is, that the correct responses are accompanied by numerous incorrect and/or inexact one.

2.4.5. DISCUSSION

Results obtained for the two systems analysed, START and MedQA, allowed us to evaluate their effectiveness and their use of different information sources. Despite certain limitations on the part of both systems (a lack of accessibility for the general public, and insufficient development in some specific areas), we were able to confirm that both are very useful in the retrieval of valid definitional healthcare information, with responses from both proving coherent and precise to an acceptable degree. They also help in understanding the information collected and are set to become one of the key tools available to index and organise health information.

As one might expect, the answers supplied by MedQA were more reliable than those of START in the sense that they came from specialised clinical or academic sources, and gave links to research articles on the subject in hand.

Another interesting finding is that the responses do not appear under a truly representative ranking of relevance, but rather, with both systems, results are shown in a pre-established order according to the source. The systems give priority in the display of results to sources that consistently provide answers (like Wikipedia or Google), regardless of the reliability and

credibility that should be demanded of scientific information. Notwithstanding, we did observe that MedQA always makes use of Medline in responding to queries, which can be interpreted as a sign of reliability, yet not necessarily of precision.

Results are encouraging in that they point to the potential for this type of tool in the more general realm of information access. They are a good, reliable and reasonably precise alternative to help with information overload. They provide concrete results quickly and easily, enabling users to spend less time in the retrieval information. Recent studies (Sing *et al.*, 2005; Buitelaar *et al.*, 2008; Cruchet *et al.*, 2009) have explored various possible means of enhancing the performance of such QA systems, for instance through the incorporation of ontology, which would heighten the quality of the answers obtained by structuring, inter-relating and formalising all relevant information from the thematic domain. In addition, other approaches such as computational grammars are slowly attracting experienced researchers in handling the results they produce.

This data suggests that we may see unexpected changes in the future. This area deserves to be studied and evaluated in future research.

2.4.6. CONCLUSIONS

Health information and libraries need current terminological information to organise and index information. Different studies in Information Retrieval have shown that QA systems are a useful tool for retrieving information quickly and accurately. In this study, we have investigated the effectiveness of these systems in the retrieval of health information, and the main differences between an open-domain QA system, like START, and a restricted-domain QA system, like MedQA.

Question-answering systems may provide a different way for physicians and users in general to seek biomedical information and identify tools to limit human work.

2.5.OPEN- VS. RESTRICTED-DOMAIN QA SYSTEMS IN THE BIOMEDICAL FIELD (PUBLICACIÓN 5)

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011) "Open- vs. Restricted- Domain Question Answering Systems in the Biomedical Field". *Journal of Information Science*. 37 (2): 152-162

ABSTRACT

Question answering systems (QA systems) stand as a new alternative for information retrieval systems. We conducted a study to evaluate the efficiency of QA systems as terminological sources for physicians, specialized translators and users in general. To this end we analysed the performance of two open-domain and two restricted-domain QA systems. The research entailed a collection of 150 definitional questions from WebMed. We studied the sources that QA systems used to retrieve the answers, and later applied a range of evaluation measures to mark the quality of answers. Through analysing the results obtained by asking the 150 questions in the QA systems MedQA, START, QuALiM and HONqa, it was possible to evaluate the systems' operation through applying specific metrics (MRR, FHS, TRR, Precision, Recall). Despite the limitations demonstrated by these systems, it has been confirmed that these four QA systems are valid and useful for obtaining definitional medical information in that they offer coherent and precise answers.

Keywords: biomedical information; evaluation; open-domain question answering systems; question answering systems; restricted-domain question answering systems

2.5.1. INTRODUCTION

Information retrieval (IR) is a discipline focused on the problems of information items' selection from a storage system in order to facilitate retrieval for the users' needs (Baeza-Yates & Ribeiro-Nieto, 1999; Korfhage, 1997; Salton, 1970; Salton & Mc Gill, 1983; Rijsbergen, 1979). Traditionally, IR is understood as a fully automatic process that responds to a user query by examining a collection of documents and returning a sorted document list that should be relevant to the user

requirements as expressed in the query (Baeza-Yates and Ribeiro-Nieto, 1999). Simply stated, it could be said that retrieval implies finding certain requested information in a storage system or database of information (Meadow, 1979). An optimal IR system recovers all the relevant documents (implying an exhaustive search, i.e. a high recall) and only the relevant documents (implying perfect accuracy, that is to say, a high precision). This traditional model involves many implied restrictions: the assumption that users want full-text documents, rather than answers, and that the query will be satisfied with these documents; that the process is direct and unidirectional rather than interactive; and finally, that the query and document share the same language.

Information overload is felt more strongly on the web than elsewhere. All too often a query made with a web search tool (search engine or meta-search engine) results in the retrieval of too many pages – many of which are useless or irrelevant to the user. Question answering systems (QA systems) are an evolutionary improvement in IR systems. As alternative traditional IR systems, they give correct and understandable answers to factual questions (Pérez-Coutiño *et al.*, 2004) – rather than just offering a list of documents related to the search. The benefit is that users do not have to read whole documents to find the desired information. Therefore, professionals from various areas are beginning to recognize the usefulness of these systems, for quickly and effectively finding specialized information (Crouch *et al.*, 2005; Lee *et al.*, 2006; Yu *et al.*, 2007).

We conducted a study to evaluate the efficiency of QA systems as terminological sources for physicians, specialized translators and users in general. To this end we analysed the performance of two open-domain QA systems, START and QuALiM, and two restricted-domain QA systems, MedQA and HONqa. The research entailed a collection of 150 definitional questions (What is...?), either general or specialized, from WebMed. We studied and analysed the performance of each QA system through to the answers retrieved by them in order to assess their efficacy in the biomedical field, using multiple evaluation measures.

2.5.2. OPEN- VS RESTRICTED-DOMAIN QA SYSTEMS ON THE WEB

QA systems have attracted major attention since the TREC-8 (Text Retrieval) conference on information retrieval (Voorhees, 1999). They are based on short-answer models (Blair-Goldensohn *et al.*, 2004) and their aim is to find an exact and correct answer – in the form of a number, a noun, a short phrase or a brief piece of text (Pérez-Coutiño *et al.*, 2004) – for users'

questions. Analysis of question, search and choosing an answer are three important issues in a QA system so it includes at least the three following processes: question processing, document processing and answer processing (Kangari *et al.*, 2008).

Although there are various templates for making queries in QA systems, most of these systems understand questions expressed with interrogative particles (who, what, where, why, when and how), while some understand the imperative form (tell me). Then QA systems proceed to do coherent questions according to natural language (Costa & Santos, 2007). When a query is entered into the interface, the system proceeds to analyse the question by separating the words or keywords. The system then locates and extracts one or several answers from different sources of information, depending on the question's specialized area (Olvera-Lobo & Gutiérrez-Artacho, 2010). Subsequently, the system evaluates and eliminates redundant information, or information that does not respond correctly to the question, and submits one or more prepared responses to the user (Cui *et al.*, 2004; Tsur, 2003). Evaluation is one of the most important dimensions in QA systems, as the process of assessing, comparing and ranking is key to monitoring progress in the field (Sing *et al.*, 2005). The main component of these systems consists of measuring modules, which analyse the tagged sentences in the documents selected, and compare them with the question in order to find the most similar sentence (Alfonseca *et al.*, 2002; Jacquemart & Zweigenbaum, 2003). Generally speaking, QA systems feature very simple and user-friendly interfaces, and rely on methods of linguistic analysis and natural language processing in the different phases of operation. Multilingual QA systems allow users to query in different languages.

These systems usually have a simple interface where users can enter their queries, while some offer a list of recent queries to help users understand how the system works. QA systems handle these queries by applying algorithms and methods of linguistic analysis, as well as using natural language processing to identify the components and determine the expected response (Zweigenbaum, 2005). This analysis usually uses a variety of standard questions in which certain words are replaced by labels accepted by the system (Pérez-Coutiño *et al.*, 2004).

While the development of QA systems represents progress, the systems nevertheless suffer restrictions. Many were only developed as prototypes, or demonstration versions, and few were marketed. Some researchers have designed and created systems that were presented and discussed at various forums and conferences. However, because the usefulness of the systems was limited to very specific contexts, or because of problems of implementation, only a few of these systems were later developed for end users.

QA systems may be general domain or domain-specific, also known as open- and restricted-domain QA systems. General domain systems answer questions from diverse fields and domain-specific systems, which focus on a specialized area, use specific linguistic resources that enable more precise answers to be given. It can find several examples of both types of QA systems available on the web, i.e. open-domain QA system TextMap Question Answer and a restricted-domain QA systems EAGLI (Engine for question-Answering in Genomics Literature), among others.

We analysed: START, QuALiM, MedQA and HONqa. We chose these four QA systems because they have been used in previous research works about this field, they have been available for few years, and they have been developed by significant research groups.

START, Natural Language Question Answering System, is a publicly accessible information access system that has been available for use on the internet since 1993. START answers natural language questions by presenting components of text and multi-media information drawn from a set of information resources that are hosted locally or accessed remotely through the web. These resources contain structured, semi-structured and unstructured information (Katz, 1990; Kaisser, 2009). It has a dynamic yet easy interface. Information is retrieved from a very wide list of sources, such as World Book, The World Factbook 2008, START KB, Internet Public Library, and many others.

Another open-domain QA system is QuALiM. This is a question answering demo, financed by Microsoft, which retrieves textual information by means of Wikipedia and graphic information, through the Google image search engine. QuALiM is a pattern-based QA system that searches the web for answers. Each of its patterns contains a syntactic description that matches a subclass of questions, a set of syntactic descriptions of potential answer sentences, and semantic information concerning the appropriate answer type for the question class. When asked a question, QuALiM will search all of the patterns' question descriptions and retain those that match the question. The matching pattern's information about potential answer sentence formulations is used to create rather specific quoted search queries that are sent to a web search engine (either Google or Yahoo) (Kaisser, 2008; Fahmi, 2009).

In a study by Ely and colleagues (2000), participating physicians spent on average less than two minutes looking for information to resolve clinical queries, although many of their questions remained unanswered. Regarding this point, some researchers have shown that physicians trust QA systems as search methods for specialized IR (Lee *et al.*, 2006; Yu *et al.*, 2007). The general public also increasingly consults knowledge resources like the web: before or after seeing a doctor, for themselves or for relatives, to obtain information about the nature of a disease, the indications and contraindications of a treatment, and so on (Zweigenbaum, 2005).

HONqa is a domain-restricted QA system operated by the Health On the Net Foundation, which is a Swiss non-profit organization that aims to promote the development of quality, reliable medical information. The HON Foundation has dedicated itself to the maintenance and improvement of the quality of online medical information. With this view, HON has developed a QA applied to medical research where responses obtained are from all the sites certified by the foundation (Cruchet *et al.*, 2009). HONqa is the only multilingual system analysed in our study, as it provides information in English, Italian and French, although we have only focused on the information retrieved in English. The sources of information used by the system are varied, and its search engines are usually general medical or specialized towards a certain illness.

Finally, MedQA, Medical Question Answering System (nowadays this QA system is not available on the webpage) is a specialized QA system that analyses thousands of documents to arrive at a coherent response. Because it works specifically in the area of healthcare, its sources are more specialized (Blair-Goldensohn *et al.*, 2003). It also has a user-friendly interface. It retrieves information from a wide array of sources, including Wikipedia, Medline or Medline Plus. To identify definitional sentences, MedQA implemented a set of lexico-syntactic patterns generated automatically from a large set of precompiled definitional sentences that were collected using Google. The lexicon-syntactic patterns were ranked based on the ratio of their occurrences in the definitional and in the non-definitional sentences. This system suffered from incorrect noun phrases being extracted by its shallow syntactic chunked from the questions (Fahmi, 2009).

Although researchers have recently studied different aspects of QA systems, there are not yet enough studies to evaluate how these tools work. In our study, we have focused on how these four QA systems work to generate phrases with dynamic, coherent definitions and with the most interesting information (Cui *et al.*, 2004; Fahmi, 2009).

2.5.3. METHODOLOGY

In total, a sample of 150 definitional questions about different medical issues formed the basis of this study (Figure 39). We decided to use definitional questions because they do not demand specific data (such as a date, name or place) as the QA system builds an answer. So, these systems have to identify and obtain relevant information in their sources used, and have to summarize and develop the best answer possible as a definition. The questions were obtained

from the webpage WebMD (<http://www.webmd.com/>), a US health portal providing valuable health information and support, tools for managing health problems and specialized background on a number of illnesses. It was created by health specialists who aspire to explain, briefly yet credibly, in-depth medical information, reference material, and online community programmes. The set of questions used passed the internal consistency test with a Cronbach's alpha coefficient of 0.997.

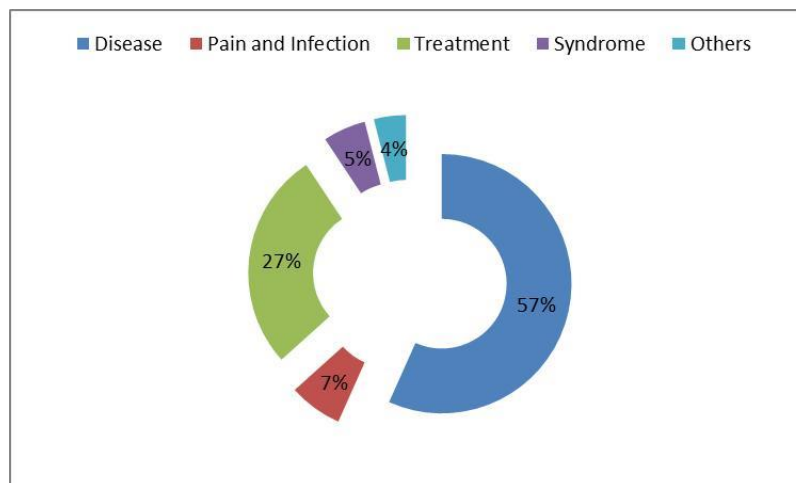


Figure 39. Category of question reference

We used START and QuALiM as open-domain QA systems – although allowing users to pose questions about various health issues, they can respond to even very specialized questions within the area of healthcare (Yu and Kaufman, 2007) – and, as restricted-domain QA systems, MedQA and HONqa.

After asking questions in the four systems, a group of two medical professionals and three medical students evaluated the answers as incorrect, inexact or correct. Correct answers were those that answered the question adequately, were expressed in fewer than 100 words and did not contain information that was irrelevant to the question. Responses that answered the question correctly but that did not meet the rest of the criteria were evaluated as inexact. The evaluations performed on each answer created a base for the application of the evaluation measures on the system's operation (Katz *et al.*, 2002), measures which are described next.

Mean Reciprocal Rank (MRR) is a statistical tool for evaluating any process that produces a list of possible answers to a query. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer (for example, if a question gets the correct answer in first place, it will receive a score of 1, it will be 1/2 if it is in the second place, 1/3 in the third place, and so on). If the answer is not found, a score of 0 is assigned. MRR can be used with several correct answers, but it only takes into account the first correct answer found.

$$MRR = \frac{1}{q} \sum_{i=1}^q \frac{1}{far_i}$$

Total Reciprocal Rank (TRR) is useful when there is more than one correct answer to a question. In these cases, it is not sufficient to consider the first correct answer in evaluations; instead, TRR takes into consideration all the correct answers and assigns a weight to each according to its ranking in the list provided by the system. For example, if the QA system provides two correct answers (the first and the third ones), the TRR will be $1/1 + 1/3$.

First Hit Success (FHS) assigns 1 if the first answer returned by the system is correct and 0 if it is not. This measure, then, only accepts the first questions in the list of results.

We have also used the classic measures of evaluation in IR: precision and recall (Salton & McGill, 1970). Precision is understood as the capacity of a system to retrieve documents, or answers (in the case of QA systems), relevant to the query.

$$\text{Precision} = \frac{|\{\text{relevant answers}\} \cap \{\text{retrieved answers}\}|}{|\{\text{retrieved answers}\}|}$$

Recall is the fraction of the documents, or answers, that are relevant to the queries that are successfully retrieved. Recall is not calculated automatically because this measure depends on all the answers retrieved and the data 'lost' or not retrieved (Taylor, 1999). The method of polling (combining) of all the outputs for the same request is used to establish relative recall:

$$\text{Recall} = \frac{|\{\text{relevant answers}\} \cap \{\text{retrieved answers}\}|}{|\{\text{relevant answers}\}|}$$

The total number of relevant answers that would be analysed in this study is 750 answers per QA system, because we evaluated a maximum of five answers for each of the 150 questions. So an exhaustive and precise QA system will retrieve a total of 750 correct answers.

2.5.4. RESULTS AND DISCUSSION

After asking the 150 questions in the four QA systems, the first five answers from each system were analysed, as five was the average number of answers retrieved by most of the systems, and because users would, presumably, use these types of systems to obtain information quickly, paying most attention to the first answers. Nonetheless, some QA systems offered more answers

for some questions, while others did not offer the five answers. The average number of answers retrieved by each QA system was very significant, because it is higher in HONqa (Table 25).

	Total number of answer retrieved	Average answers retrieved	Total number of answer assessed*	Total correct answers	% correct answers	Total incorrect answers	% incorrect answers	Total inexact answers	% inexact answers
HONqa	6635	44.23	726	343	47.24	325	44.76	58	7.99
QuALiM	441	3	406	166	40.88	180	44.33	60	14.78
MedQA	802	5.34	705	329	46.66	245	34.75	131	18.58
START	236	1.6	234	164	70.08	27	11.54	43	18.38

* Only the first five answers retrieved by QA systems in each question were assessed.

Table 25. Measures for evaluating the quality of answers

The open-domain systems retrieved fewer answers than the restricted-domain systems, with START in last place (with an average of 1.6 answers), followed closely by QuALim (with 3 answers). In the restricted-domain QA systems, the results increased substantially, especially in the case of HONqa (with 44.23 answers), while MedQA was slightly better than the open-domain systems (with an average of 5.34 answers for each question).

The correct answers were present to the greatest degree in START (70.08%). In the two restricted-domain QA systems, this average decreased – MedQA (46.66%) and HONqa (47.24%) – and QuALim was the most deficient, with 40.88% of answers being correct. Figure 40 shows the number of correct, inexact and incorrect answers from the four QA systems analysed.

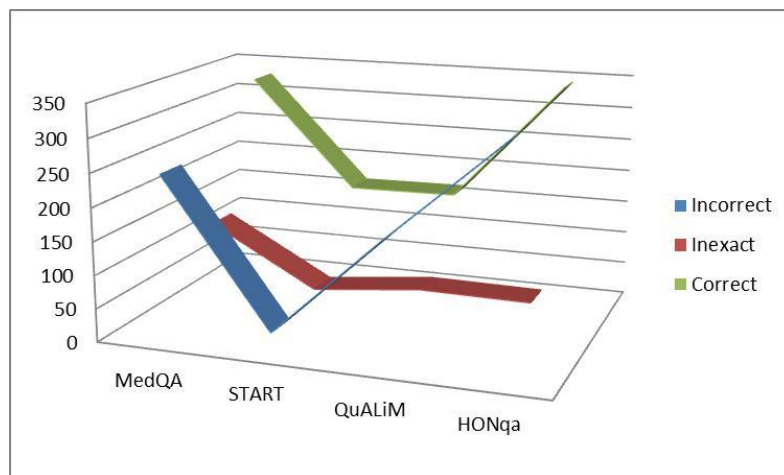


Figure 40. Incorrect, inexact and correct answer

In relation to the inexact answers, the QA system with the lowest average was HONqa (7.99%). The remaining QA systems offered similar averages (MedQA, 18.58%, START, 18.38% and QuALiM, 14.78%).

The incorrect answers in HONqa (44.76%) and QuALiM (44.33%) were similar figures. In MedQA, the value was somewhat lower (34.75%) although still high, and START (11.54%) was the system that presented the fewest inexact answers.

The value obtained by applying the MRR and FHS evaluation measures indicates that MedQA best ranks answers, as the first correct answer appears at the top of the list of results (Table 26). This is significant in that the system does not use any sort of rank algorithm to carry out this process but rather it always recurs to the same order of answers based on the knowledge source. FHS turns out to be a very relevant metric, as users often tend to focus on the first answer obtained, disregarding the rest.

The other restricted-domain QA system, HONqa, is in the opposite situation, as it offers the lowest FHS value but has the second highest average in MRR. The precision of the two open-domain QA systems is quite similar, and there is not a great deal of difference between the two metrics, which is explained by the fact that the answers retrieved for each question usually fluctuate from one to three.

However, the TRR metric is greater in the restricted-domain systems as it considers the value of each correct answer based on its position in the list of results. Naturally, this value increases when more results are offered.

	MRR*	TRR*	FHS*	Precision	Recall
HONqa	0.75	1.15	0.55	47.24%	45.73%
QuALiM	0.65	0.77	0.59	40.88%	22.13%
MedQA	0.87	1.29	0.76	46.66%	43.87%
START	0.67	0.81	0.64	70.08%	21.87%

* MRR: Mean Reciprocal Rank; TRR: Total Reciprocal Rank; FHS: First Hit Success

Table 26. Evaluation measures for QA systems

The precision value in the open-domain QA system START was higher than in the restricted-domain QA systems because the latter ones demonstrate a high rate of noise. However, the recall shows us that the two open-domain QA systems are less exhaustive than the restricted-domain ones. So START presents a recall value of 21.87% and an average of retrieved answers of 1.6, and the values of QuALiM are 22.13%, referring to the recall measure, and an average of three retrieved answers.

As can be observed, none of the applied metrics present very high values, a situation which has clearly been influenced by the high standards set for an answer to be evaluated as correct. In many cases, for example in MedQA and START, there were correct answers with more than 100 characters, which caused them to be considered inexact, as indicated in Section 3. Likewise, in HONqa, there were correct answers that were considered inexact because they did not respond

to the question in a totally specific and precise way. So if we considered the inexact answers as correct to calculate the precision, the values would increase, in some cases considerably (Table 27).

	HONqa	QuALiM	MedQA	START
Precision	52,75%	42,66%	65,24%	88,46%

Table 27. Precision considering correct and inexact answers

Finally, the correlation between the metrics used in this study (Table 28) shows a significant correlation between TRR and MRR because they focus on similar characteristics of the answers retrieved by QA systems.

		MRR	TRR	FHS	Precision	Recall
MRR	Pearson	1.000	0.959*	0.700	-0.265	0.837
	Sig. (2-tailed)		0.041	0.300	0.735	0.163
TRR	Pearson	0.959*	1.000	0.472	-0.316	0.957*
	Sig. (2-tailed)	0.041		0.528	0.684	0.043
FHS	Pearson	0.700	0.472	1.000	0.093	0.197
	Sig. (2-tailed)	0.300	0.528		0.907	0.803
Precision	Pearson	-0.265	-0.316	0.093	1.000	-0.387
	Sig. (2-tailed)	0.735	0.684	0.907		0.613
Recall	Pearson	0.837	0.957*	0.197	-0.387	1.000
	Sig. (2-tailed)	0.163	0.043	0.803	0.613	

* The correlation is significant at 0.05

Table 28. Correlation of measures

2.5.5. CONCLUSIONS

Today, users are more demanding of the retrieved information, as much with regard to the quality as with regard to the quantity, just as in response time. QA systems could therefore be one of the future systems of IR on the internet as they intend to meet the needs and demands of the current users. Nonetheless, despite an increase in QA system research, the open-domain systems are scarce and not all of them function adequately. In as much, it is necessary to continue analysing and evaluating such systems and their characteristics, in order to promote their development and commercialization.

Through analysing the results obtained by asking the 150 questions in the QA systems MedQA, START, QuALiM and HONqa, it was possible to evaluate the systems' operation through applying specific metrics. Despite the limitations demonstrated by these systems, as they are not accessible to everyone and they are not always completely developed, it has been

confirmed that these four systems are valid and useful for obtaining definitional medical information in that they offer coherent and precise answers.

Another interesting aspect concerns the sources of information used by each of these QA systems. In previous work (Olvera-Lobo & Gutiérrez-Artacho, 2010), we compared the sources used by the restricted-domain QA systems with those used by the general-domain systems to ascertain if we could see significant differences in the typology and specialization. A comparison of the sources used by START and QuALiM confirms the absence of significant differences, as both systems use Wikipedia as their main source, although START recurs to more sources for its retrieval. Nonetheless, we see big differences in the sources used by the two restricted-domain QA systems. While MedQA uses dictionaries, encyclopaedias and databases specializing in the biomedical field, HONqa opts for websites specializing in this area.

The results are encouraging because they present this type of tool as a new possibility for gathering precise, reliable and specific information in a short period of time. In this sense, some authors (Saracevic, 1995) have explored various possibilities for improvements, such as the use of ontologies, which will increase the quality of the answers obtained by formalizing the relevant information from the domain in question. In addition, these approaches together with others are slowly attracting more researchers who are experienced in handling the results they produce. These data suggest that we may see unexpected changes in the future and this area deserves to be studied and evaluated in future research.

QA systems have been extended in recent years to explore critical new scientific and practical dimensions (Buitelaar *et al.*, 2008). Additional aspects, such as interactivity (often required for clarification of questions or answers), answer reuse, and knowledge representation and reasoning to support question answering, have been explored. Future research may explore what kinds of questions can be asked and answered about social media, including sentiment analysis. Another key aspect of these systems is that the system–user relationship is two-way. Establishing an interaction helps QA systems find better answers and, in turn, the QA system helps users find answers more quickly. However, it is still necessary to deepen the interactive design of these systems and enable true feedback between questions and answers so that users communicate with the system in a conversational manner. Finally, retrieval has shown that the QA systems are a useful tool to retrieve information quickly and accurately.

Today, users are more demanding of the retrieved information, as much in regards to the quality as in regards to the quantity, just as in response time. Because of all this, the QA systems are the future of information retrieval on the Internet since they intend to meet the needs and demands of the current users. None the less, despite an increase in QA system research, the

open domain systems are scarce and not all of them function adequately. In as much, it is necessary to continue analyzing and evaluating such systems and their characteristics, in order to promote their development and commercialization

2.6. EVALUACIÓN DEL RENDIMIENTO DE LOS SISTEMAS DE BÚSQUEDA DE RESPUESTAS DE DOMINIO GENERAL (PUBLICACIÓN 6)

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2010) "Question-Answering Systems as Efficient Source of Terminological Information: Evaluation". *Health Information and Library Journal*. 27 (4): 268 – 276

RESUMEN

Resumen: Los sistemas de búsqueda de respuestas (SBR) son una alternativa a los tradicionales sistemas de recuperación de información tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de presentar al usuario una lista de documentos relacionados con su búsqueda. Se ha evaluado la eficacia de cuatro SBR disponibles en la Web –QuaLiM, SEMOTE, START, y TrueKnowledge–, mediante una amplia muestra de preguntas de definición, factuales y de lista, pertenecientes a distintos dominios temáticos. Se utilizó una colección de 500 preguntas cuyas respuestas fueron valoradas por los usuarios y, posteriormente, se aplicaron varias medidas para su evaluación (MRR, TRR, FHS, MAP y precisión). Se observa que START y TrueKnowledge presentan un nivel aceptable de respuestas correctas, precisas y en una secuencia bien ordenada. Los resultados obtenidos revelan el potencial de esta clase de herramientas en el ámbito del acceso y la recuperación de información de dominio general.

Palabras clave: Recuperación de información, sistemas de búsqueda de respuestas, evaluación, medidas de evaluación, World Wide Web.

Performance Analysis in Web-based Question Answering Systems

ABSTRACT

Abstract: Information overload is felt more strongly on the Web than elsewhere. Question-answering systems (QA systems) are considered as an alternative to traditional information retrieval systems, because they give correct and understandable answers rather than just offering a list of documents. Four answer search systems available online have been analyzed: START, QuaLiM, SEMOTE, and TrueKnowledge. They were analyzed through a wide range of questions

that prompted responses of definitions, facts, and closed lists pertaining to different thematic areas. The answers were analyzed using several specific measurements (MRR, TRR, FHS, MAP and precision). The results are encouraging and they show that these systems, although each one different, are potentially valid for precise information retrieval of diverse types and thematic areas.

Keywords: Question-Answering Systems, performance analysis, definitional questions, factoid questions, list questions, evaluation.

2.6.1. INTRODUCTION

En el entorno de la Web la sobrecarga de información se deja sentir aún más que en otros contextos. De esta forma, en demasiadas ocasiones, al plantear una determinada consulta en las herramientas de búsqueda de información web (buscadores, directorios o metabuscadores) el número de páginas web recuperadas resulta excesivo y no todas ellas son relevantes ni útiles para los objetivos del usuario. Los sistemas de búsqueda de respuestas o SBR (en inglés, *question-answering systems*) se presentan como una alternativa a los tradicionales sistemas de recuperación de información tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de mostrar al usuario una lista de documentos relacionados con su búsqueda (Jackson y Schilder, 2005). El funcionamiento de los SBR se basa en los modelos de respuestas cortas (Blair-Goldensohn, 2004), y la ventaja principal que ofrece al usuario es que éste no ha de consultar documentos completos para obtener la información requerida puesto que el sistema ofrece la respuesta correcta en forma de un número, un sustantivo, una frase corta o un fragmento breve de texto (Pérez-Coutiño *et al.*, 2004).

Puesto que la búsqueda de respuestas se presenta como un avance destacado en la mejora de la recuperación de información (Kolomiyets y Moens, 2011) se hace necesario determinar su eficacia para el usuario final. Con este objetivo se ha realizado un estudio donde se evalúa el rendimiento y la calidad de las respuestas de los principales SBR de dominio general disponibles en la Web (QuALiM⁷³, SEMOTE⁷⁴, START⁷⁵ y TrueKnowledge⁷⁶) ante preguntas de diversos tipos (de definición, factuales y de lista) y temas (Arte y Literatura, Biología, Personajes, Historia,

⁷³ <http://demos.inf.ed.ac.uk:8080/qualim/> (Disponible hasta noviembre de 2011)

⁷⁴ <http://www.semote.com> (Disponible hasta agosto de 2011)

⁷⁵ <http://start.csail.mit.edu/>

⁷⁶ <http://www.trueknowledge.com>

Economía o Deportes, entre otros), para lo que se aplican diferentes medidas de evaluación. A continuación se detalla el análisis realizado. Los objetivos del trabajo es comparar y evaluar las respuestas ofrecidas por cada SBR de dominio general ante 500 preguntas factuales, de definición y de lista, de modo que podamos confirmar su relevancia y eficacia.

2.6.2. SISTEMAS DE BÚSQUEDA DE RESPUESTAS

Desde el punto de vista de la recuperación de información, el uso del lenguaje natural favorece el acceso a los contenidos al permitirle al usuario recurrir a su forma habitual de expresión. Los SBR normalmente presentan una sencilla interfaz con un motor de búsqueda mediante el que los usuarios pueden formular su pregunta, e incluso algunos proporcionan un listado de las últimas cuestiones introducidas para ayudarles a entender cómo han de plantearlas. Ciertamente, estos sistemas intentan emular el comportamiento del lenguaje humano por lo que tratan de entender la pregunta formulada en lenguaje natural y proporcionar respuestas adecuadas. En otras palabras, la interpretación del lenguaje natural por el sistema es un proceso esencial en el desarrollo de los SBR (Belkin y Vickery, 1985; Sultan, 2006). Tanto es así que el análisis de la pregunta, así como la búsqueda y la extracción de las respuestas son tres importantes tareas llevadas a cabo por los SBR, las cuáles implican, al menos, el procesamiento de las preguntas, el procesamiento de los documentos y el procesamiento de las respuestas (Kangavari *et al.*, 2008).

Los primeros sistemas de búsqueda de respuestas surgieron en los años 60 y utilizaban bases de datos de dominio restringido con información estructurada. Ejemplos clásicos son Baseball (Green *et al.*, 1961), una base de datos de partidos de béisbol –How many games did the Yankees play in July?, Lunar (Woods *et al.*, 1972), una base de datos de análisis químicos de las misiones lunares de Apollo –What is the average concentration of aluminium in high alkali rocks?- o Chat-80 (Warren, 1981), una base de datos geográficos –Which is the largest African country?- con una versión moderna que convierte la pregunta en SQL. Otro tipo de SBR son los sistemas de diálogo como el clásico Eliza (Weizenbaum, 1966). Este sistema simulaba un psicoanalista y puede considerarse precursor de los actuales chatterbot –software diseñado para simular una conversación inteligente con uno o más humanos por medio de texto y/o audio-. Por último, los antecesores más inmediatos de los sistemas web de búsqueda de respuestas, en los que aquí nos centramos, son los sistemas de búsqueda en documentos de texto, los cuáles tomaron un

importante impulso a partir de la conferencia TREC-8 (Text Retrieval Conference) (Voorhees, 1999).

En el tratamiento y la gestión de las preguntas, los SBR aplican algoritmos y métodos de análisis lingüístico y de procesamiento del lenguaje natural para identificar sus componentes y determinar la clase de respuesta esperada (Zweigenbaum, 2005). El tipo de preguntas que suelen permitir son las denominadas preguntas factuales, de definición y de lista. Las preguntas factuales son las relacionadas con datos o hechos concretos, nombres propios, etc., se expresan mediante partículas interrogativas (*who, what, where, when, how*) y persiguen una respuesta concreta y rápida (un nombre, una fecha, un lugar, una cantidad). Este tipo de preguntas constituye la mayoría de las consultas (“*who won the Nobel Prize for Literature in 1994?*”, “*what actress starred in “The Lion in Winter”?*”, “*when was the telegraph invented?*”, “*how did Jimi Hendrix die?*” “*where are the Rocky Mountains?*”). Las preguntas de definición, como su nombre indica, persiguen obtener la definición de un término, organización, etc., y están formuladas como *what is..?* (*what is angiotensin?*). En estos casos, las respuestas más relevantes serán las que ofrezcan información de manera eficiente, con el menor número de palabras, pero de construcción similar a las entradas de una enciclopedia (Greenwood y Saggion, 2004; Olvera-Lobo y Gutiérrez-Artacho, 2011^a). Finalmente, las preguntas de lista son aquellas que solicitan un cierto número de respuestas de un mismo tipo y suelen plantearse de forma imperativa (“*tell me...*”, “*name all of London’s airports*” o “*List 5 pharmaceutical companies that manufactures antibiotics*”).

Como parte de la arquitectura de los SBR, el módulo de procesamiento de documentos se encarga de realizar una primera selección de los documentos o párrafos que se pueden considerar como relevantes para la pregunta planteada (véase figura 1). Las fuentes de información que utilizan los sistemas para seleccionar estos documentos son de lo más variadas y van desde la omnipresente Wikipedia hasta enciclopedias, diccionarios o bases de datos especializadas de gran prestigio como Medline (Olvera-Lobo y Gutiérrez-Artacho, 2011^b). La elección de las fuentes de información es una decisión habitualmente condicionada por el hecho de que se trate de un SBR de dominio general –y, por tanto, capaz de atender consultas de temas muy diversos, como START Natural Language Question Answering o NSIR Question Answering System⁵– o de dominio específico –si se centran en un ámbito temático determinado, como HONQA Health On the Net Foundation⁶ (Crouch *et al.* 2005; Olvera-Lobo y Gutiérrez-Artacho, 2011^c) or EAGLi Engine for question-Answering in Genomics Literature ⁷ (Abdou *et al.*, 2006) –.

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

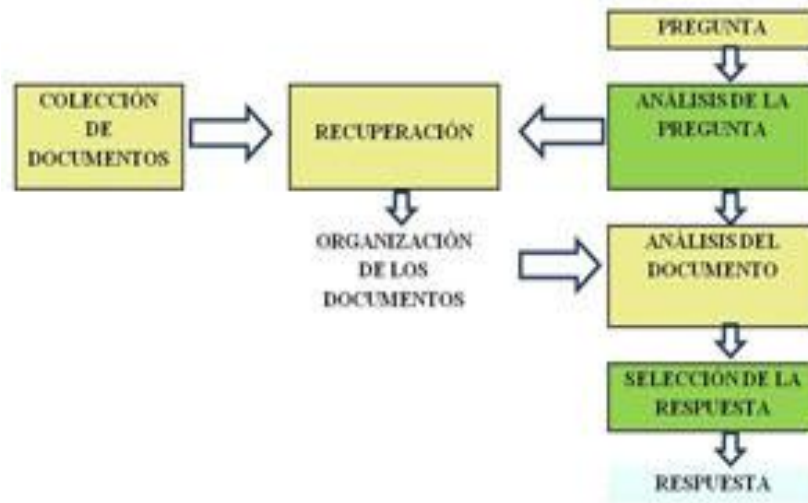


Figura 41. Arquitectura general de un SBR

Por último, el módulo de procesamiento de la respuesta lo forman dos importantes componentes, el destinado a la extracción de la respuesta y el dedicado a la validación de la misma. Las respuestas candidatas se extraen de los documentos que son recuperados por el motor de búsqueda del SBR. Tras ello, mediante el filtrado y la ordenación de las respuestas candidatas, se validan las respuestas que serán las que finalmente se muestren al usuario (Kangavari *et al.*, 2008). El objetivo de esta etapa es eliminar cualquier pasaje incorrecto o redundante que se encuentre en la lista recuperada por el SBR. No siempre la respuesta será única y el sistema puede proveer varias respuestas correctas que satisfagan la necesidad el usuario (Cui *et al.*, 2004; Rodrigo *et al.*, 2010; Tsur, 2003).

2.6.3. MÉTODO Y MATERIALES

La colección de preguntas que se ha utilizado en este estudio incluye preguntas factuales, de definición y de lista en lengua inglesa, y se creó a partir de las colecciones de preguntas de evaluación propuestas por dos de las principales conferencias sobre recuperación de información a nivel internacional, TREC (Text Retrieval Conference) y CLEF (Cross-Language Evaluation Forum). Las colecciones de evaluación generadas en estos foros son utilizadas por los participantes para llevar a cabo la evaluación de sus sistemas de manera que los resultados obtenidos puedan compararse con los de los demás. Partiendo de las colecciones de preguntas de los años 2000 a 2004 se obtuvo una serie de casi 2000 preguntas de definición, factuales y de

lista –para las que existen métodos de evaluación claramente definidos (Voorhees, 2002)– que versaban sobre diferentes temas y especialidades (véanse tablas 29 a 31).

	CLEF	TREC	Total
Nº Preguntas	597	1383	1980

Tabla 29. Procedencia de las preguntas de la muestra

Año						
	2000	2001	2002	2003	2004	Total
Nº Preguntas	730	475	100	475	200	1980

Tabla 30. Preguntas por año

Temas de las preguntas de evaluación									
Arte y Literatura	269	Ciencia	251	Deportes	91	Economía	156	General	324
Geografía	178	Historia	255	Medicina	86	Personajes	219	Política	151
Total								1980	

Tabla 31. Temas a los que se refieren las preguntas de la muestra

Finalmente, se utilizaron para la evaluación las 500 preguntas (véase tabla 32), tanto de dominio general como específico, que obtuvieron respuesta por parte de los cuatro sistemas analizados, a saber, QuaLiM, SEMOTE –sistemas que recientemente han dejado de estar operativos– START, y TrueKnowledge. Se trata de SBR gratuitos, monolingües, de dominio general, disponibles en la Web, y que ofrecen una amplia cobertura temática ante diferentes tipos de preguntas.

	P. definición	P. factuales	P. de lista
Nº Preguntas	127	348	25

Tabla 32. Preguntas según el tipo de respuesta esperada

QuaLiM era un sistema financiado por Microsoft y desarrollado por el investigador Michael Kaiser de la Universidad de Edimburgo. Aunque se definía como una demo, se trataba de un sistema que contaba con un funcionamiento aceptable y recuperaba tanto información textual – para lo que utilizaba únicamente la enciclopedia Wikipedia– como gráfica –extraída del buscador de imágenes de Google– (Kaiser, 2008). Se caracterizaba por presentar una interfaz muy sencilla y breves explicaciones con ejemplos. Por su parte, SEMOTE era un sistema que permitía a los usuarios plantear preguntas sobre diferentes dominios temáticos. Se caracterizaba por utilizar una amplia variedad de recursos –desde páginas web dedicadas a temas específicos hasta portales web de más amplia cobertura– para extraer las repuestas a las preguntas planteadas. Además, los resultados ofrecidos solían ser bastante exhaustivos. Estos dos últimos sistemas no se encuentran actualmente operativos ya que se tratan de demos creados por Universidades para su investigación y tras tenerlos varios años sin actualizar han decidido quitarlos. Sin embargo, los dos sistemas funcionaban correctamente ya que aunque los desarrolladores no trabajasen en

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

ellos, las fuentes de información de donde extraían la información están en continua actualización.



Figura 43. Interfaz de QuALiM: Página de resultados



Figura 44. Interfaz de SEMOTE: Página principal

START es un sistema desarrollado por el Massachusetts Institute of Technology que permite a los usuarios plantear preguntas sobre temas muy diversos, ya sean especializados o no (Katz *et al.*, 2007). Cuenta con una sencilla interfaz y sus tiempos de respuesta son considerablemente plausibles (Olvera-Lobo y Gutiérrez-Artacho, 2010). Las fuentes de información de las que extraen las respuestas son muy variadas, entre las que se encuentran sitios web de cobertura amplia como Wikipedia, diccionarios de uso general, Internet Public Library, WorldBook, The World Factbook 2008, entre otros, así como sitios web dedicados a un determinado ámbito temático como diccionarios y enciclopedias especializados, etc.

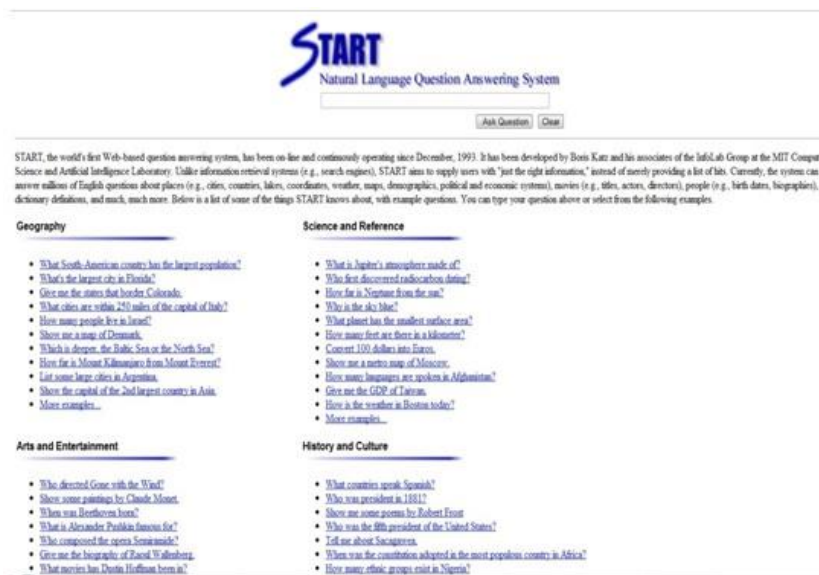


Figura 45. Interfaz de START: Página principal

Por último, TrueKnowledge, desarrollado por una empresa londinense, se caracteriza por extraer las respuestas de numerosos recursos utilizando tanto su base de datos, como diversos sitios web, y las propias respuestas ofrecidas por los usuarios. Al igual que QuALiM, en ocasiones, también recupera información visual.



Figura 46. Interfaz de TrueKnowledge: Página principal

Las respuestas ofrecidas por cada sistema fueron juzgadas por un grupo de estudiantes de tres diferentes grados de la Universidad de Granada y de edades variadas. El grupo de estudiantes se dividió en grupos de tres personas (uno de cada especialidad) y analizaron una parte de la muestra conjuntamente como incorrectas, inexactas o correctas siguiendo la metodología de evaluación propuesta en CLEF (Peters, 2009). Se consideraron correctas aquellas preguntas que

respondían a la consulta de forma adecuada y no añadían información irrelevante. Todas las respuestas que satisfacían la consulta pero incorporaban información irrelevante fueron consideradas inexactas. Finalmente, se calificaron como incorrectas las respuestas cuyo contenido era irrelevante a la pregunta formulada al sistema. A partir de la valoración de las respuestas obtenidas, se aplicaron diferentes medidas para la evaluación del funcionamiento de los SBR.

Una de las medidas utilizadas para evaluar los sistemas de RI en general, que también se aplica a los SBR en particular (Fukumoto *et al.*, 2004; Voorhees y Tice, 1999) es Mean Reciprocal Rank (MRR). Esta medida asigna el valor inverso de la posición en la que la respuesta correcta fue encontrada (1 si es la primera, 1/2 si es la segunda, 1/3 si es la tercera, y así sucesivamente), o cero si la respuesta correcta no fue encontrada. Según esta medida, solamente hay una respuesta correcta dentro de la lista de resultados ofrecidos por el sistema, y el valor final es el promedio de los valores obtenidos para cada pregunta. MRR asigna un valor alto si las respuestas correctas se encuentran en las posiciones más altas del ranking de resultados.

$$MRR = \frac{1}{q} \sum_{i=1}^q \frac{1}{rank_i}$$

En este análisis se han aplicado además otras medidas específicamente desarrolladas para la evaluación de SBR en el entorno de la Web como son FHS y TRR (Radev *et al.*, 2001). Efectivamente, en los SBR la importancia de recuperar los resultados más relevantes en primer lugar es tan esencial como la recuperación en sí. Y es que estos sistemas suelen ser utilizados por usuarios que persiguen recuperar información rápida y eficaz en un breve espacio de tiempo y, por ello, no suelen examinar un número muy alto de respuestas. La medida *First Hit Success* (FHS) asigna valor 1 si la primera respuesta ofrecida es correcta, y valor 0 si no lo es (por lo que sólo considera la respuesta que aparece en primer lugar en la lista de resultados).

Por su parte, *Total Reciprocal Rank* (TRR) resulta una medida bastante útil para evaluar la existencia de varias respuestas correctas ofrecidas por un sistema ante una misma pregunta y asigna un valor a cada respuesta de acuerdo con su posición en la lista de resultados recuperados. Si en una lista de respuestas aparece varias veces repetida la correcta, el usuario puede considerarla como más fiable. En estos casos no es suficiente tener en cuenta únicamente la primera respuesta correcta en las evaluaciones, por lo que TRR las tiene en cuenta a todas. Así, si la primera y la tercera respuesta de una lista de resultados son correctas para una pregunta el valor de TRR será $1/1 + 1/3$.

La precisión, basada en la relevancia, es una de las medidas tradicionales de la RI (Harman, 1998) que más viene utilizándose para la evaluación del funcionamiento de los sistemas de RI

desde los años 50 (Cleverdon, 1997). Impulsada por Salton y McGill (1983), sigue contando en la actualidad con gran aceptación y consenso en la comunidad investigadora, tal como lo demuestra el hecho de que reputados foros como Text Retrieval Conference la incorporen a su modelo de evaluación. La precisión refleja la capacidad del sistema para recuperar documentos (respuestas, en el caso de los SBR) que sean relevantes a la consulta (o pregunta) planteada.

$$\begin{array}{ll} \text{Precisión de la recuperación de información} & \text{Precisión} = \frac{\{\text{Número de documentos relevantes}\}}{\{\text{Número de documentos recuperados}\}} \\ \text{Precisión de la búsqueda de respuestas} & \text{Precisión} = \frac{\{\text{Número de respuestas relevantes}\}}{\{\text{Número de respuestas recuperadas}\}} \end{array}$$

Las medidas tradicionales de evaluación se han ido enriqueciendo con otras que las completan y complementan. Una de las medidas más comunes en la “comunidad TREC” es la *Mean Average Precision* (MAP), la cual genera un único valor que resume el rendimiento de un sistema a distintos niveles de cobertura. Efectivamente, para los sistemas que devuelven una secuencia ordenada de documentos o respuestas, es necesario también considerar el orden en el que se presentan los documentos recuperados. MAP mide el promedio medio de precisión para un conjunto de preguntas formuladas cuyas respuestas son ordenadas por el sistema siguiendo un ranking de relevancia. Cuando se realiza la evaluación utilizando MAP, para cada consulta se calcula la media de los valores de precisión obtenidos cada vez que se encuentra un documento relevante. El valor final para el conjunto de consultas permite determinar qué sistema demuestra una mejor eficacia en la recuperación (Buckley y Voorhees, 2000). En MAP, Q es el conjunto total de preguntas, q se refiere a una única pregunta.

$$MAP = \frac{\sum_{q=1}^Q AveP(q)}{Q}$$

Por su parte, la precisión promedio (*Average Precision* o AveP), incluida en MAP, permite medir la eficacia en la ordenación de las respuestas relevantes en la lista de resultados recuperados puesto que calcula el valor promedio de precisión para cada lugar del ranking en el que aparecen las respuestas relevantes. Un sistema con buen funcionamiento situará las respuestas relevantes en las primeras posiciones de la lista. P_r ¿Qué es la precisión del sistema en la posición r de la lista de resultados y rel_r ¿Qué es una función binaria que indica si el documento recuperado r es relevante a la consulta (valor 1) o no (valor 0).

$$AveP = \frac{\sum_{r=1}^n (P(r) \times rel(r))}{\text{número de respuestas relevantes}}$$

2.6.4. RESULTADOS Y DISCUSIÓN

Tras plantear las 500 preguntas en los cuatro SBR evaluados se analizaron las respuestas ofrecidas por cada uno de ellos. Los resultados indican que el número total de respuestas recuperadas en SEMOTE –5000 y un promedio de 10 respuestas para cada pregunta– fue bastante superior al resto, seguido , aunque de lejos, por QuaLim con algo más de un tercio de respuestas recuperadas –1871 y 3,7 de promedio–. En los otros dos sistemas evaluados el total de respuestas recuperadas fue similar (744 respuestas para START y 766 en TrueKnowledge, es decir, un promedio de 1,5 y 1,5 respectivamente). Los resultados obtenidos en trabajos previos (Olvera-Lobo y Gutiérrez-Artacho, 2011a) ofrecían un número aproximado de respuestas recuperadas por QuALiM y START (3 y 1,6 respectivamente).

Sistemas de búsqueda de respuestas	Total de respuestas	Promedio respuestas	Respuestas correctas	Respuestas inexactas	Respuestas incorrectas
QuaLiM	1871	3,7	47,7% (892)	16,1% (302)	36,2% (677)
SEMOTE	5000	10	31,8% (1588)	10,8% (538)	57,5% (2874)
START	744	1,5	84,3% (627)	7,4% (55)	8,3% (62)
TrueKnowledge	766	1,5	67,4% (516)	19,5% (149)	13,2% (101)

Tabla 33. Respuestas recuperadas en los cuatro SBR

Si se tiene en cuenta la ratio de respuestas correctas respecto al número total de respuestas recuperadas por cada sistema se observa que es START el de funcionamiento más eficaz, con el 84,3% de respuestas correctas, mientras que SEMOTE, el sistema con más respuestas totales recuperadas, ha sido el que presenta un porcentaje inferior (31,8%). Es decir, los sistemas que menos respuestas promedio recuperaron (START, TrueKnowledge y QuaLiM) fueron sin embargo más eficaces, lo que constata que una larga lista de respuestas no garantiza que éstas sean mejores ni más precisas.

En lo que a respuestas incorrectas se refiere SEMOTE –con un 57,5%– es el que mayor índice presenta seguido de QuaLiM con una proporción también bastante considerable –36,2%–. Frente a éstos, tanto START como TrueKnowledge –8,3% y 13,2%, respectivamente– revelan una ratio de respuestas incorrectas manifiestamente inferior. Por último, y en relación al tipo de respuestas que aquí se ha considerado como inexactas, la presencia de las mismas en general no ha sido demasiado elevada, si bien en algún sistema ha superado el 20%. Estos datos mejoran sutilmente los resultados obtenidos en evaluaciones previas de los SBR (Olvera-Lobo y Gutiérrez-Artacho, 2010, 2011a), en donde QuALiM y START fueron evaluados y comparados con SBR de

dominio especializado. Aunque las preguntas en las evaluaciones previas se restringían exclusivamente a las de definición y en un ámbito de especialización, se comprueba que los sistemas siguen siendo eficaces en la recuperación de información.

Los datos que arrojan las medidas de evaluación utilizadas ilustran el comportamiento de estos sistemas considerando además la eficacia en la ordenación de las respuestas. Los resultados, en general, pueden considerarse bastante satisfactorios y, por tanto, evidencian que estos sistemas son potencialmente útiles para recuperar información concisa de distinto tipo y dominio temático (véase Tabla 34). El valor de MRR, medida que considera únicamente el lugar donde aparece la primera respuesta correcta en la lista de resultados, es bastante elevado en todos los sistemas analizados excepto en SEMOTE (0,4). En este sentido, FHS –aún más exigente con la ordenación de las respuestas correctas– es una medida muy destacada puesto que los usuarios, en muchas ocasiones, tienden a centrarse en la primera respuesta recuperada obviando el resto. Se observa que más del cincuenta por ciento de las primeras respuestas ofrecidas en todos los sistemas han sido correctas. El sistema que presenta un valor superior en FHS ha sido START (0,9) frente a SEMOTE, sistema que obtuvo el valor inferior (0,6). En TRR los valores mejoran para todos los sistemas puesto que se tiene en cuenta el lugar que ocupan en el ranking todas las respuestas correctas y no sólo las que aparecen en primer lugar.

Sistemas de búsqueda de respuestas	MRR	FHS	TRR	P	P*	MAP
QualiM	0,72	0,68	1,05	0,48	0,64	0,59
SEMOTE	0,38	0,55	1,01	0,32	0,42	0,35
START	0,91	0,89	1,05	0,84	0,92	0,93
TrueKnowledge	0,82	0,83	0,89	0,67	0,87	0,78

MRR: Mean Reciprocal Rank; FHS: First Hit Success; TRR: Total Reciprocal Rank; P: precisión; P*: precisión incluyendo también a las respuestas inexactas; MAP: Mean Average Precision

Tabla 34. Medidas de evaluación

Al analizar la precisión se observa que, como para las otras medidas, son START y TrueKnowledge los SBR que ofrecen mejores resultados. Además, si se flexibiliza el nivel de exigencia y se incluye en el cálculo de la precisión, no sólo las respuestas valoradas como correctas sino también las denominadas inexactas –que igualmente incluyen la información requerida pero con cierto ruido– los valores se incrementan, en algunos casos considerablemente. Por su parte, MAP, una medida ampliamente usada que ofrece una idea global del funcionamiento del sistema, muestra el mismo patrón de comportamiento que las medidas anteriores. Efectivamente, se observa que, excepto para el caso de TRR, hay una alta correlación entre las medidas usadas en este estudio (véase Tabla 35).

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

Sistemas de búsqueda de respuestas	MRR	TRR	FHS	Precisión	MAP
MRR	1	-0,09	0,96*	0,96*	0,97*
TRR		1	-0,21	-0,14	-0,11
FHS			1	0,99**	0,99**
Precisión				1	0,99**
MAP					1

MRR: Mean Reciprocal Rank; FHS: First Hit Success; TRR: Total Reciprocal Rank; MAP: Mean Average Precision

* La correlación es significativa a 0,05 (p<0,05).

** La correlación es significativa a 0,01 (p<0,01).

Tabla 35. Correlación entre medidas

2.6.5. CONCLUSIONES

El usuario actual confía en recuperar información específica y de calidad que responda a sus necesidades. Los SBR presentan una interesante alternativa a la recuperación de información en Internet intentando satisfacer sus exigencias y demandas. Sin embargo, a pesar del aumento de esta clase de sistemas y del avance que supone el poder contar con herramientas de búsqueda de información de este tipo, los SBR disponibles en la Web son escasos y no todos proporcionan una cobertura adecuada. De hecho, las investigaciones que se vienen realizando y que culminan en interesantes propuestas plasmadas en diferentes publicaciones, foros y congresos, salvo contadas excepciones –y ya sea porque su utilidad se limita a contextos muy concretos, o bien por sus dificultades de implementación–, no se desarrollan para el usuario final.

En este análisis se han evaluado cuatro SBR de dominio general accesibles desde la Web mediante una colección de 500 preguntas cuyas respuestas, conforme a la metodología TREC, fueron juzgadas como correctas, incorrectas o inexactas por estudiantes y especialistas en diferentes campos temáticos. En base a estas valoraciones se aplicaron diferentes medidas de evaluación mediante las que se ilustra claramente la eficacia del funcionamiento de los sistemas analizados. Los dos sistemas que obtuvieron peores resultados (QuALiM y SEMOTE), recientemente han desaparecido y ya no se encuentran operativos en la Web para el usuario final. Uno de los principales problemas que presentan estos sistemas es que las bases de datos internas del sistema no se actualizan con regularidad, presentando en ocasiones resultados obsoletos. Sin embargo, como la mayor fuente de información de los SBR son portales y páginas web de reconocido prestigio, cuya gestión y financiación principalmente depende de organismos gubernamentales, suple los problemas iniciales de la estructura.

No obstante, el estudio realizado revela resultados alentadores debido a que presentan este tipo de herramienta como una nueva posibilidad para obtener información precisa y fiable en un corto período de tiempo.

2.7.MULTILINGUAL QUESTION-ANSWERING SYSTEM IN BIOMEDICAL DOMAIN ON THE WEB: AN EVALUATION (PUBLICACIÓN 7)

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011) "Multilingual Question-Answering System in biomedical domain on the Web: an evaluation". *Multilingual and Multimodal Information Access Evaluation. Lecture Notes in Computer Science*, 6941: 83-88, Berlín: Springer International Publishing.

ABSTRACT

Abstract. Question-answering systems (QAS) are presented as an alternative to traditional systems of information retrieval, intended to offer precise responses to factual questions. This study assesses the effectiveness of these systems with terminological sources for specialists, as well as general users, in the context of multilingual searches within a restricted domain. With this aim, an analysis has been made of the results offered by the QA multilingual biomedical system HONqa, available in the Web. The study has used a set of 120 biomedical definitional questions (What is...?), taken from the medical website WebMD, which were formulated in English, French, and Italian. The answers have been analysed using a series of specific measures (MRR, TRR, FHS, precision, recall, MAP), as well as sources used to get the responses.

The study confirms that for all the languages analysed the functioning effectiveness needs to be improved, although in the multilingual context analysed the questions in the English language achieve better results for retrieving definitional information than in French and Italian. This may be because the information sources from which the answers are elicited are more numerous in English and have a more appropriate structure for this purpose than in the other languages studied.

Keywords: Multilingual data, Multilingual Question Answering Systems, Restricted-domain Question Answering Systems, HONqa, Evaluation of Information Retrieval, Biomedical data, Evaluation measures

2.7.1. INTRODUCTION

The advent of the Web and its subsequent expansion have provided the general public with access to enormous volumes of information, offering unquestionable benefits. Nevertheless, this has also brought disadvantages such as overloads of information—which in this environment is even more acute— or the fact that much of the information is incorrect, incomplete, or inaccurate, whether intentionally so or not. Consequently, it becomes indispensable to develop tools and procedures that enable the user to acquire reliable information that is relevant for a particular consultation. This is the challenge that faces Information Retrieval (hereafter IR)

In recent years, some of the efforts to improve IR in the Web has focused on the design and development of the so-called question-answering systems (hereafter QAS). The development of the QAS gained strong impetus in the conference on information retrieval TREC (*Text Retrieval Conference*)⁷⁷—primarily beginning with TREC-8 (Vorhees, 1999)— which since 1992 has constituted an important international forum to unite and foment research in different areas of information retrieval.

This work evaluates the multilingual search for definitional responses in the context of restricted-domain QAS. For this, the results offered by the multilingual biomedical QAS HONqa, available on the Web, were analysed. A set of 120 questions (in English, French, and Italian) related to this thematic field, together with the responses received, were assessed by a series of measures applicable to such systems. Below, the characteristic principles of the QAS are briefly described, the methodology used is detailed, and the main results are presented.

2.7.2. QUESTION-ANSWERING SYSTEMS: BEYOND INFORMATION RETRIEVAL

According to a study by Ely (1999), medical specialists invest an average of more than two minutes searching for information related to questions that arise and, despite the time taken up, adequate answers are often not found. In this sense, several works have demonstrated the confidence of medical specialists in the use of QAS as a method of searching and retrieving specialized information (Lee *et al.*, 2006; Yu & Kaufman, 2007). Patients have also increasingly consulted these systems, before and after seeing the doctor, to gather information on the nature of the illness, treatment recommendations, contraindications, etc. (Zweigenbaum, 2005)

⁷⁷ Access to Trec (Text REtrieval Conferecen): <http://trec.nist.gov/>

QAS are designed to offer understandable responses to factual questions of specialized content rapidly and precisely in such a way that the user does not have to read the complete documents to satisfy a particular query. These systems begin with the user's question in order to construct coherent answers in everyday language (Costa & Santos, 2007).

The functioning of the QAS is based on short-answer models (Blair-Goldensohn, 2004), since the potentially correct answer does not go beyond a number, a noun, a short phrase, or a brief fragment of text. Normally, they accept consultations expressed as queries that include interrogative particles (what, how, why, when, where) or those presented in an imperative manner (tell, answer, reply). When a question is posed in the system, a certain analysis begins to determine the type of response the user seeks. Then the QAS locates and extracts one or several responses from different sources, according to the subject of the consultation (Olvera-Lobo & Gutiérrez-Artacho, 2011b). Finally an evaluation is made and information that is redundant or that does not answer the question properly is eliminated in order to present specific responses designed to satisfy the information needed by the user (Cui *et al.*, 2004; Tsur, 2003).

Therefore, the QAS must generate sentences with data and/or dynamic and consistent definitions that summarize the most descriptive information to be found in the set of documents pertaining to the term or focus of the user's query (Blair-Goldensohn, 2004; Cui *et al.*, 2004). The answers are presented according to a ranking established by the system based on its own relevance criteria (Olvera-Lobo & Gutiérrez-Artacho, 2010; Sing *et al.*, 2005).

Aside from its numerous advantages, the QAS nevertheless present a number of restrictions. For the environment of the Web, some of these have been developed only as prototypes or demos, and are very rare in systems available to the final user. Nonetheless, some QAS of a general domain can currently be consulted, as they are capable of addressing questions on very diverse subjects (such as START⁷⁸ or QuALiM⁷⁹) and from very specific domains (such as EAGLi⁸⁰ o HONqa⁸¹), which focus on a given context. In addition, the QA systems try to overcome the limitations of the traditional tools of information retrieval, such as the consultations being monolingual. For this reason, studies focusing on the sphere of multilingual QAS (such as HONqa) recently began to emerge as systems that provide information in different languages.

Nevertheless, in this ambit, there are still different matters that require further development, such as cross-language searches and interactivity. The appropriate retrieval tools that enable the

⁷⁸ Access to START, <http://start.csail.mit.edu/>

⁷⁹ Access to QuALiM, <http://demos.inf.ed.ac.uk:8080/qualim/>

⁸⁰ Access to to EAGLi, <http://eagl.unige.ch/EAGLi/>

⁸¹ Access to HONqa, <http://services.hon.ch/cgi-bin/QA10/qa.pl>

procedure known as cross-lingual information in retrieval (CLIR) would enable consultations in several languages with information retrieval in all the languages accepted by the system (Diekema, 2003). The translingual QAS, still not widely used, could present an architecture composed of several QAS monolingual systems while also incorporating a translation module and/or a tool for linguistic resource that would provide cross-lingual retrieval (Tsur, 2003).

In relation to interactiveness, the establishment of a system-user relationship that is not one-directional, as well as interaction in the search process, would help the QAS prepare better answers and aid the user in finding the answer more quickly. Nevertheless, it is still necessary to progress in the design of interactive systems that offer true feedback between questions and answers, so that the user can communicate with the system at a conversational level.

Within this framework, our study evaluates the effectiveness of these tools, with the aim of contributing to the development of this emerging subarea of the IR.

2.7.3. METHOD SECTION

The methodology applied used 120 biomedical questions concerning definitions on diverse medical subjects. The questions were formulated as consultations of the type “What is” in the search engine of the website WebMD⁸², created in the USA by medical specialists to resolve doubts held by patients. The information offered by this portal covers a wide spectrum of medical subjects of different degrees of specialization. From the questions that this portal provides, an ample set was selected to be translated by a team of professional translators of French and Italian, and from this initial set, 120 questions that elicited responses in the three languages in the system were selected. These constituted the body of the questions used. The main biomedical aspects related to the selected questions were diseases, operations, treatments, syndromes, and symptoms. The set of questions used passed the validity test with a Chronbach’s alpha of 0.936.

HONqa, the QAS evaluated in this work, was developed by the *Health On the Net Foundation*⁸³, an Swiss non-profit organization that promotes the availability of reliable and quality online medical information. It is a multilingual system that retrieves information in English, French, and Italian (Cruchet *et al.*, 2009). The information sources used are very diverse, though usually general medical portals or those specializing in a certain disease were used. This

⁸² Access to WebMD, <http://www.webmd.com/>

⁸³ Access to Health On The Net Foundation, <http://www.hon.ch/>

system has a very simple interface to make the consultations, though it offers the possibility of an advanced search that permits specifying the language, the information source, and the number of responses. The results page indicates the total number of answers retrieved, the language of the responses, whether the question is definitional or factual, and the thematic area of the medical question posed. In the results page, the user can positively or negatively appraise each answer retrieved, offering this information to the developers.

The responses offered by the system were evaluated by a group of experts from different medical areas, as incorrect, inexact, or correct, according to the evaluation methodology proposed in CLEF (Peters, 2009). Questions that were answered properly and did not add irrelevant information were considered correct. All the answers that resolved the question but added irrelevant information were considered inexact. Finally, answers that contained irrelevant information with regard to the question were considered incorrect. From the evaluation of the answers retrieved, the evaluation measures were applied as described below (Raved *et al.*, 2001). *Mean Reciprocal Rank* (MRR) assigns the inverse value of the position in which the correct answer is found (1 if the first, $\frac{1}{2}$ if the second, $\frac{1}{4}$ if the fourth, and so on), or zero if there is no correct response. This measure considers only the first correct response shown on the list of results offered by the system, and the final value is the average of the values found for each question. MRR assigns a high value to the responses that were in the highest classification positions.

$$MRR = \frac{1}{q} \sum_{i=1}^q \frac{1}{f ar_i}$$

Total Reciprocal Rank (TRR) is useful for evaluating the existence of several correct responses offered by a system to the same query. In these cases, it is not sufficient to consider only the first correct response in the evaluations, and thus TRR takes all of them into account and assigns a weight to each answer according to its position in the list of results retrieved. Thus, if the second and fourth answers on the list of results are correct for a question, the TRR value would be $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$.

First Hit Success (FHS) assigns a value of 1 if the first answer offered is correct, and a value of 0 if it is not (thus it considers only the answer that appears in the first place on the list of results).

Also, we used the classical measures of evaluation in Information Retrieval: precision and recall. *Precision* is defined as the capacity of system to retrieve documents, or answers (in the case of QA systems), relevant to the query.

$$\text{Precision} = \frac{|\{\text{relevant answers}\} \cap \{\text{retrieved answers}\}|}{|\{\text{retrieved answers}\}|}$$

Recall is the fraction of the documents, or answers, that are relevant to the queries that are successfully retrieved.

$$\text{Recall} = \frac{|\{\text{relevant answers}\} \cap \{\text{retrieved answers}\}|}{|\{\text{relevant answers}\}|}$$

Recall was not calculated automatically because this measure depends on all the answers retrieved and the data “lost” or not retrieved. Here, we considered 5 to be the total possible number of relevant answers, since it is the maximum number of results analysed for each query.

Mean Average Precision or *MAP*: For systems that return a ranked sequence of documents, it is desirable also to consider the order in which the returned documents are presented. MAP measures the average precision of a set of queries for which the answers are arranged by relevance.

$$\text{MAP} = \frac{\sum_{q=1}^Q \text{AveP}(q)}{Q}$$

Thus, the average precision (AveP) must be calculated for each question formulated. Average precision emphasizes the higher ranking of relevant documents. It is the average precision computed at the point of each of the relevant documents in the ranked sequence. A well-functioning system situates the relevant answers in the first few positions on the list. The AveP seeks to measure the order of the relevant answers in the list of results retrieved.

$$\text{AveP} = \frac{\sum_{r=1}^n (P(r) \times \text{rel}(r))}{\text{number of relevant answers}}$$

We considered 600 answers per QA system as the total number of relevant answers that would be analysed in this study, because we evaluated a maximum of 5 answers for each of the 120 questions. Thus, a perfectly exhaustive and precise QA system will retrieve a total of 600 correct answers.

2.7.4. RESULTS SECTION

After 120 questions were posed in the multilingual QAS, the first five answers in each of these systems were analysed. Although the mean of the answers retrieved by the system in the three languages approached and in some cases exceeded 30, only the first answers offered were considered, as the users enter the QA systems for their quickness and precision, and the user is not likely to read a long list of answers for each question. The assessment measures applied also took into account the arrangement of the answers and, on the other hand, the sources from which the answers were drawn were analysed in each language. Below, the results are shown.

2.7.4.1. EVALUATION MEASURES

The average of the total answers retrieved by the system was 47.46 in the case of English, 27.36 for French, and 25.03 for Italian.

	Total answers	Average of Answers	Answers analyzed	Correct answers	Inexact answers	Incorrect answers
English	5695	47,46	589	287 (48.73%)	67 (11.37%)	235 (39.9%)
French	3283	27,36	573	52 (9.07%)	124 (21.64%)	397 (69.28%)
Italian	3123	25,03	585	32 (5.47%)	95 (11.63%)	458 (82.9%)

Table 36. Answers retrieved by HONqa in the three languages

The volume of answers retrieved in English was substantially higher (5695 answers retrieved) than in the other cases, the other two languages registering similar values (3283 for French and 3123 for Italian).

The correct answers were present in greater measure in the English version of the system, which properly responded to more than 48% of the cases, whereas French offered a low rate of 9.07% and Italian provided only 2.05%. The number of imprecise answers was higher in French (21.64%), followed by Italian (16.24%). In relation to the incorrect answers, the number was very high in all three languages, exceeding 50% of the total in French (397) and Italian (458). Figure 46 shows the percentage of correct, imprecise, and incorrect answers to the questions formulated in the three languages available in HONqa.

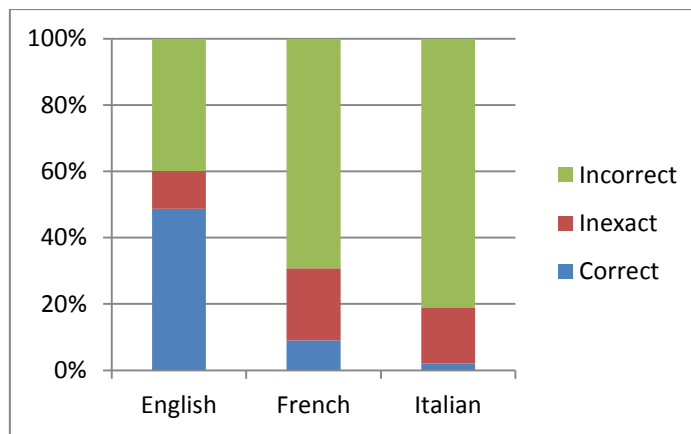


Figure 46. Correct, inexact and incorrect answers

This behaviour directly influenced the results found when applying the evaluation measures proposed. The MRR value for the responses offered in the three languages reflect the above comments. While the results of the English option were quite plausible, at 0.76, the other two languages offered very poor results (0.19 for French and 0.13 for Italian), indicating the low reliability of the first response by the system for these languages.

In relation to the TRR measure, which considers all the answers correct among the first 5 results analysed, it was found that, except for English the results did not substantially improve. FHS is an important measure, as the users often tend to focus on the first response retrieved, skipping the rest. It was found that more than 50% of the answers offered in English (0.575) provided an initial correct answer while the other cases were not encouraging (0.12 in French and 0.06 in Italian).

MAP is a widely used measure that offers an overall idea of the functioning of the system by the calculation of the mean value of the average precision for the set of queries. The evaluation of the system did not register an adequate level for any of the languages analysed.

	MRR	TRR	FHS	P	P*	P@3	P@3*	R	R*	MAP
English	0.76	1.55	0.575	0.55	0.65	0.57	0.67	0.59	0.59	0.25
French	0.19	0.27	0.12	0.10	0.31	0.11	0.32	0.17	0.44	0.05
Italian	0.13	0.15	0.06	0.05	0.16	0.06	0.15	0.13	0.28	0.03

(P=Precision, P*=Precision considering also the inexact answers, P@3=Precision of the 3 first results, P@3*=Precision of the 3 first results including inexact answers; R=Recall; R*= Recall considering also the inexact answers)

Tabla 37. Evaluation measures

The precision value is closely related to the rest of the measures discussed above. The precision was measured, on the one hand, considering as relevant only the responses scored as correct (measures P and P@3) and, on the other hand, considering also the imprecise answers (measures P* and P@3*) as relevant –that is, being more flexible to evaluate an answer as adequate. In this latter case, clearly, the precision values significantly increased in some cases. Nevertheless, as with the rest of the measures, there was a marked difference between English and the other languages. On considering P@3 or only the precision of the first three results, the values found for this new measure only weakly improved though not very different from the previous values. This indicates that the arrangement of the answers retrieved according to their relevance to the question was not the best. The small number of correct answers in some cases made the recall values of the QAS very low, except in the case of English.

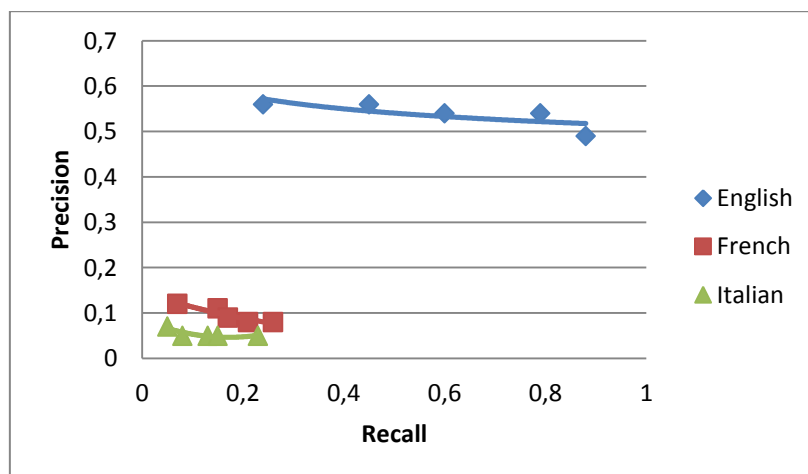


Figura 47. Recall-precision relationship

Figure 47 shows the recall-precision relationship of the first five answers retrieved illustrating and corroborating the results found by the rest of the measures applied.

2.7.4.2. INFORMATION SOURCES

The sources of the information used by the system were analysed to determine whether such uneven results for the three languages were influenced by the resources from which the information was retrieved, or whether the differences might be due to linguistic problems or to the system architecture itself. Table 48 shows the 10 sources of information most widely used

for the English language in the QAS analysed. For the option of searches in English, highly regarded medical sources from GB and the USA were used, except for one. The most widely used was WebMD (representing 18% of the total answers analysed). With a similar proportion of correct and incorrect responses (38.68% and 46.23%, respectively), this is a reputable website of medical outreach directed at a non-specialized audience who consult medical specialists to broaden their knowledge and resolve doubts regarding treatments, diseases, operations, and other concerns. Given the characteristics of this website, it becomes difficult to attribute the high number of incorrect answers to the source used.

Sources (% of total answers)	Answers retrieved	Corrects	Inexacts	Incorrects
WebMD (18%)	106	41 (38.68%)	16 (15.09%)	49 (46.23%)
About.com (13.24%)	78	34 (43.59%)	10 (12.82%)	34 (43.59%)
National Institutes of Health (8.32%)	49	29 (59.18%)	2 (4.09%)	18 (36.73%)
MedicineNet.com (7.64%)	45	26 (57.78%)	5 (11.10%)	14 (31.12%)
WrongDiagnosis.com (6.62%)	39	26 (66.67%)	3 (7.76%)	10 (25.64%)
FamilyDoctor.org (3.90%)	23	11 (47.83%)	2 (9.14%)	10 (43.48%)
LiveStrong.com (3.74%)	22	9 (40.91%)	1 (4.54%)	12 (54.55%)
KidsHealth (3.23%)	19	11 (57.89%)	0	8 (42.11%)
Revolution Health (2.38%)	14	7 (50%)	1 (7.15%)	6 (42.85%)
Netdoctor (2.38%)	14	9 (64.29%)	1 (7.14%)	4 (28.57%)
Otras (30.56%)	180	84 (46.67%)	26 (14.44%)	70 (38.89%)
Total of answers analysed	589 (100%)			

Table 38. Sources used in English

*About.com*⁸⁴, belonging to *The New York Times Company*, is a website directed mainly at Americans, where hundreds of experts collaborate and frequently update their contents. This is a semi-specialized source and, in our study, it has offered the same proportion of correct and incorrect answers. The *National Institutes of Health*⁸⁵, constitute a governmental source of the USA serving a large number of institutions of the federal administration involved in medical research. It is one of the sources with the highest percentages of correct answers. Meanwhile, *MedicineNet*⁸⁶ is a portal created to offer basic medical information for doctors as well as patients and families. The results of this source were acceptable, since almost half of the answers were correct. *WrongDiagnosis.com*⁸⁷ is one of the main free portals online for medical information both for patients as well as for professionals. The answers from this source were quite adequate, since more than half of them were correct. The last five sources of information used by HONqa to formulate its answers were characterized also by medical portals specialized

⁸⁴ Access to About.com, www.about.com

⁸⁵ Access to National Institutes of Health, www.nih.gov

⁸⁶ Access to MedicineNet, www.medicinenet.com

⁸⁷ Access to Wrong Diagnosis, www.wrongdiagnosis.com

in providing more or less specialized information. Some are websites specializing in a certain segment of the public, as in the case of *KidsHealth*⁸⁸ or *LiveStrong*⁸⁹. The results were quite appropriate except in the case of *LiveStrong*, where the number of incorrect answers outweighed the correct ones. This could be due to the nature of the source itself, since the portal is dedicated to *fitness* and to providing nutritional information.

It bears mentioning the good performance of the questions made to the sources of information used in HONqa in the case of English. In most of the cases, the replies were taken from the most appropriate sources, according to the topic and the characteristics of the query.

Sources (% of total answers)	Answers retrieved	Corrects	Inexacts	Incorrects
L'Internaute (10.47%)	60	5 (8.33%)	19 (31.67%)	36 (60%)
Bonjour-Docteur (10.47%)	60	15 (27.27%)	10 (18.18%)	35 (63.64%)
CareVox (3.14%)	18	1 (5.56%)	1 (5.56%)	16 (88.88%)
Partageons L'Info Santé (2.69%)	13	1 (6.69%)	4 (30.77%)	8 (61.54%)
LaSante.net (2.69%)	13	6 (46.15%)	1 (7.70%)	6 (46.15%)
SantePratique.fr (1.92%)	11	2 (18.18%)	2 (18.18%)	7 (63.64%)
Prevention.ch (1.57%)	9	1 (17.17%)	3 (27.27%)	5 (55.56%)
AXA Santé (1.40%)	8	2 (25%)	3 (37.5%)	3 (37.5%)
AstraZeneca (1.40%)	8	0	3 (37.5%)	5 (62.5%)
Association Francois Aupetit (1.49%)	8	0	1 (12.5%)	7 (87.5%)
Others (64.92%)	370	19 (6.49%)	77 (20.81%)	269 (72.70%)
Total answers analysed	573 (100%)			

Table 39. Sources used in French

The sources used for the searches in French were not all exclusively medical and there were various portals of different thematic domains. *L'Internaute*⁹⁰ is an electronic journal created in the year 2000 by *Benchmark Group*, an independent French company which offers a great variety of interactive services. The results of this source were poor, with a minimal percentage of correct answers.

*Bonjour-Docteur*⁹¹, edited by *Pulsations Multimedia, SARL*, offers videos with medical information explained by specialists. *CareVox Partageons L'Info Santé*⁹² is a website dedicated to topics of health and well being that include information on medicines, diseases, scientific research, and well being. It was created by the initiative of *Pharmagest Interactive*, a company dedicated to the development and marketing of computer solutions for the pharmaceutical

⁸⁸ Access to Kids Health, www.kidshealth.org

⁸⁹ Access to Live Strong, www.livestrong.com

⁹⁰ Access to L'Internaute, www.linternaute.com

⁹¹ Access to Bonjour – Docteur, www.bonjour-docteur.com

⁹² Access to CareVox Partageons L'Info Santé, www.carevox.fr

industry. They adhere to the principles of *HONcode*, a system of quality certification for websites dedicated to the area of health.

*LaSante.net*⁹³ is a portal with articles, a directory, and general social and preventive medical information. *SantePratique.fr*⁹⁴ is a portal that offers information on taking medicine correctly, symptoms of obesity, infantile diseases, and age-related illnesses. In addition, it provides advice for a balanced diet and a guide on food complements. Also, it adheres to the principles of *HONcode*.

*Prevention.ch*⁹⁵ is a Swiss website created in 1996 and written in French. *AXA Santé*⁹⁶ is a portal devoted to promoting prevention in health matters. *AstraZeneca*⁹⁷ offers information on medicines to doctors and patients. All these websites follow the principles of *HONcode*. Finally, *AFA*⁹⁸ is a website of a French company dedicated to promoting and financing research on inflammatory intestinal illnesses, and *Le Collège des médecins de famille du Canada*⁹⁹, available in English and French is the website of a Canadian organization founded in 1954 made up of family doctors that offer ongoing training.

Although all the sources used are characterized by being quite specialized and being developed by large organizations, institutions, or companies, all the answers retrieved were in general quite unsatisfactory. In all the sources, the numbers of incorrect answers exceed the correct ones, and in many cases even the sum of the correct and imprecise answers.

Sources (% of total answers)	Answers retrieved	Corrects	Inexacts	Incorrects
Medicitalia (40.24%)	235	2 (0.85%)	11 (4.68%)	222 (94.47%)
Farmaco e Cura (5.48%)	32	3 (9.37%)	11 (34.38%)	18 (56.25%)
Lab Tests Online (3.14%)	17	3 (17.65%)	3 (17.65%)	11 (64.7%)
Cup 2000 (2.57%)	15	0	0	15 (100%)
Ambulatorio (2.57%)	15	3 (20%)	1 (6.67%)	11 (73.33%)
Sport Medicina (2.49%)	14	1 (7.15%)	1 (7.15%)	12 (85.7%)
AIMAC (1.88%)	11	1 (9.1%)	2 (18.18%)	8 (72.72%)
Sos tumori (1.88%)	11	2 (18.19%)	5 (45.45%)	4 (36.36%)
San Raffaele (1.88%)	11	0	1 (9.09%)	10 (90.91%)
Vita di Donna (1.54%)	9	0	2 (22.22%)	7 (77.78%)
Otras (38.18%)	215	17 (7.9%)	58 (26.98%)	140 (65.12%)
Total of answers analysed	585 (100%)			

Table 40. Sources used in Italian

⁹³ Access to La Santé, www.lasante.net

⁹⁴ Access to Santé Pratique, www.santepratique.fr

⁹⁵ Access to Prevention, www.prevention.ch

⁹⁶ Access to AXA Santé, www.axasante.fr

⁹⁷ Access to Astra Zeneca, www.astrazeneca.fr

⁹⁸ Access to AFA (Association Francois Aupetit), www.afa.asso.fr

⁹⁹ Access to Le Collège des médecins de famille du Canada, www.cmfc.ca

The Italian sources, like those of the French language, offered more incorrect answers than correct ones. *MedicItalia*¹⁰⁰ is one of the most distinguished online medical portals in the Italian panorama and has received noted prizes of recognition. This is a website directed both at specialized doctors as well as students and to other types of internauts to delve into, broaden, and update their knowledge on health topics. It bears a HONcode certification. Meanwhile, *Farmaco e Cura*¹⁰¹, a website created in 2007 to provide outreach information on health in Italian, seeks to offer complete and reliable information to people who have medical knowledge, in an effort to encourage good communication between doctors and patients.

*Lab Test Online*¹⁰² is a multilingual website created by the *American Association for Clinical Chemistry* in which numerous international organizations participate. It is dedicated to clinical laboratory tests, and abides by the *HONcode*. *CUP 2000*¹⁰³ offers services related to the medical sphere, such as the ability to pose online questions to specialists.

*Ambulatorio.com: Il mondo della salute a casa tua*¹⁰⁴ is a medical portal with information on illnesses and symptoms, etc. The website *Medicina Sport On-line: Il sito di chi ama lo sport*¹⁰⁵ focuses on sports medicine from a preventive perspective. The aim of the *AIMaC (Associazione Italiana Malati di Cancro)*¹⁰⁶ is to provide information on cancer and its treatment as well as to guarantee psychological support for patients and their families. All follow the *HONcode*.

Finally, it bears mentioning *SOS Tumori*¹⁰⁷, which provides information and support in the field of oncology; *San Raffaele S.p.A.*¹⁰⁸, distinguished in Italy in the sphere of the rehabilitation of health assistance; and *Vita di Dona: L'associazione per la tutela della salute femminile*¹⁰⁹, a non-profit organization that offers free consultation related to health problems in women. The first two also adhere to the *HONcode*.

¹⁰⁰ Access to *MedicItalia*, www.medicitalia.it

¹⁰¹ Access to *Farmaco e Cura*, www.farmacoecura.it

¹⁰² Access to *Lab Test Online*, www.labtestonline.com

¹⁰³ Access to *CUP 2000*, www.cup2000.it

¹⁰⁴ Access to *Ambulatorio.com: Il mondo della salute a casa tua*, www.ambulatorio.com

¹⁰⁵ Access to *Medicina Sport Online*, www.sportmedicina.com

¹⁰⁶ Access to *AIMaC: Associazione Italiana Malati di Cancro*, www.aimac.it

¹⁰⁷ Access to *Sos tumori*, www.sostumori.it

¹⁰⁸ Access to *Sant Raffaele*, www.sanraffaele.it

¹⁰⁹ Access to *Vita di Dona: L'associazione per la tutela della salute femminile*, www.vitaididona.it

2.7.5. CONCLUSIONS

The analysis of the results from posing 120 questions in the QA system of the biomedical domain HONqa has enabled the evaluation of its functioning in the retrieval of multilingual information by applying specific measures and analysing the information sources used for each language. Despite the restrictions that these systems show, the study indicates that this QA system is valid and useful for the retrieval of definitional medical information, mainly in the English language, although it is not yet the most advisable resource to gather multilingual information in a quick and precise way.

Another noteworthy aspect refers to the information sources used to elicit the answers. When the sources used by each language were compared, it was found that most of them are portals or websites specializing in medical topics though the type of portal differed from one language to the other. In English, most of the portals presented their content in the form of questions, posed by the developers themselves or the users of the system, and their corresponding answers. In other portals, the information offered was of quality but not all showed definitions or information relevant to diseases, treatments, etc. Therefore, this is probably one of the causes of the great differences in the results for the different languages.

The search for multilingual responses in the context of the Web still needs to progress a long way to reach the effectiveness levels of general retrieval systems, and especially in monolingual ones. Furthermore, the searches in the English language continue to offer the best results primarily because the information sources from which the answers are taken are more numerous and have a more appropriate structure for this purpose than in other languages. The search for multilingual answers still needs many improvements—for example, in relation to the retrieval of information and to the gathering and presentation of responses. Nevertheless, the results are promising as they show this type of tool to be a new possibility within the sphere of precise, reliable, and specific information retrieval in a brief period of time.

2.8. SATISFACCIÓN DE USUARIOS DEL ÁMBITO DE LA TRADUCCIÓN EN EL USO DE SISTEMAS DE BÚSQUEDA MULTILINGÜE DE RESPUESTAS COMO RECURSO DE INFORMACIÓN TERMINOLÓGICA (PUBLICACIÓN 8)

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2014) "Satisfacción de usuarios del ámbito de la traducción en el uso de sistemas de búsqueda multilingüe de respuestas como recurso de información terminológica". *TIC, trabajo colaborativo e interacción en Terminología y Traducción*. Interlingua, Granada: Comares

RESUMEN

Con el rápido crecimiento de Internet y el desarrollo de las nuevas tecnologías en los últimos años, los sistemas de búsquedas de respuestas (SBR) se han convertido en una alternativa a los tradicionales sistemas de Recuperación de Información. Aunque existe una prolifera producción sobre estos sistemas, pocos son los trabajos que se han desarrollado en la evaluación centrada en el usuario. El presente trabajo se centra exclusivamente en la evaluación de los SBR multilingüe ya que permite al usuario acceder a información terminológica no disponible en su lengua, y en la evaluación centrada en el usuario para entender las necesidades del usuario e identificar las dimensiones y factores en el desarrollo de un sistema de información con el fin de mejorar su aceptación.

El objetivo es conocer el grado de satisfacción del usuario para el SBR multilingüe HONqa (inglés, francés e italiano) utilizando la herramienta de los investigadores ONG *et al.* (2009) basado en un examen de modelos y teorías. Para ello, se ha analizado el grado de satisfacción de los profesionales de la traducción como recurso de información terminológica.

Un total de 122 especialistas en inglés, francés o italiano ha utilizado y evaluado el recurso terminológico desde la perspectiva de la traducción.

Tras el análisis de los resultados obtenidos respecto al SBR de HONqa los cuales han permitido evaluar la satisfacción de los alumnos de la Facultad de Traducción e interpretación de Granada se ha comprobado que el sistema les resulta a los usuarios-alumnos fácil y útil para la recuperación de información terminológica en todos los idiomas.

Palabras clave: Información multilingüe, Sistemas QA multilingües, Sistemas QA de dominio especializado, HONqa, Evaluación de la satisfacción de los usuarios, Información terminológica.

2.8.1. INTRODUCCIÓN

La aparición y el consiguiente desarrollo de la Web han concedido al público general la capacidad de acceder a grandes volúmenes de información, lo que supone indudables beneficios. No obstante, esto también conlleva inconvenientes como la sobrecarga de información –que en este entorno es aún más acusada–, o el hecho de que buena parte de dicha información sea incorrecta, incompleta o imprecisa –ya sea de forma intencionada o no–. En consecuencia, se hace imprescindible avanzar en el desarrollo de herramientas y procedimientos que permitan al usuario obtener información fiable que sea relevante para su consulta en particular. Este es el reto al que se enfrenta la Recuperación de Información (en adelante, RI).

En los últimos años, algunos de los esfuerzos en la mejora de la RI en el entorno de la Web se han centrado en el diseño y desarrollo de los llamados sistemas de búsqueda de respuestas (en adelante SBR). El desarrollo de los SBR toma un importante impulso en el seno de la conferencia sobre recuperación de información TREC (Text Retrieval Conference¹¹⁰) –principalmente a partir de TREC-8 (Vorhees, 1999)– la cual, desde 1992, constituye un importante foro internacional para aunar e incentivar la investigación en diferentes ámbitos de RI.

En este trabajo se ha evaluado la satisfacción de los usuarios como información terminológica en el ámbito de los SBR de dominio restringido. Para ello, se ha utilizado la herramienta desarrollada por Ong *et al.* (2009) para la evaluación del grado de satisfacción de los usuarios del SBR multilingüe HONqa. A continuación se describen brevemente las principales características de los SBR, se detalla la metodología utilizada y se presentan los principales resultados obtenidos.

2.8.2. SISTEMAS DE BÚSQUEDA DE RESPUESTAS: MÁS ALLÁ DE LA RECUPERACIÓN DE INFORMACIÓN

En el entorno de la Web la sobrecarga de información se deja sentir aún más que en otros contextos. De esta forma, en demasiadas ocasiones, al plantear una determinada consulta en las herramientas de búsqueda de información Web (buscadores, directorios o metabuscadores) el

¹¹⁰ TREC (Text REtrieval Conference), <http://trec.nist.gov/>

número de páginas Web recuperadas resulta excesivo y no todas ellas son relevantes ni útiles para los objetivos del usuario. Por ello, los profesionales de diversos ámbitos comienzan a reconocer la utilidad de otros tipos de sistemas, como los SBR, como método para la obtención de información especializada de forma rápida y efectiva (Crouch *et al.*, 2005; Lee *et al.*, 2006).

La importancia de la aparición de los SBR, en el ámbito de la información digitalizada, es actualmente innegable. Se presentan como una alternativa a los tradicionales sistemas de RI tratando de ofrecer respuestas precisas y comprensibles a preguntas factuales, en lugar de presentar al usuario una lista de documentos relacionados con la búsqueda (Jackson y Schilder, 2005), de modo que el usuario no ha de leer documentos completos para obtener la información requerida. La tarea de los SBR se puede definir como la acción de encontrar respuestas concretas a preguntas precisas y arbitrarias formuladas por los usuarios en lenguaje natural dentro de grandes volúmenes de información (como por ejemplo la World Wide Web). A pesar del avance que supone el poder contar con herramientas de búsqueda de información de este tipo, los SBR presentan algunas restricciones.

El funcionamiento de los SBR se basa en los modelos de respuestas cortas (Blair-Goldensohn, 2004), ya que la pregunta es dividida en palabras claves a las que se le asigna una etiqueta que permite distinguir y recuperar la información deseada. El sistema reemplaza esa etiqueta por las palabras adecuadas para poner a disposición de los usuarios una selección de textos que responden correctamente a la consulta (Pérez-Coutiño *et al.*, 2004). La ventaja principal es que el usuario no ha de leer documentos completos para obtener la información requerida puesto que el sistema ofrece la respuesta correcta en forma de un número, un sustantivo, una frase corta o un fragmento breve de texto. Planteada la pregunta en el motor de búsqueda del sistema, se procede a analizar la pregunta separando la palabra o palabras claves, luego se localiza y extrae una respuesta a partir de diferentes fuentes dependiendo de la especialización del sistema se utilizarán unas u otras fuentes (Olvera-Lobo y Gutiérrez-Artacho, 2010)–, y finalmente, se evalúa y elimina aquella información redundante o que no responde correctamente a la pregunta planteada para, posteriormente, elaborar y presentar una o varias respuestas concretas que supuestamente satisfacen la necesidad del usuario (Cui *et al.* 2004; Tsur 2003).

En cuanto a los tipos de la lengua hay un gran número de clasificaciones propuestas (Adiwibowo y Adriana, 2007; Izquierdo *et al.*, 2007; Roger *et al.*, 2007; Solorio *et al.*, 2005) que permite la distinción entre sistemas monolingües y multilingües, aunque en algunos casos un tercer tipo podrían ser incorporados relacionados con los SBR en varios idiomas que utilizan el inglés como lengua fundamental (García-Cumbreras *et al.*, 2006). Por otro lado, la cobertura temática de los documentos en la base de datos es también un aspecto a tener en cuenta

(Harabagiu et al, 2000; Magnini et al, 2001; Moldovan et al, 2003.; Roger *et al.*, 2007). Así, mientras que unos pocos son de dominio abierto, es decir, los SBR permiten a los usuarios realizar preguntas sobre temas de carácter general o multidisciplinarios, sin restricciones, otros tienen una colección especializada de un determinado campo temático.

En un entorno multilingüe como el de la Web la mayoría de los sistemas de RI tienen la limitación de encontrar documentos sólo en el idioma en el que se escribe la consulta o bien incorporan sistemas de traducción automática que únicamente resultan útiles cuando los documentos ya han sido localizados y tampoco siempre, pero no facilitan un medio efectivo para salvar la barrera del idioma en el proceso de búsqueda. Dentro de los SBR, hay que mencionar los vinculados a la RI multilingüe, los cuáles implican, al menos, dos lenguas, y permiten plantear las consultas en varios idiomas y recuperar información en todas las lenguas aceptadas por el sistema (Diekema, 2003).

Estos sistemas son capaces de operar en una colección de documentos multilingües dada una consulta determinada, y recuperar aquella información relevante que responda a la misma, independientemente del idioma utilizado al plantear la consulta (Grefenstette, 1998). Dentro de la RI multilingüe, se debe destacar el caso especial del objeto del estudio, los sistemas multilingües de SBR, que son aquellos sistemas donde el idioma en el que se plantea la pregunta, puede ser diferente a la lengua en la que está escrito el documento recuperado, pero se diferencian del resto de sistemas de la RI multilingües en que éstos no recuperan documentos completos sino que responden con una respuesta corta a la consulta planteada. Normalmente el funcionamiento de estos sistemas es muy similar a los de SBR monolingües, solamente se incorpora el módulo de la traducción y/o la herramienta o recurso lingüístico que llevará a cabo la recuperación translingüe.

2.8.3. INTERACTIVIDAD DE LOS SISTEMAS DE BÚSQUEDA DE RESPUESTAS

La capacidad de interacción con el usuario es un aspecto muy importante de los SBR. Aunque es notable el interés creciente por los SBR capaces de mantener un diálogo continuado con los usuarios, la mayor parte de los sistemas aún no alcanzan un grado de comunicación conversacional. La ventaja de los sistemas plenamente interactivos es que utilizan cada nuevo input para conseguir cierta retroalimentación y afinar en la búsqueda de la respuesta más adecuada. Por ejemplo, están más preparados para resolver con éxito las preguntas enlazadas, que incluyen referencias a preguntas o a respuestas anteriores (¿quién descubrió la penicilina?,

¿dónde nació?). Los sistemas interactivos de búsqueda de respuestas encajan perfectamente con el enfoque de los tradicionales servicios de referencia. Las aplicaciones en este ámbito (Pomerantz, 2005) parten de la premisa de que la consulta primera del usuario suele ser errónea, ya que éste puede tener cierta dificultad para expresar su necesidad de información. Se cuestiona la validez de la pregunta inicial y se busca una reformulación más exacta ofreciendo mayor orientación durante la consulta. Por el contrario, los sistemas de búsqueda de respuestas que no cuentan con la perspectiva del servicio de referencia, es decir, la inmensa mayoría, consideran que el usuario formula correctamente su pregunta.

A la par de los SBR, se han desarrollado medidas para evaluar el desempeño de estos sistemas. Estas medidas se han propuesto desde diferentes perspectivas, tales como: la utilización de colecciones de prueba (Voorhees y Tice, 2000), el uso de pruebas de lectura y comprensión de textos (Charniak *et al.*, 2000) y la aplicación de sistemas automáticos que evalúan la validez de las respuestas dadas por los sistemas, esto es mediante su comparación con las respuestas generadas por los humanos a las mismas preguntas (Breck *et al.*, 2000). La perspectiva de mayor éxito ha sido la utilización de colecciones de prueba, la cual consiste de un conjunto de documentos, un conjunto de preguntas y respuestas y una medida del rendimiento del sistema (Vicedo *et al.*, 2003).

2.8.4. EVALUACIÓN CENTRADA EN EL USUARIO

Partimos de un modelo de evaluación de SBR a partir de modelos de sistemas de satisfacción del usuario y aceptación de la tecnología. El concepto fundamental fue inspirado por la Teoría de la Acción Razonada (TRA) (Fishbein y Ajzen, 1980), elaborado de la psicología social, que se ha utilizado para predecir una amplia gama de los comportamientos ya que se plantea la hipótesis de que las actitudes están determinadas por las características que los observadores asocian con un objeto se basan en las características del contexto y de la muestra. También se ha basado en la teoría del comportamiento planificado (TPB) (Ajzen, 1991) que ayuda a entender cómo podemos cambiar el comportamiento de la gente.

Se adoptan tres dimensiones de la calidad (DeLone y McLean, 2003) de los modelos de éxito de los sistemas de información: calidad de la información, calidad de los sistemas y calidad del servicio. La satisfacción también tiene tres dimensiones: satisfacción de la información, satisfacción de los sistemas y satisfacción del servicio.

Existen creencias acerca de que la calidad afecta a la satisfacción. En el que la satisfacción de la información y la satisfacción del sistema se cree que influye en la utilidad y la facilidad de uso percibida. Muchos investigadores consideran que la intención de un individuo para utilizar un sistema se encuentra correlacionada significativamente con su uso real que a su vez es un indicador de la aceptación de un sistema de información. En resumen la satisfacción del usuario puede ser considerada como una función de la facilidad y utilidad de uso percibida y es probable que conduzca a una aceptación y éxito. El éxito de la evaluación de un SI y la satisfacción del usuario son temas importantes en el campo de la gestión de la información, especialmente en el caso de los sistemas de gestión *online*. Básicamente, se utilizan modelos de evaluación para comprender las necesidades de los usuarios e identificar importantes dimensiones y factores que intervienen en el desarrollo de sistemas con el fin de ampliar su aceptación.

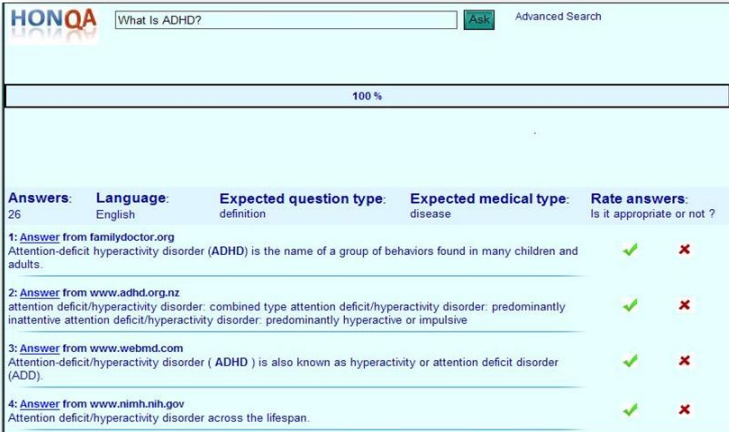
Con el rápido crecimiento en los últimos años, los SBR han surgido importantes aplicaciones. Por ello, han recibido la atención de los investigadores, en particular la de los de recuperación de información y procesamiento del lenguaje natural. Hay SBR que están enfocados a la interacción con el usuario, permitiendo que el proceso de búsqueda vaya refinándose con la intervención del usuario. Este tipo de sistemas tienen en cuenta a los usuarios y sus interacciones con las preguntas y los documentos donde se realiza la búsqueda (Hersh, 2006). El principal objetivo de la recuperación de la información de los SBR es satisfacer las necesidades de sus usuarios. El primer intento de definir “la satisfacción del usuario” como concepto, Tessier *et al.* (1977) afirmó que la satisfacción es “En última instancia una vivencia en el interior de la cabeza del usuario” y por lo tanto fue una respuesta “que puede ser a su vez emocional e intelectual”. La evaluación centrada en el usuario se utiliza para entender las necesidades del usuario e identificar importantes dimensiones y factores en el desarrollo de un sistema de información con el fin de mejorar su aceptación.

2.8.5. METODOLOGÍA

En este trabajo se ha llevado a cabo un estudio para evaluar la opinión de los usuarios acerca del uso de una herramienta que puede ser útil como fuente de información de traducción terminológica en un sistema biomédico. Con este fin, se ha analizado y evaluado el funcionamiento del SBR, de dominio especializado en el dominio biomédico HONqa. HONqa está desarrollado por la *Health On the Net Foundation* (Fundación de la salud en la red), una organización suiza sin ánimo de lucro. El usuario accede directamente a la información requerida

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

y no tiene que pasar por todos los documentos propuestos por el motor de búsqueda, una de las características principales de este sistema es que permite al usuario evaluar positiva o negativamente la respuesta recuperada, ofreciendo esa información a los desarrolladores (véase figura 1). Además, para cada pregunta planteada al sistema por el usuario se indica de qué tipo es (definición, factual o de lista), el número total de respuestas recuperadas, la lengua en la que están, así como sobre qué aspecto médico versa (tratamiento, síntomas, diagnóstico...). Los objetivos del sistema son los de ofrecer respuestas pertinentes a una pregunta en lenguaje natural, promover la creación de información médica de calidad y veraz y facilitar el acceso a los datos médicos más recientes e importantes a través del uso de Internet.



Answers:	Language:	Expected question type:	Expected medical type:	Rate answers:
26	English	definition	disease	Is it appropriate or not ?
1: Answer from familydoctor.org				
Attention-deficit/hyperactivity disorder (ADHD) is the name of a group of behaviors found in many children and adults.				✓ ✗
2: Answer from www.adhd.org.nz				
attention deficit/hyperactivity disorder: combined type attention deficit/hyperactivity disorder: predominantly inattentive attention deficit/hyperactivity disorder: predominantly hyperactive or impulsive				✓ ✗
3: Answer from www.webmd.com				
Attention-deficit/hyperactivity disorder (ADHD) is also known as hyperactivity or attention deficit disorder (ADD).				✓ ✗
4: Answer from www.nimh.nih.gov				
Attention deficit/hyperactivity disorder across the lifespan.				✓ ✗

Figura 48. Página de resultados de sistema de búsqueda de respuestas HONqa

El modelo de evaluación propuesto proporciona un marco de trabajo para el diseño de SBR desde la perspectiva del usuario y que podría ayudar a incrementar la aceptación del usuario de los SBR. En nuestro estudio partimos de la herramienta propuesta por los investigadores Ong *et al.* (2009), en su trabajo “The measurement of user satisfaction with question answering systems” que permite medir la satisfacción de los usuarios. La herramienta propuesta ha sido probada de una forma rigurosa proporcionando un alto grado de confianza en la fiabilidad y la validez de las escalas, los resultados confirmaron la existencia de cuatro factores con valores propios superiores a 1, que en conjunto representaron el 78,5% de la varianza total, confirmando la carga significativa de todos los elementos en un solo factor indicado de forma unidimensional, esto junto al hecho de que no haya carga cruzada de dichos ítems, apoyan la validez discriminante del instrumento. La fiabilidad se determinó mediante el uso de alfa de Cronbach para evaluar la consistencia interna de los elementos que representan a cada factor. El

instrumento de 18 ítems (véase Tabla 41) tenían una alta fiabilidad (0,92) superando el estándar mínimo sugerido para la investigación básica (Anexo 3). La escala utilizada para medir el grado de satisfacción es una escala Likert de siete puntos que va desde “Totalmente en desacuerdo” a “Muy de acuerdo”. Esta medida supone que los usuarios no estarán satisfechos con un sistema si este no proporciona información de forma satisfactoria. La calidad técnica de un sistema es irrelevante, ya que un sistema no se consideraría exitoso si no satisface las necesidades de los usuarios. Esto explica por qué la calidad del sistema no juega un papel en la medición de la satisfacción del usuario.

Para ello se ha utilizado un cuestionario que se estructura en tres partes, una primera donde se presenta los datos demográficos (edad, sexo y curso), una segunda parte que es la autopercepción de los usuarios respecto al grado de experiencia con herramientas de recuperación y al grado de conocimientos médicos, que se evalúa con una escala de tres puntos (alto=3, medio=2, bajo=1). Y por último, una tercera parte que consta de 17 preguntas en donde se recoge la herramienta diseñada para medir la satisfacción de los usuarios en los SBR (Ong *et al.*, 2010). Estas preguntas han sido evaluadas por una escala de Likert de 7 puntos (0= Totalmente en desacuerdo, 1= Muy en desacuerdo, 2 = En desacuerdo en ciertos aspectos, 3 = Indeciso, 4 = De acuerdo en ciertos aspectos, 5 = Muy de acuerdo, 6= Totalmente de acuerdo).

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

CODIGO	ORIGINAL	DESCRIPCIÓN
E1	Q27	<i>La interacción con el sistema de búsqueda de respuestas es clara y comprensible</i>
E2	Q22	<i>Resulta sencillo aprender a utilizar el sistema</i>
E3	Q25	<i>Es fácil llegar a estar completamente familiarizado con el uso del sistema</i>
E4	Q24	<i>Encuentro fácil utilizar el sistema para que realice lo que yo quiero que haga</i>
E5	Q26	<i>Encuentro fácil utilizar el sistema</i>
U1	Q20	<i>Utilizar el sistema podría mejorar mi eficacia en el desarrollo de mi trabajo</i>
U2	Q18	<i>El sistema podría ser útil en el desarrollo de mi trabajo</i>
U3	Q16	<i>El uso del sistema podría mejorar el rendimiento en mi trabajo</i>
U4	Q19	<i>El uso del sistema en mi trabajo podría incrementar mi productividad</i>
U5	Q21	<i>El uso del sistema podría hacer más fácil mi trabajo</i>
S1	Q12	<i>Este sistema de búsqueda de respuestas es fiable</i>
S2	Q13	<i>Los técnicos del sistema de búsqueda de respuestas ofrecen un servicio rápido a los usuarios</i>
S3	Q11	<i>La tecnología utilizada por el sistema y la información que ofrece están actualizados</i>
S4	Q14	<i>Los técnicos de los sistemas de búsquedas de respuestas están bien formados para hacer bien su trabajo</i>
I1	Q2	<i>La información facilitada por el sistema es fácil de comprender</i>
I2	Q4	<i>La información facilitada por el sistema es relevante</i>
I3	Q1	<i>La información facilitada por el sistema es completa</i>
I4	Q3	<i>La información facilitada por este sistema está personalizada</i>

E =Facilidad de uso, **U** =Utilidad, **S** =Calidad del Servicio, **I** =Calidad de la Información
Qn= número de pregunta en nuestro cuestionario

Tabla 41. Herramienta de evaluación para medir la satisfacción de los usuarios

Estas 17 preguntas se agrupan en cuatro factores: un primer factor, la facilidad de uso del sistema (la cinco primeras preguntas), un segundo factor, la utilidad del sistema (de la sexta a la décima), un tercero, la calidad del servicio (de la undécima a la décima tercera) y un cuarto, la calidad de la información (de la décimo cuarta a la décimo séptima) (véase Tabla 42). Los cuestionarios se han realizado en la Facultad de Traducción e Interpretación de la Universidad de Granada con alumnos de 1ª, 2º, 3º y 4º curso de la especialidades de Inglés, Francés e Italiano, la elección de estas tres lenguas es porque coinciden con las del sistema multilingüe HONqa.

La interacción con el sistema de búsqueda de respuestas es clara y comprensible: *Facilidad de uso*
 Resulta sencillo aprender a utilizar el sistema: *Facilidad de uso*
 Es fácil llegar a estar completamente familiarizado con el uso del sistema: *Facilidad de uso*
 Encuentro fácil utilizar el sistema para que realice lo que yo quiero que haga: *Facilidad de uso*
 Encuentro fácil utilizar el sistema *Facilidad de uso*
 Utilizar el sistema podría mejorar mi eficacia en el desarrollo de las tareas académicas: *Utilidad*
 El sistema podría ser útil en el desarrollo de mis tareas: *Utilidad*
 El uso del sistema podría mejorar el rendimiento en mi trabajo: *Utilidad*
 El uso del sistema en mis tareas podría incrementar mi productividad: *Utilidad*
 El uso del sistema podría hacer más fácil mi trabajo: *Utilidad*
 Este sistema de búsqueda de respuestas es fiable: *Calidad del servicio*
 La ayuda ofrecida por el sistema es suficiente: *Calidad del servicio*
 La tecnología utilizada por el sistema y la información que ofrece están actualizados: *Calidad del servicio*
 La información facilitada por el sistema es fácil de comprender: *Calidad de la información*
 La información facilitada por el sistema es relevante: *Calidad de la información*
 La información facilitada por el sistema es completa: *Calidad de la información*
 La información facilitada por este sistema está personalizada: *Calidad de la información*

Tabla 42. Tabla de ítems del cuestionario agrupada en cuatro factores

2.8.6. RESULTADOS Y DISCUSIÓN

Los resultados se presentan en los siguientes cuatro apartados: datos demográficos, autopercepción de los usuarios respecto al grado de experiencia con herramientas de recuperación y al grado de conocimientos médicos, satisfacción de los usuarios respecto al sistema según la herramienta de evaluación aplicada, y correlaciones entre el grado de satisfacción en relación a la facilidad de uso, utilidad, calidad del servicio y calidad de la información.

2.8.6.1. Datos demográficos

Los datos recogidos de un total de 122 alumnos se dividen en las siguientes lenguas: 58 alumnos de inglés, 32 alumnos de francés y 32 alumnos de italiano. En los 122 encuestados las edades oscilan entre los 17 y los 39 años obteniendo una media de edad de 21,48. Aunque existe cierta variación en la edad de los alumnos según la lengua, la media de edad en los tres idiomas es prácticamente igual (20,50) en francés, (21,83) en inglés y (21,84) en italiano (véase tabla 53).

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

Estadísticos

Edad		
N	Válidos	122
	Perdidos	0
Media		21,48
Mediana		21,00
Moda		20
Mínimo		17
Máximo		39

Tabla 43. Estadísticas de la edad

De los 122 alumnos 27 son hombres y 95 son mujeres. Esta diferencia entre el número mujeres y hombres se ve reflejada en el número de alumnos y alumnas del Grado y Licenciatura en Traducción e Interpretación, siendo casi el triple el número de alumnas que el de alumnos. La mayoría de los estudiantes son de tercero (89 alumnos). Le siguen en número los alumnos de cuarto (20) y primero (12) (véase tabla 2). La mayoría de los alumnos son de la especialidad de inglés (58) le sigue a continuación los alumnos de francés e italiano ambos con igual número de alumnos (32).

Curso

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Primero	12	9,8	9,8	9,8
	Segundo	1	,8	,8	10,7
	Tercero	89	73,0	73,0	83,6
	Cuarto	20	16,4	16,4	100,0
	Total	122	100,0	100,0	

Tabla 44. Estadísticas de los alumnos según curso

2.8.6.2. La autopercepción de los usuarios respecto al grado de experiencia con herramientas de recuperación y al grado de conocimientos médicos.

El grado de experiencia global (1,75) corresponde según la escala de 3 puntos utilizada (bajo, medio, alto) a un valor bajo. El grupo de italiano (2,07) supera esta media obteniendo un nivel medio en cuanto al grado de experiencia en el uso de herramientas de recuperación, le sigue el grupo de francés (1,73) y en último lugar el de inglés (1,60). La media global (1,22) obtenida en el

grado de conocimientos médicos no es significativa, los resultados son muy similares en los tres grupos obteniéndose un nivel de conocimientos médicos bajo (Figura 49).

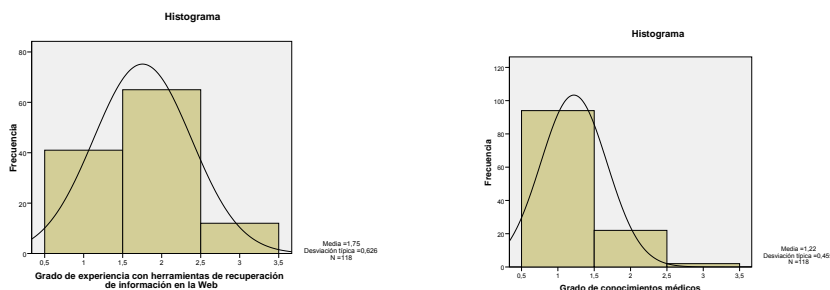


Figura 49. Histograma del grado de conocimiento de los estudiantes

2.8.6.3. La satisfacción de los usuarios respecto al sistema, según la herramienta de evaluación aplicada

El grado de satisfacción de los alumnos de la Facultad de Traducción e interpretación que tienen acerca del SBR HONqa se evaluó en el cuestionario en el cual se encontraban dos diferentes escalas una escala de tres puntos y una escala de Likert de 7 puntos los resultados se explicaran a continuación a través de la descripción de las diferentes relaciones resultantes del análisis entre los tres grupos de las 3 especialidades de idioma (inglés, francés e italiano) realizado con la ayuda del programa SPSS (*Statistical Program For Social Sciences*) A la pregunta realizada, “la interacción con el sistema de búsqueda de respuestas es clara y comprensible” la media global obtenida es de 4,04 mostrándose de acuerdo en ciertos aspectos. El grupo de inglés (4,59) al igual que el de francés (4,38) muestran el mismo grado de conformidad en ciertos aspectos y el grupo de italiano se muestran indecisos (2,70). Los resultados obtenidos a la pregunta, “resulta sencillo aprender a utilizar el sistema” la media global obtenida (4,88) de los tres grupos reflejan que se encuentran de acuerdo en ciertos aspectos. A los alumnos del grupo de inglés le resulta sencillo el aprendizaje (5,34), en cambio al grupo de francés (4,75) y al de italiano no les resulta sencillo en su totalidad (4,16).

El resultado obtenido a la pregunta es “fácil llegar a estar completamente familiarizado con el uso del sistema”, la media global (4,49) obtenida por los tres grupos indican acuerdo en ciertos aspectos. Los del grupo de inglés dan una puntuación mayor (4,95) les siguen los del grupo de francés (4,23) y por último los del grupo de italiano (3,90) que se muestran indecisos.

A la pregunta “encuentro fácil utilizar el sistema para que realice lo que yo quiero que haga” la media global obtenida en (3,47) los tres grupos muestra una clara indecisión. El grupo de inglés (3,98) da una puntuación superior al de francés y al grupo de italiano (2,69) que se muestra en desacuerdo en ciertos aspectos. A la pregunta “encuentro fácil utilizar el sistema, la media global del grupo (4,61) nos indica que están casi de acuerdo en este aspecto. Por el contrario el grupo de inglés (5,00) está muy de acuerdo con la facilidad del sistema, mientras que el grupo de francés (4,69) opina al igual que la media grupal y el grupo de italiano (3,84) se encuentra indeciso. A la pregunta “utilizar el sistema podría mejorar mi eficacia en el desarrollo de las tareas académicas” la media global grupal (3,70) indica indecisión. El grupo de inglés (4,14) se muestra más de acuerdo en ciertos aspectos, mientras que el grupo de francés (3,43) y el de italiano (3,13) se encuentran indecisos. A la pregunta “el sistema podría ser útil en el desarrollo de mis tareas” la media global (3,91) grupal muestran indecisión. El promedio entre las tres lenguas no es significativo, no hay diferencias entre los tres grupos. A la pregunta “el uso del sistema podría mejorar el rendimiento en mi trabajo” la media global (3,63) grupal indica vacilación. El grupo de inglés (4,12) muestra algo más de seguridad, mientras que el grupo de francés y el de italiano vuelven a mostrar esa indecisión. A la pregunta “el uso del sistema en mis tareas podría incrementar mi productividad” la media grupal (3,58) sigue mostrando indecisión. Los tres grupos por separado muestran el mismo grado de satisfacción puntuando mejor el grupo de inglés (3,98). A la pregunta “el uso del sistema podría hacer más fácil mi trabajo” la media global (3,77) continua con esa indecisión ante las preguntas relacionadas con el sistema.

El grupo de inglés (4,29) sigue estando entre de acuerdo e indeciso. El grupo de francés (3,56) y el de italiano (3,06) comparten el mismo grado de indecisión. A la pregunta “este sistema de búsqueda de respuestas es fiable” la media global (3,68) grupal indica un desconocimiento sobre la fiabilidad del sistema. El grupo de inglés (4,25) muestra un mayor grado de confianza en la fiabilidad mientras que esta confianza va decayendo en el de francés (3,55) e italiano (2,81).

A la pregunta “la ayuda ofrecida por el sistema es suficiente” se obtienen los segundos peores resultados en el cuestionario con una media global (2,99) muy baja mostrando su desacuerdo la ayuda ofrecida por el programa. El promedio entre los tres grupos no es significativo no hay diferencias entre las tres lenguas. A la pregunta “la tecnología utilizada por el sistema y la información que ofrece están actualizados” la media global (3,54) indican desconocimiento sobre este aspecto. El promedio de los tres grupos no es significativo, expresan el mismo grado de satisfacción de indecisión. A la pregunta “La información facilitada por el sistema es fácil de comprender” la media grupal (4,04) indica mejor valoración relativa, se muestran de acuerdo en

ciertos aspectos. El grupo de inglés (4,58) tiene el mismo grado de satisfacción que la media, pero el grupo de francés (3,90) y el de italiano (3,22) se muestran indecisos. A la pregunta “la información facilitada por el sistema es relevante” la media grupal (3,66) indica desconocimiento de la materia. El grupo de inglés (4,29) muestra mayor conocimiento que los del grupo de francés (3,34) y mucho más que los del grupo de italiano (2,88) que están en desacuerdo parcialmente. A la pregunta “la información facilitada por el sistema es completa” la media obtenida (3,22) grupal considera que no es del todo completa ni incompleta. El grupo de inglés (3,76) y el de francés (3,10) valoran con un grado de satisfacción igual a la media y el grupo de italianos (2,38) muestra su desacuerdo parcial. A la pregunta “la información facilitada por este sistema está personalizada” se obtiene la puntuación más baja estando la media global grupal (2,77) en desacuerdo. El grupo de inglés (3,05) se muestra indeciso, el grupo de francés (2,71) y el de italiano (2,34) muestran su desacuerdo. Los resultados obtenidos de los 17 ítems agrupados en 4 factores muestran valoraciones dispares, dando puntuaciones muy altas a la facilidad y utilidad del sistema y más bajas a la calidad del servicio y de la información prestada por el mismo. Estos resultados son similares tanto a nivel global como por lengua.

2.8.6.4. Correlaciones entre el grado de satisfacción en relación a la facilidad de uso, utilidad, calidad del servicio y calidad de la información.

Antes de comentar el grado de correlación entre los cuatro factores realizaré un breve comentario sobre la correlación de la variable edad y sexo en relación a las respuestas de los alumnos. No hay correlación significativa entre la variable edad y la respuesta de los alumnos. En cuanto a la variable sexo no hay grandes diferencias en la respuestas a excepción de las preguntas: “Utilizar el sistema podría mejorar mi eficacia en el desarrollo de las tareas académicas” ($P= 0,047$) y “Este sistema de búsqueda de respuestas es fiable” ($P= 0,038$) en el que la mujeres dan mejor puntuación. Los resultados utilizados para interpretación se encuentran contenidos en las tablas, de estos se tomaron en cuenta aquellos cuya relación tuviera un valor superior o igual a 0,6 debido a que se considera que este valor implica una relación fuerte. A nivel global los cuatro factores nos dan como resultado una correlación más alta entre la calidad de servicio y la calidad de la información (668) cuanto mayor sea el grado en que un alumno muestra su comprensión y está de acuerdo con la situación, capacidad de respuesta y tiempo de reacción del servicio, más satisfecho se mostrará con la calidad de los datos proporcionados por el sistema. Esta correlación también se refleja a nivel grupal de francés (763) e italianos (754).

2.8.7. CONCLUSIONES

Tras el análisis de los resultados obtenidos respecto al SBR de HONqa los cuales han permitido evaluar la satisfacción de los alumnos de la Facultad de Traducción e interpretación de Granada se ha comprobado que el sistema les resulta a los usuarios-alumnos fácil y útil para la recuperación de información, pero se muestran indecisos al evaluar la calidad del servicio y la calidad de la información proporcionada. Esto es debido a que el interfaz de sistema es simple y fácil de manejar y su utilidad se aprecia en seguida. Pero por otro lado no se ven capaces de evaluar la calidad de la información proporcionada puesto que ellos muestran un nivel bajo de conocimientos médicos. Otro aspecto interesante es que si comparamos los resultados obtenidos en los tres grupos de lengua, los del grupo de inglés en general se muestran más satisfechos con el sistema ya que han dado a los 17 ítems los valores más altos. A continuación le sigue el grupo francés y por último los de italiano. En definitiva desde el punto de vista del usuario, la satisfacción de sus necesidades de información, es la medida, de eficacia del sistema por excelencia. La opinión de los usuarios sobre la satisfacción o insatisfacción, deriva directamente de la calidad de un servicio, de sus características y de la información que le proporciona. Constituyendo una información muy importante acerca del éxito o fracaso del sistema para cubrir sus expectativas. No es suficiente con que el sistema funcione, es necesario que el servicio que se presta sea satisfactorio para el usuario y que éste así lo perciba.

PARTE III. CONCLUSIONES

CAPÍTULO 3

3.1. CONCLUSIONES GENERALES

El usuario actual confía en recuperar información específica y de calidad que responda a sus necesidades. Los SBR presentan una interesante alternativa a la RI intentando satisfacer sus exigencias y demandas. Junto con el desarrollo de estos sistemas, los avances en MLIR y CLIR ofrecen un panorama de innovación y progreso que alienta a diseñar nuevos SRI que se aproximen mejor a la comunicación humana. Sin embargo, a pesar de las numerosas investigaciones realizadas y del aumento de SBR monolingües y multilingües, todavía es necesario focalizar los estudios en aspectos aún no resueltos como es la comunicación translingüe efectiva, y por ende, la implementación de los recursos y las herramientas que favorecen la traducción de las consultas, palabras claves, documentos y/o respuestas.

Ante la situación expuesta surge la necesidad de desarrollar esta tesis doctoral, la cual pretende analizar y evaluar los SBR monolingües y multilingües desde diferentes parámetros para conocer mejor su funcionamiento y poder emitir un juicio razonable sobre la situación de la traducción en MLIR y CLIR.

En primer lugar, el presente trabajo cuenta con un primer capítulo formado por los antecedentes y la justificación de la misma. El apartado de los antecedentes se divide en tres capítulos en donde se ha ofrecido una visión general del estado de la cuestión que avala los estudios realizados. En el apartado 1.1.1. "Recuperación de Información", se hace un breve recorrido por la historia de estos sistemas y sus características para poder comprender la aparición de nuevos sistemas más sofisticados que tambalean los cimientos de los sistemas tradicionales en RI. Este apartado se divide a su vez en tres secciones en donde se especifican la historia y funcionamiento de los sistemas de RI monolingües, multilingües y translingües.

El apartado 1.1.2. "Sistemas de búsqueda de respuestas" trata de desgranar la evolución, arquitectura, funcionamiento y particularidades de los sistemas objeto de esta tesis doctoral. A su vez está dividida en seis secciones en donde se incluye un listado actualizado de los SBR desarrollados y disponibles en la Web.

El apartado 1.1.3. La Traducción en CLIR se describe la relación existente actualmente entre los sistemas CLIR con la traducción, centrandó nuestro estudio en los recursos y herramientas

lingüísticos. Para ello, en primer lugar, se definen las principales características de la disciplina y profesión de la Traducción, así como las relaciones interdisciplinares con otras ramas. Las investigaciones realizadas en cualquier rama pueden ser de mejora en el resto. En este apartado se incluye una importante sección en donde se tratan los recursos utilizados por estos sistemas para la traducción. Se ha realizado una clasificación propia que se dividen en dos grandes grupos: recursos lingüísticos y herramientas lingüísticas. En la sección “Recursos lingüísticos”, se ha realizado una descripción de todos los recursos utilizados por los SBR multilingües como son las bases de datos, los diferentes tipos de corpus, las diferentes clases de diccionarios, las enciclopedias, EuroWordNet, las ontologías, las páginas web y los tesauros. Por otra parte, en la sección “Herramientas lingüísticas”, se han descrito las dos herramientas más usuales en estos sistemas, gramáticas computacionales y traductores automáticos, y los diferentes tipos de cada uno de ellos.

En el apartado 1.2. “Justificación” se describe en primer lugar las motivaciones que han impulsado a realizar los diferentes estudios presentados, las hipótesis iniciales que se plantearon al inicio del trabajo y los objetivos generales y específicos que han marcado las pautas de la presente tesis doctoral. En el siguiente apartado de estas conclusiones, se desgranarán los objetivos cumplidos.

Tras ello, se resume la metodología general de la tesis doctoral que es ampliada con tres subapartados donde se explican los métodos y materiales específicos de los diferentes estudios realizados. Se ha dividido en tres grandes bloques: en primer lugar, los dos primeros estudios que versan sobre el análisis de los recursos y herramientas lingüísticas utilizados en los SBR; en segundo lugar, el grueso principal de la tesis con la evaluación de los sistemas, que lo compone un total de cinco artículos realizados; y, finalmente, el último estudio que analiza la satisfacción de los usuarios de los SBR desde una perspectiva terminológica.

El capítulo II “Estudios realizados” está formado por ocho estudios que han sido publicados en editoriales nacionales y revistas internacionales de prestigio que siguen el sistema de evaluación por pares, estando seis de ellas indexadas en las principales bases de datos científicas.

El primer bloque lo conforman los artículos “Language resources used in multi-lingual Question Answering Systems” y “Language Resources for Translation in Multi-lingual Question Answering Systems”, publicados respectivamente en la revista internacional *Online Information Review* y la revista online *Translation Journal*. El objetivo de estos estudios ha sido analizar los recursos y herramientas lingüísticos utilizados desde la perspectiva doble de la traducción y la RI, con la intención de comprender las características, ventajas y desventajas de cada uno de los recursos, y extraer los puntos de convergencia y divergencia en sus mecanismos.

El segundo apartado lo componen los estudios centrados en el funcionamiento de los sistemas. Para ello, en el artículo “Question Answering Track Evaluation in TREC, CLEF and NTCIR”, publicado en la serie *Advances in Intelligent Systems and Computing*, se ha analizado qué medidas de evaluación han sido durante estos últimos quince años las más utilizadas para el análisis y evaluación de la efectividad de estos sistemas. Como se explicará a continuación, las medidas tradicionales con sus respectivas variantes siguen siendo las más utilizadas por los evaluadores en las diferentes *tracks* de los tres principales congresos. En los cuatro estudios posteriores se han evaluado de diferentes modos los SBR monolingües y multilingües de dominio general y específicos, todos accesibles en la Web. Los artículos presentados son “Question-Answering Systems as Efficient Source of Terminological Information: Evaluation”, publicado en *Health Information and Library Journal*; “Open- vs. Restricted- Domain Question Answering Systems in the Biomedical Field”, publicado en *Journal of Information Science*; “Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general”, publicado en la *Revista Española de Documentación Científica*; y, “Multilingual Question-Answering System in biomedical domain on the Web: an evaluation”, publicado en la edición *Multilingual and Multimodal Information Access Evaluation* de *Lecture Notes in Computer Science*. Las colecciones de preguntas utilizadas fueron juzgadas conforme a la metodología TREC, y en base a estas valoraciones se aplicaron diferentes medidas de evaluación mediante las que se ilustra claramente la eficacia del funcionamiento de los sistemas analizados.

Finalmente, el último estudio que completa esta investigación se denomina “Satisfacción de usuarios del ámbito de la traducción en el uso de sistemas de búsqueda multilingüe de respuestas como recurso de información terminológica”, publicado en el libro *TIC, trabajo colaborativo e interacción en Terminología y Traducción* de la editorial Comares. Este estudio se centra en un concepto que en los últimos años ha adquirido una notable importancia tanto en el ámbito de los servicios como de las unidades de información, la satisfacción de los usuarios. Si no se estudia y analiza en profundidad las necesidades y opiniones de los usuarios sobre los SBR, todos los estudios anteriores pierden relativamente su interés.

3.2. CONCLUSIONES VERSUS HIPÓTESIS

Durante toda la historia de humanidad el conocimiento, se comunica, se guarda y se maneja en la forma de lenguaje natural. Actualmente, con el desarrollo de procesamiento natural, las

computadoras pueden ayudar al procesamiento de este conocimiento. Sin embargo, lo que es conocimiento para nosotros –los seres humanos– no lo es para las computadoras. Para combatir esta situación, se dedica mucho esfuerzo al desarrollo de la ciencia que se encarga de habilitar a las computadoras a entender el texto. Se entiende al PLN como la tarea de estudiar, diseñar e implementar sistemas computacionales capaces de utilizar y comprobar el lenguaje natural tal y como lo usamos los humanos, es decir, posibilitando una comunicación fluida y eficaz (García Cumbreras, 2009).

Como resulta habitual en otros ámbitos científicos, en el campo de la RI y, concretamente, en lo que se refiere a CLIR, la investigación y la industria transcurren por caminos que nunca llegan a encontrarse. Efectivamente, muchas de las cuestiones que ya están resueltas por parte de los investigadores nunca ven su aplicación en el mundo real y los avances realizados no llegan a comercializarse ni difundirse.

Aunque esta situación responde a causas diversas y complejas, una de ellas puede estar relacionada con el hecho de que las más importantes campañas de evaluación (TREC, CLEF y NTCIR) tienden a centrarse en aquellos aspectos del funcionamiento del sistema que pueden medirse fácilmente de forma objetiva e ignoran otros que son igualmente o, incluso, más importantes para el desarrollo general del sistema (Braschler *et al.*, 2007). De esta forma, características tales como la calidad lingüística o la estabilidad del sistema (es decir, los criterios de eficiencia) han sido casi siempre ignoradas.

Por otro lado, es llamativa la escasez de estudios de usuarios en esta área. Y más aún si se tiene en cuenta que aspectos tales como la usabilidad cobran un enorme protagonismo para los usuarios de los sistemas MLIR. En este entorno, los usuarios pueden necesitar orientaciones en el proceso de la formulación de las preguntas y de la asistencia en la interpretación de los resultados.

Con el fin de conseguir una mayor conexión entre investigación y desarrollo se ideó la puesta en marcha de esta tesis doctoral, donde se plantearon unas hipótesis y objetivos que se satisfacen en las ocho publicaciones presentadas desde el periodo comprendido de 2010 a 2015.

La primera hipótesis planteada hacía referencia a la necesidad de aportar nuevas visiones en el problema de la RI monolingüe, MLIR y CLIR mediante el análisis y evaluación de los SBR multilingües. Con los ocho estudios presentados y avalados en editoriales y revistas internacionales, hemos avanzado en la investigación y análisis de estos sistemas. Las citas e interés surgido en la comunidad internacional acreditan esta hipótesis.

La segunda hipótesis afirmaba que una evaluación real del funcionamiento de los SBR monolingües y multilingües desde diferentes perspectivas proporciona un *feedback* de la situación real. Los ocho estudios realizados presentan tres vertientes acotadas de la situación ya

que se han realizado estudios exhaustivos sobre los recursos lingüísticos utilizados, sobre el funcionamiento real de los SBR disponibles en la Web y sobre la opinión de los usuarios finales desde una perspectiva terminológica.

Desde el principio del presente trabajo hemos elogiado las ventajas fehacientes que presentan los SBR (monolingües y multilingües) frente a los tradicionales SRI. Estas ventajas han sido apreciadas por las grandes empresas tecnológicas, tales como IBM, que han creado sus propios sistemas privados para el uso de sus trabajadores. Como hemos mostrado en los cuatro estudios realizados en la evaluación de los sistemas, efectivamente estos presentan y diversifican las posibilidades de las necesidades de búsqueda de los usuarios, ya sean de dominio abierto o de dominio especializado. No obstante, los problemas actuales que presentan muchos de ellos, debido a la no actualización del software y las bases de datos internas, así como las restricciones propias de cualquier tipo de SRI, nos confirma que todavía queda muchos avances que realizar y que tal vez hasta dentro de unos años no podamos acceder con plenitud a ellos.

En relación con la última afirmación realizada, se encuentra el estado de la siguiente hipótesis planteada. En nuestra introducción abogábamos por la necesidad de determinar cuáles son las herramientas lingüísticas que más afectan a la eficacia de su funcionamiento para la mejora y avance de los SBR multilingües. Efectivamente, en los artículos 1 y 2 se presenta una exhaustiva revisión de las herramientas y los recursos lingüísticos usados por los desarrolladores para solventar el problema de la traducción en estos sistemas.

Por último, en la quinta hipótesis se subrayaba la importancia de los usuarios en la evaluación real de los SBR monolingües y multilingües. Aunque todas las evaluaciones del funcionamiento de los SBR disponibles en la Web han sido realizadas por especialistas en el área en cuestión, veíamos necesaria analizar las opiniones y satisfacción de los usuarios mediante una *checklist* previamente ratificada.

3.3. CONCLUSIONES PARTICULARES

3.3.1. SOBRE LA IDENTIFICACIÓN, EL ANÁLISIS Y LA EVALUACIÓN DE LOS RECURSOS Y HERRAMIENTAS LINGÜÍSTICOS EN LOS SBR (OBJETIVOS 1 Y 2)

A pesar de la inmensa cantidad y variedad de herramientas de búsqueda web existentes en la red el acceso a la información desde una perspectiva multilingüe es algo que aún no está resuelto

adecuadamente. Posiblemente la capacidad multilingüe de Internet no se esté aprovechando convenientemente. El hecho de que no se hayan integrado múltiples mecanismos de apoyo lingüístico en los SRI provoca que el acceso a recursos en diferentes lenguas no se esté realizando de una forma verdaderamente efectiva.

En este sentido, se plantean los dos primeros trabajos de esta tesis doctoral. La necesidad de realizar búsquedas multilingües o translingües es un hecho, y la demanda de este tipo de búsquedas continúa aumentando junto al crecimiento de la Web. Los experimentos realizados han demostrado que la recuperación translingüe es perfectamente factible con un nivel de eficiencia cercano al de las búsquedas monolingües.

Cuestiones tales como la mejora de las herramientas lingüísticas utilizadas y la interacción con el usuario son asuntos clave para optimizar el funcionamiento de los sistemas CLIR. Como se ha comprobado, las herramientas que ofrecen actualmente los SBR analizados siguen siendo demasiado pobres, ya que solventan el problema del sistema en cuestión, pero no se pueden extrapolar de manera generalizada al resto de SBR. Algunos autores (Kishida, 2008) proponen adoptar una aproximación semi-automática, que incorpore algunas capacidades humanas, es decir, aumentar la eficacia mediante un SBR interactivo que permita a los usuarios seleccionar la traducción correcta. El SBR multilingüe HonQA, objeto de estudio de nuestra publicación 7, ofrece en cierto modo esta opción, ya que el usuario puede calificar las respuestas como correctas o incorrectas, pero no hay ningún informe de los desarrolladores sobre las mejoras o cambios realizados.

Para mejorar los SBR multilingües o translingües desde una perspectiva lingüística, se debería explorar simultáneamente estos dos enfoques —el automático y el que propone la participación activa del usuario—. Esto nos lleva a la falta de una adecuada metodología de evaluación para tareas interactivas. Los actuales parámetros de evaluación, diseñados para sistemas automáticos, no reflejan aspectos específicos interactivos que son de especial interés en este tipo de tarea. Sería necesario tener en cuenta todo el proceso de búsqueda y de razonamiento llevado a cabo por el usuario, en lugar de considerar sólo la respuesta final del sistema y el documento que la soporta (López-Ostenero *et al.* 2008).

A veces se ha llamado a CLIR “el problema de encontrar los documentos que usted no puede leer,” lo que, así planteado tiene un valor discutible. El debate no debe centrarse únicamente en si se puede leer lo que se encuentre, sino que la pregunta que debe responderse es si podemos permitirnos el lujo de no saber lo que existe en otras lenguas. Tal vez esa pregunta podría haber sido contestada afirmativamente en el pasado, pero parece poco probable que el siglo XXI sea tan tolerante con esa miopía (Oard *et al.*, 2008).

3.3.2. SOBRE LA EVALUACIÓN DE LOS SBR (OBJETIVO 3)

En los SBR, tanto monolingües como multilingües, la importancia de recuperar los resultados más relevantes en primer lugar es tan esencial como la recuperación en sí. Y es que estos sistemas suelen ser utilizados por usuarios que persiguen recuperar información rápida y eficaz en un breve espacio de tiempo y no suelen examinar un número muy alto de respuestas.

Por ello, se han realizado cinco estudios que han evaluado de diferentes modos los SBR de dominio general y específicos accesibles desde la Web mediante dos colecciones de preguntas. Para la evaluación de estos sistemas, se han usado las medidas de evaluación que cuentan en la actualidad con gran aceptación y consenso en la comunidad investigadora, tal como lo demuestra el hecho de que los reputados foros evaluados en el tercer estudio las incorporen a su modelo de evaluación. Los estudios realizados revelan resultados alentadores debido a que presentan este tipo de herramientas como una nueva posibilidad para obtener información precisa y fiable en un corto período de tiempo.

De gran interés resultan los SBR de dominio especializado cuya arquitectura afecta al tipo de preguntas que permiten plantear y a las respuestas que se esperan obtener. Las características de este tipo de sistemas beneficia a las diferentes técnicas de SBR, de modo que algunos dominios son particularmente apropiados para el desarrollo de las mismas (Mollá y Vicedo, 2007). Un sistema parcialmente libre que ilustra bastante bien esta situación es *Wolfram Alpha*, un buscador desarrollado por la compañía *Wolfram Research* que se dedica a realizar operaciones matemáticas y a responder a preguntas computacionales.

En el plano de MLIR, el diseño y desarrollo de herramientas multilingües, o en otros idiomas diferentes al inglés, han fomentado la investigación y la creación de nuevos instrumentos para la búsqueda y RI multilingüe (Ferrández Escámez, 2008). No obstante, este interés suscitado ha tenido simplemente un tímido reflejo en los SBR disponibles en la Web.

3.3.3. SOBRE LA EVALUACIÓN DE LA SATISFACCIÓN DE LOS USUARIOS (OBJETIVO 4)

Los estudios realizados en los primeros siete artículos presentados se han centrado en el funcionamiento real de los SBR, lo cual nos ha permitido valorar objetivamente las ventajas y

restricciones que ofrecen al usuario como SRI. Sin embargo, el reconocimiento y aceptación como un SRI pleno no será otorgado hasta confirmar la eficacia de la interactividad de los usuarios con los sistemas. Como ha sucedido en otros populares sistemas de RI más comunes, véase el caso de *Yahoo* o *Altavista*, los usuarios han ido progresivamente abandonando su uso a favor de otros sistemas más completos, como es el caso de *Google*. Las propias empresas que gestionan estos sistemas ofrecen datos cuantitativos de su popularidad que nos permite emitir afirmaciones como las anteriores.

Sin embargo, el hecho de que los SBR no hayan sido todavía explotados en demasía y tengamos pocos disponibles en la Web, nos obliga a realizar estudios que se centren en la satisfacción de los usuarios. Esta evaluación nos permite entender las necesidades de los internautas e identificar las dimensiones y factores en el desarrollo de un SRI con el fin de mejorar su aceptación.

En definitiva desde el punto de vista del usuario, la satisfacción de sus necesidades es otra medida de eficacia para el análisis y la evaluación de los SBR monolingües y multilingües. En el caso de nuestro estudio, nos hemos centrado exclusivamente en los multilingües ya que consideramos que estos sistemas son una estupenda alternativa terminológica a los tradicionales SRI, además en cuatro de los ocho estudios presentados fueron especialistas los que valoraron objetivamente la información recuperada.

Por todo ello, concluimos que la opinión de usuarios sobre su grado de satisfacción o insatisfacción deriva positivamente en la calidad del servicio, en sus características y en la información que presta, proporcionándonos tanto a los desarrolladores como a los investigadores una información muy importante acerca del éxito o fracaso del sistema para cubrir expectativas. No es suficiente con que el sistema funcione, es necesario que el servicio que se presta sea satisfactorio para el usuario y que éste así lo perciba.

3.5. CONSIDERACIONES FINALES Y LINÉAS DE INVESTIGACIÓN FUTURAS

La presente tesis doctoral ofrece una visión amplia de la influencia de la traducción en la situación actual de los SBR monolingües y multilingües. Para ello, durante un periodo de cinco años los directores de esta tesis doctoral, la Dra. D^a María Dolores Olvera Lobo y el Dr. D^a Bryan John Robinson Fryer, y yo, como autora, hemos conjuntamente investigado sobre el ámbito en cuestión. Los perfiles diferenciados, Información y Comunicación por un lado y,

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

Traducción e Interpretación por otro, nos ha permitido abordar todos los estudios desde dos ámbitos distintos, lo que ha enriquecido enormemente las aportaciones presentadas.

En esta tesis presentamos ocho de los catorce estudios que hemos publicado sobre el tema durante estos años en revistas internacionales, editoriales de prestigio y congresos internacionales. A pesar de que la mayoría de los expertos en esta área son especialistas en Ciencias de la Tecnología de la Computación o Inteligencia Artificial, la comunidad investigadora ha respaldado ampliamente nuestros estudios ya que todos los trabajos han pasado exhaustivas revisiones por pares y han sido citados por expertos en numerosas obras.

A pesar de las numerosas ventajas que ofrecen a los usuarios estos nuevos SRI, las restricciones latentes de los SBR (amplias y actualizadas base de datos, actualización de recursos lingüísticos, coste económico, etc.) limitan su futuro más inmediato, así como su puesta en marcha de manera gratuita en la Web. Algunos SRI, como *Google*, presentan en su interfaz una versión alternativa a los SBR pero que en esencia permite al usuario acceder a información breve del tema en cuestión a partir de los datos publicados en la enciclopedia Wikipedia.

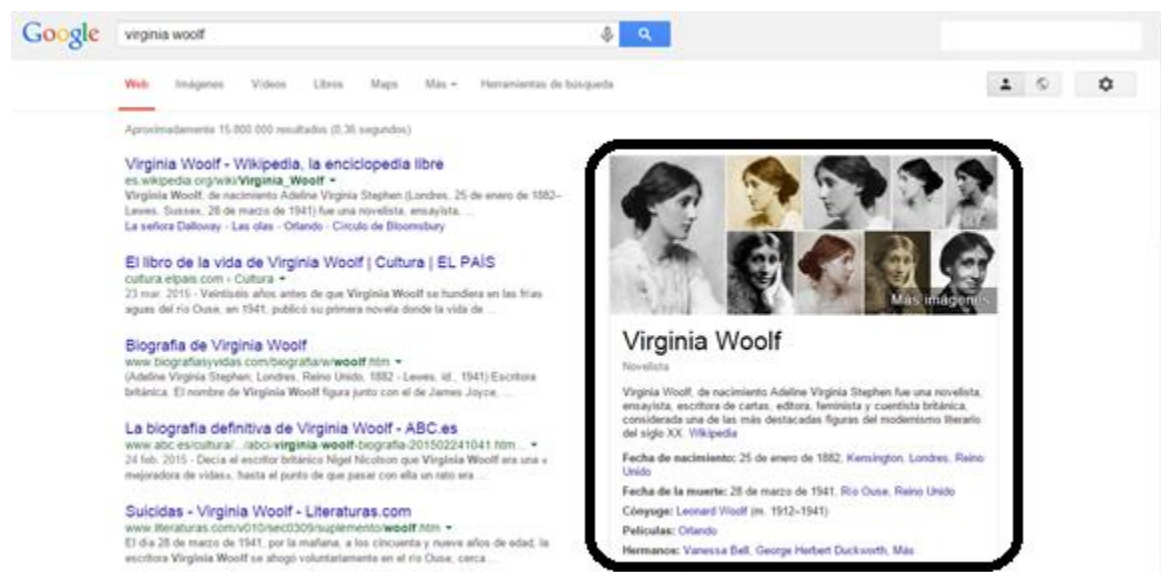


Figura 50. Interfaz de *Google* al buscar información sobre Virginia Woolf

Consideramos que esta línea de trabajo tiene continuidad ampliando el análisis y evaluación de los nuevos sistemas multilingües que han aparecido en los últimos años. La relación de CLIR con la disciplina de la Traducción nos parece de extrema importancia dado que los recursos translingües van a seguir creándose en los próximos años. La realización de estudios interdisciplinarios y transdisciplinarios en RI y CLIR deben ser continuados para fomentar el

enriquecimiento de esta área de investigación. Por todo ello, creemos que los futuros trabajos deben abordar las siguientes perspectivas:

- ✓ Ampliar los estudios previamente realizados por Olvera-Lobo y Gutiérrez-Artacho (2010, 2011a, 2011b, 2011c, 2012, 2013a, 2013b, 2014a, 2014b, 2014c, 2014d, 2015) y Narasimha Raju y colegas (2014) sobre la situación de CLIR desde el enfoque de la traducción.
- ✓ Analizar las nuevas tendencias en el ámbito de RI que incorporen técnicas y herramientas de traducción.
- ✓ Evaluar la calidad de las traducciones proporcionadas por los diferentes sistemas CLIR disponibles fijando el punto de interés en los recursos lingüísticos utilizados.
- ✓ Analizar el impacto de la traducción y sus consecuencias en el correcto funcionamiento de los sistemas CLIR.

REFERENCIAS

A

Abaitua, J. (2002). *Introducción a la traducción automática (en diez horas)*. Disponible en http://paginaspersonales.deusto.es/abaitua/konzeptu/ta/mt10h_es/ta10h-1es.htm

Abdou, S.; Savoy, J.; Ruch, P. (2006). "Dépister efficacement de l'information dans une banque documentaire: L'exemple de MEDLINE". En: *Actes du XXIVème Congrès INFORSID*. 129-143.

Aceves Pérez, R.M. (2008). *Búsqueda de Respuestas en Fuentes Documentales Multilingües*. Tesis doctoral. Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico.

Adiwibowo, S.; Adriani, M. (2007). "Finding answers using resources in the internet". En: *Working notes for the CLEF 2007 Workshop*. Disponible en: http://www.clef-campaign.org/2007/working_notes/adiwibowoCLEF2007.pdf

Adafre, S.F.; Rijke, de M. (2006). "Finding Similar Sentences across Multiple Languages in Wikipedia". En: *Proceedings of the EACL Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, 62-69.

Ajzen, I. (1991). "The theory of planned 229recisión". *Organizational and Human Decision Processes* 50, 179-211.

Alcántara Plá, M. (2007). *Introducción al análisis de estructuras lingüísticas en corpus. Aproximación semántica*. Madrid: UAM Ediciones. (ISBN: 978-84-8344-069-8)

Alfonseca, E.; De Boni, M.; Jara, J.L.; Manandhar, S. (2002). "A prototype Question Answering system using syntactic and semantic information for answer retrieval". En: *Proceedings of the 10th Text Retrieval Conference (TREC-10)*.

Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swan, R., Xu, J. (1998). "INQUERY does battle with TREC-6". *NIST SPECIAL PUBLICATION SP*, 169-206.

Amati, G.; Van Rijsbergen, C.J. (2002). "Probabilistic models of information retrieval based on measuring the divergence from randomness". *ACM Transactions on Information Systems* 20 (4), 357-389.

Athenikos, S.J.; Han, H. (2010). "Biomedical question answering: A survey". *Computer Methods and Programs in Biomedicine* 99 (1), 1-24

B

Baeza-Yates, R.; Ribeiro-Nieto, B. (1999). *Modern Information Retrieval*. Nueva York: ACM Press.

Baeza-Yates, R.; Ribeiro-Nieto, B. (2011) *Modern Information Retrieval: the Concepts and Technology behind search (2nd Edition)*. Nueva York: ACM Press.

Baker, M. (1995). "Corpora in translation studies: an overview and some suggestion for future research". *Target* 7 (2). 223-243.

Bel, N.; Marimon, M. (2008). "Clarin: Common Language Resources and Technology Infrastructure". *Procesamiento del Lenguaje Natural*, 41, 311-312.

Belkin, N.J. (1980). "Ineffable concepts in information retrieval. In K. Sparck Jones (ed.), *Information Retrieval Experiment*". London, Butterworths, 44-58.

Belkin, N.J.; Croft, W. B. (1987) "Retrieval techniques". *Annual Review of Information Science and Technology*, 22, 109-146.

Belkin, N.J.; Vickery, A. (1985). *Interaction in Information Systems: A Review of Research from Document Retrieval to Knowledge-based systems (LIR Report No 35)*. Londres: The British Library.

Blair, D.C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier Science Publishers.

Blair-Goldensohn, S.; McKeown, K.; Schlaikjer, A.H. (2004). "Answering Definitional Questions: A Hybrid Approach", *New Directions in Question Answering*, 4, 47-58.

Blair-Goldensohn, S.; McKeown, K.; Schlaikjer, A.H. (2003). "A hybrid Approach for QA Track Definitional Questions". *Proceedings of TREC 2003*, Gaithersburg, Maryland, 336-343.

Braschler, M.; Di Nunzio, G.; Ferro, N.; Gonzalo, J.; Peters, C.; Sanderson, M. (2007). "From CLEF to treble CLEF: Promoting technology transfer for multilingual information retrieval". En C. Thanos, F. Borri y A. Launaro (Eds.), *Working Notes of the Second DELOS Conference on Digital Libraries*, en: www.trebleclef.eu/publications.php

Breck, E.J.; Burger, J.D.; Ferro, L.; Hirschman, L.; House, D.; Light, M. (2000). "How to Evaluate Your Question Answering System Every Day and Still Get Real Work Done". *Proceedings of Second International Conference on Language Resources and Evaluation*. Atenas: Grecia.

Bos, J.; Nissim M. (2006). "Cross-Lingual Question Answering by Answer Translation". En C. Peters (ed.), *Working Notes for the Cross Language Evaluation Forum (CLEF) 2006*.

Buitelaar, P.; Cimiano, P.; Frank, P.; Hartung, M.; Racioppa, S. (2008). "Ontology-based information extraction and integration from heterogeneous data sources". *International Journal of Human-Computer Studies*, 66, 759-788.

Buckley, C.; Voorhees, E. M. (2000). "Evaluating evaluation measure stability". *SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 33-40.

C

- Callan, J.P.; Croft, W.B.; Stephen, M. (1992). "The INQUERY Retrieval System". *DEXA*, 78-83.
- Cao, Y.G.; Ely, J.; Antieau, L.; Yu, H. (2009). "Evaluation of the Clinical Answering Presentation". *Proceedings of the Workshop on BioNLP*, 171-178.
- Cabré i Castellví, M.T. (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona: Editorial Antártida/ Empuries.
- Cabré i Castellví, M.T. (2000). "El traductor y la terminología: necesidad y compromiso". *Panace@*, 1 (2), 2-4.
- Chang, G., Healey, M., McHugh, J.A.M., Wang, T.L. (2001). *Mining the World Wide Web: an information search approach*. Norwell, Massachusetts: Kluwer Acad. Publishers.
- Charniak, E.; Altun, Y.; Salvo Braz, R.; Garrett, B.; Kosmala, M.; Moscovich, T.; Pang, L.; Pyo, C.; Sun, Y.; Wy, W.; Yang, Z.; Zeller, S.; Zorn, L. (2000). "Reading Comprehension Programs in a Statistical-Language-Processing Class". *ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*. Washington: Seattle.
- Chen, A.; Gey, F.C. (2001). "Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval". *The Tenth Text Retrieval Conference (TREC 2001)*, 529-533.
- Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I. (2008). "Novelty and diversity in information retrieval evaluation". *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 659-666
- Cleverdon, C. (1997). "The Cranfield tests on index languages devices". En: Sparck Jones, K.; Willett, P. (eds.), *Readings in information retrieval*. San Francisco: Morgan Kaufmann. 47-59.
- Chowdhury, G.G. (1999) *Introduction to modern information retrieval*. Londres: Library Association.

Codina Bonilla, L. (2011). "Sistemas de búsqueda y obtención de información". En F. Casheda Seijo, F., J.M. Fernández Luna y J.F. Francisco Huete Guadix (eds.), *Recuperación de información: un enfoque práctico y multidisciplinar*, Madrid: Ra-Ma

Costa L.F., Santos, D. (2007). "Question Answering Systems: a partial answer". *SINTEF*, Oslo.

Crestani, F.; Pasi G. (2000). *Soft Computing in Information Retrieval: Techniques and Applications*. Nueva York: Physica-Verlag (A Springer-Verlag Company).

Croft, W.B.; Metzler, D.; Strohman, T. (2010). *Search Engines: Information Retrieval in Practise*. Addison Wesley.

Crouch, D.; Saurí, R.; Fowler, A. (2005). "AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines". *Palo Alto Research Center*. Disponible en:
http://www2.parc.com/isl/groups/nltt/papers/aquaint_kb_pilot_evaluation_guide.pdf

Cruchet, S.; Gaudinat, A.; Rindflesch, T.; Boyer, C. (2009). "What about trust in the Question Answering world?" *AMIA 2009 Annual Symposium* (San Francisco).

Cui, H.; Kan, M. Y.; Chua, T.S.; Xiao, J. (2004). "A Comparative Study on Sentence Retrieval for Definitional Question Answering". *SIGIR Workshop on Information retrieval for Question Answering (IR4QA)*, Sheffield.

Cucerzan, S.; Agichtein, E. (2005). Factoid Question Answering over Unstructured and Structured Web Content. In TREC, 72, 90.

D

DeLone, W.H.; McLean, E.R. (2003). "Model of Information Systems Success: A Ten-Year Update". *Journal of Management Information Systems*, 19 (4), 9-30

Diéguez, M.I.; Lazo, R.M. (2004). "Tecnologías de la Información y de la Comunicación (TIC) al Servicio del Traductor Profesional". *Onomázein: Revista de Lingüística, Filología y Traducción*, 9(1), 54-74.

Diekema, A.R. (2003). *Translation Events in Cross-Language Information Retrieval: Lexical ambiguity, lexical holes, vocabulary mismatch, and correct translations*. Dissertation. University of Syracuse.

Diekema, A.R.; Yilmazel, O.; Liddy, E.D. (2004). "Evaluation of restricted domain Question-Answering systems". *Center for Natural Language Processing*. Paper 3

Díez Orzas, P. L. (1999). "La relación de meronimia en los sustantivos del léxico español: contribución a la semántica computacional". *Estudios de lingüística del español*, 2., 000-0.

Dwivedi, S.K.; Singh, V. (2013). "Research and Reviews in Question Answering System". *Procedia Technology*, 10, 417—424

E

Ellis, N. C. (1996). "Analyzing language sequence in the sequence of language acquisition: Some comments on Major and Loup". *Studies in Second Language Acquisition*, 18(03), 361-368.

Ely, J.W.; Osheroff, P.N.; Ebell, M.; Bergus, G.; Barcey, L.; Chambliss, M.; Evans, E. (1999). "Analysis of questions asked by family doctors regarding patient care". *British Medical Journal*, 319, 358–361.

Ely, J.W.; Osheroff, J.A.; Gorman, P.N.; Ebell, M.H.; Chambliss, M.L.; Pifer, E.A.; Stavri, P.Z. (2000). "A taxonomy of generic clinical questions: classification study". *British Medical Journal*, 321, 429–432.

F

Fahmi, I. (2009). *Automatic term and relation extraction for medical question answering system*. Tesis doctoral. Groningen Dissertations of Linguistics, 72.

Ferrández Escámez, S. (2009). *Arquitectura multilingüe de sistemas de búsqueda de respuestas basada en ILI y Wikipedia*. María Teresa Vicente-Díez, Paloma Martínez, Ángel Martínez-González, 42, 127-128.

Filatova, I. (2010). *Evaluación de herramientas y recursos informáticos (TAO y ofimática) para la traducción profesional: hacia la configuración de un entorno óptimo de trabajo para el traductor anónimo* (Tesis doctoral, Universidad de Málaga).

Fluhr, C. (1996). *Multilingual Information Retrieval. Survey of the State of the Art in Human Language Technology*. Disponible en: cslu.cse.ogi.edu/HLTsurvey/ch8node7.html

Forner, P.; Giampiccolo, D.; Magnini, B.; Peñas, A.; Rodrigo, Á.; Sutcliffe, R. (2010). "Evaluating multilingual question answering systems at CLEF". *Proceedings of 7th International Conference on Language Resources and Evaluation*. LREC 2010, Malta: 2774—2781.

Frank, A.; Kirefer, H.U.; Xu, F.; Uszkoreit, H.; Crysmann, B.; Jörg, B.; Schäfer, U. (2006). "Question answering from structured knowledge sources". *Journal of Applied Logic, Special Issue on Questions and Answers: Theoretical and Applied Perspectives*, 5, 20-48.

Fukumoto, J.; Kato, T.; Masui, F. (2004). "Question Answering Challenge (QAC-I) an evaluation of question answering tasks at the NTCIRWorkshop 3". *Proceedings of AAAI Spring Symposium on New Directions in Question Answering*. 122-133.

G

García Broncano, R. (2006). *Recuperación y Organización de la Información – Modelos de Recuperación*. Universidad Carlos III de Madrid, España. [Online]. Disponible: <http://modelosrecuperacion.tripod.com/modelosrecuperacion.pdf>

García Cumbreiras, M. Á., Ureña López, L.A., Martínez Santiago, F., and Montejo Ruez, A. (2005) “Búsqueda de respuestas multilingüe: clasificación de preguntas en español basadas en aprendizaje”, *Procesamiento del lenguaje natural*, 34, 31–40, disponible en: www.sepln.org/revistaSEPLN/revista/34/03.pdf.

García-Cumbreiras, M. A.; Ureña-López, L. A.; Martínez-Santiago, F. (2006). “Bruja: question classification for Spanish. Using machine translation and an English classifier”. En: *EACL 2006 Workshop on multilingual question answering – MLQA06*, 39-44.

García-Santiago, L.; Olvera-Lobo, M.D. (2010) “Automatic Web Translators as Part of a Multilingual Question-Answering (QA) System: Translation of Questions”, *Journal of Documentation*, 66, 434–455.

Gatius i Vila, M. (2000). *Using an ontology for guiding Natural Language interaction with Knowledge Based Systems*. Tesis Doctoral. Universidad Politécnica de Cataluña.

Gey, F.C.; Kando, N.; Peters, C. (2005) “Cross-language information retrieval: the way ahead.” *Information processing & management* 41 (3), 415-431.

Graesser A.; Murachver, T. (1985). “Symbolic procedures of question answering”. En A. Graesser y J. Black (eds.), *The Psychology of Questions*, 15–88. Erlbaum.

Green, B. F.; Wolf, A. K.; Chomsky, C.; Laughery, K. (1961). “Baseball: An Automatic Question Answerer”. *Proceedings of the Western Joint Computer Conference*. 19, 219–224.

Greenwood, M. A.; Saggion, H. (2004). “A Pattern Based Approach to Answering Factoid, List and Definition Questions”. *Proceedings of the 7th RIAO Conference (RIAO 2004)*, 232-243.

Grefenstette, G. (1998). *Cross-Language Information Retrieval*. Kluwer academic publishers, Vol. 1.

Grossman, D.A.; Frieder, O. (1998). *Information retrieval: algorithms and heuristics*. Boston: Kluwer Academia Publishers.

Gruber, T.R. (1993). "A Translation Approach to Portable Ontologies". *Knowledge Acquisition* 5 (2), 199-220.

H

Harabagiu, S. M.; Pasca, M. A.; Maiorano, S. J. (2000). "Experiments with open domain textual question answering". *Proceedings of the Coling-2000*. Association for Computational Linguistics / Morgan Kaufmann.

Harman, D.K. (1993). "The first text retrieval conference (TREC-1) Rockville, MD, USA, 4-6 November, 1992." *Information Processing & Management*, 29(4), 411-414.

Harman, D.K. (1998). "Text retrieval conferences (TRECs): providing a test-bed for information retrieval systems". *Bulletin of the American Society for Information Science*, 24 (4), 11-13.

Harter, S.P.; Hert C.A. (1997). "Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods". *Annual Review of Information Science and Technology*, 32, 3-79.

Hernández-Rubio, G. (2008). *Recuperación de Pasajes Orientada a la Resolución de Preguntas con Restricción Temporal*. Tesis doctoral. México: INAOE.

Hiemstra, D.; Kraaij, W.; Pohlmann, R.; Westerveld, T. (2001). "Twenty-one at CLEF-2000: Translation resources, merging strategies and relevance feedback". En C. Peters (ed.), *Proceedings of the CLEF 2000 Cross-language Text Retrieval System Evaluation Campaigns*.

Holstein, R.C.; Lundberg; G.D. (2003) "Use of the Internet for health information and communication". *JAMA*, 290, 2255-2255.

Hovy, E.; Hermjacob, U.; Lin, C. (2001). "The use of external knowledge in factoid qa". En *Proceedings of Text Retrieval Conference (TREC-10)*.

Hunston, S. (2002). *Corpora in Applied Linguistics*; Cambridge: CUP, 2002.

Hutchins, J. (2002). *Machine translation today and tomorrow*. na.

I

Izquierdo, R.; Ferrández, O.; Ferrández, S.; Vicedo, J. L.; Martínez, P.; Suárez, A. (2007). "QALL-ME: question answering learning technologies in a multiLingual and 238recisión238 environment". *Revista procesamiento del lenguaje natural*, 38, 43-47.

J

Jackson, P.; Schilder, F. (2005). "Natural Language Processing: Overview". En Brown (ed.), *Encyclopedia of Language & Linguistics*, 2, 503-518. Amsterdam, Elsevier Press.

Jacquemart, P.; Zweigenbaum, P. (2003). "Towards a Medical Question-Answering System: a Feasibility Study". En Beux, P. L. & Baud, R. (eds.), *Proceedings of Medical Informatics Europe (MIE '03)*, vol. 95 of Studies in Health Technology and Informatics, San Palo, California: 463-468

Jones, K.S.; Jackson, D.M. (1970). "The use of automatically-obtained keyword classifications for information retrieval". *Information Storage and retrieval*, 5(4), 175-201.

Jung, H.; Lee, G.G. (2003). "Multilingual Question Answering with High Portability on Relational Databases". *IEICE Transactions on Information and Systems*, E86-D (2), 306-315.

Jurafsky, D.; Martin, J.H. (2000). "Speech and Language Processing: An Introduction to Natural Language Processing", *Computational Linguistics, and Speech Recognition*; New Jersey: Prentice-Hall.

K

Kangavari, M.R.; Ghandchi, S.; Golpour, M. (2008) "A New Model for Question Answering Systems", *World Academy of Science, Engineering and Technology*, 42.

Katz, B. (1990). "Using English for Indexing and Retrieving". En P.H. Winston y S.A. Shellard (eds.), *Artificial Intelligence at MIT: Expanding Frontiers*, 1, MIT Press, Cambridge, MA.

Katz, B.; Borchardt, G.; Felshin, S.; Shen, Y.; Zaccak, G. (2007). "Answering English questions using foreign-language, semistructured sources". *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*. Irvine: IEEE Computer Society. 439-445.
<http://dx.doi.org/10.1109/ICSC.2007.59>

Katz, B.; Felshin, S.; Yuret, D.; Ibrahim, A.; Lin, J.; Martion, G.; McFarland, A.J.; Temelkuran, B. (2002). "Omnibase: Uniform Access to Heterogeneous Data for Question Answering". *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*, 230–234.

Kaiser, M. (2008). "The QuALiM question answering demo: supplementing answers with paragraphs drawn from Wikipedia". *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*. Stroudsburg: Association for Computational Linguistics. 32–35

Kaiser, M. (2009). *Acquiring Syntactic and Semantic Transformations in Question Answering*. Tesis doctoral. University of Edinburgh, Edinburgh, Escocia, Reino Unido.

Kangavari, S.; Ghandchi, M.; Golpour, M. (2008). "A New Model for Question Answering Systems". *World Academy of Science, Engineering and Technology*, 42.

Kelly, D.; Sugimoto, C.R. (2013). "A systematic review of interactive information retrieval evaluation studies, 1967–2006". *Journal of the American Society for Information Science and Technology*, 64: 745–770.

Kelly, D. (2002). "Un modelo de competencia traductora: bases para el diseño curricular". *Puentes. Hacia nuevas investigaciones en la mediación intercultural* 1, 9-20.

Kolomityets, O.; Moens, M. F. (2011). "A Survey on Question Answering Technology from an Information Retrieval Perspective". *Information Sciences*, 181 (24), 5412–5434

Korfhage, R. R. (1997). *Information Storage and Retrieval*. New York: Wiley Computer Publishing.

L

Lee, M.; Cimino, J.; Zhu, H.R.; Sable, C.; Shanker, V.; Ely, J.; Yu, H. (2006). "Beyond Information Retrieval –Medical Question Answering". *AMIA*. Washington DC, Estados Unidos.

Li, X.; Morie, P.; Roth, D. (2004). "Identification and tracing of ambiguous names: Discriminative and generative approaches". *Proceedings of the National Conference on Artificial Intelligence* , 419-424.

Liddy, E.D. (1998). "Enhanced text retrieval using natural language processing". *Bulletin of the American Society for Information Science and Technology*, 24(4), 14-16.

Liddy, E.D. (2003). *Encyclopedia of Library and Information Science*.

Llisterri, J. (2008). *Tecnologías lingüísticas: los recursos lingüísticos* [on line]. En Universitat Autònoma de Barcelona,
http://liceu.uab.es/~joaquin/language_technology/HLT/tecnol_ling_recursos.html

López, V. (2002). "Posibilidades y realidades de la Traducción Automática". *La Linterna del Traductor*, 3.

Lopez, V.; Unger, C.; Cimiano, P.; Motta, E. (2013). "Evaluating question answering over linked data." *Web Semantics Science Services And Agents On The World Wide Web*, 21:3–13.

López-Ostenero, F. (2002) *Un Sistema Interactivo para la Búsqueda de Información en Idiomas Desconocidos por el Usuario*. Tesis doctoral, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, España.

López-Ostenero, F. (2004). "Búsqueda de información multilingüe: estado del arte". *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*. 8, 22, 11-35.

M

Marcos Marín, F. (1994). *Informática y humanidades*. Madrid: Gredos

Martínez de Sousa, J. (2009). *Manual básico de lexicografía*. Gijón: Ediciones Trea.

Martínez Méndez, J. (2002). *Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en Internet*. Tesis doctoral.

Martínez Santiago, F.; Ureña López, L.A. (2003). "Propuesta para un sistema de recuperación de información multilingüe independiente del lenguaje". *Proceedings of the II Jornadas de Tratamiento y Recuperación de la Información (JOTRI 2003)*, 100-108. Madrid, Spain: Universidad Carlos III.

Mayoral Asensio, R. (2001). "Por la renovación en la formación de traductores e intérpretes: revisión de algunos de los conceptos sobre los que el actual sistema basa su estructura y contenidos". *Sendeban* 12, 311-336.

Mayoral Asensio, R. (2005). "Presentación". En D. Sales (ed.), *La Biblioteca de Babel: Documentación para traducir*. Granada: Comares, ed.; xi-xiii.

Magnini, B. (2002). "Evaluation of Cross-Language Question Answering Systems". CLEF Workshop 2002. Disponible en:

<http://clef.iei.pi.cnr.it:2002/workshop2002/presentations/q-a.pdf>.

Magnini, B.; Negri, M.; Prevete, R. (2001). "Open domain question/answering on the Web". En: *Congress of the Italian Association for Artificial Intelligence*, 2175, 273-284.

Maybury, M. T. (2004). *New Directions in Question Answering*. AAAI/MIT Press.

Meadow, C. T. (1993). *Text Information retrieval Systems*. Academic Press, San Diego.

Miller, G.A. (1986). "Dictionaries in the Mind". *Language and Cognitive Processes*, 1(3), 171-185.

Miller, G.A.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K. J. (1990). "Introduction to wordnet: An on-line lexical database". *International journal of lexicography*, 3(4), 235-244.

Moghadas, S.I., Ravana, S.D. (2013). "Low-cost evaluation techniques for information retrieval systems: A review". *Journal of Informetrics*. 7 (2), 301-312

Moldovan, D. (2003). "Performance issues and error analysis in an open-domain question answering system". *ACM transactions on information systems*, 21 (2): 133-154.

Monz, C. (2003). *From Document Retrieval to Question Answering*. Tesis doctoral. Amsterdam: University of Amsterdam, 2003.

Mollá, D.; Vicedo, J.L. (2005). *Question-Answering in Restricted Domains*. Menlo Park, California, AAAI Press.

Mollá, D.; Vicedo, J.L. (2007). "Question Answering in Restricted Domains: An Overview". *Comput. Linguist.* 33, 1 (March 2007), 41—61

Moreno Ortiz, A. (2000). *Diseño e implementación de un lexicón computacional para lexicografía y traducción automática*. Tesis doctoral. Universidad de Málaga.

Muñoz Martínez, R. (2000). "As Time Goes By: Herramientas de traducción en la enseñanza". En E. Morillas y J. Álvarez (eds.). *Las Herramientas del Traductor*, pp. 397-416. Málaga: Ediciones del Grupo de Investigación Traductología.

N

Nguyen, D.; Overwijk, A.; Hauff, C.; Trieschnigg, D.R.B.; Hiemstra, D.; de Jong, F.M.G. (2009). "WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia". En Peters *et al.* (eds.): *CLEF 2008*, LNCS 5706, 58-65.

Nunzio, G.M.; Ferro, N. (2005). "Appendix A – Results of the Core Tracks and Domain-Specific Tracks". *Working Notes for the CLEF 2005 Workshop*.

O

Oard, D.W. (2001). "Interactive cross-language information retrieval". *SIGIR FORUM*, 35 (1), 1-3.

Oard, D.; Gonzalo, J. (2001). The CLEF 2001 Interactive Track. *Evaluation of Cross-Language Information Systems*, LNCS 2785, Springer.

Ogawa, Y.; Morita, T.; Kobayashi, K. (1991). "A fuzzy document retrieval system using the keyword connection matrix and a learning method". *Fuzzy sets and systems*, 39(2), 163-179.

Olvera-Lobo, M.D. (1999). "Evaluación de sistemas de recuperación de información: aproximaciones y nuevas tendencias". *El profesional de la información*, 11 (4), 4-14

Olvera Lobo, M.D.; Robinson, B.; Castro Prieto, R.M.; Quero Gervilla, E.; Muñoz Martín, R.; Muñoz Raya, E.; Díez Lerma, J.L. (2007). "A professional approach to translator training (PATT)". *Meta: Journal des traducteurs/Translators' Journal*, 52(3), 517-528.

Olvera-Lobo, M.D. (2009). "Cross-language Information Retrieval on the Web". En Cruz-cunha, M.M.,Oliveira, E.F., Tavares, A.J.V. y Ferreira, L.G. (eds.), *Handbook Of Research On Social Dimensions Of Semantic Technologies And Web Services*, Chapter XXXIV, 704–719.

Olvera-Lobo, M.D.; Amo-Valdivieso, M.; Gutiérrez-Artacho, J. (en prensa). "Wikipedia as a source of monolingual and multilingual information about the heritage of Spain". *Transinformação*.

Olvera-Lobo, M.D.; Robinson-García, N. (2009) "Tratamiento lingüístico de las preguntas en los sistemas de búsqueda de respuestas". *El profesional de la documentación*, 18 (2): 180–187.

Olvera-Lobo, M.D.; García-Santiago, L. (2010). "Analysis of errors in the automatic translation of questions for translingual QA systems", *Journal of Documentation*, 66 (3), 434–455.

Olvera-Lobo, M. D.; Gutiérrez-Artacho, J. (2010). "Question-answering systems as efficient sources of terminological information: Evaluation". *Health Information and Library Journal*, 27 (4): 268-274

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011a). "Evaluation of Open- vs. Restricted- Domain Question Answering Systems in the Biomedical Field". *Journal of Information Science*. 37 (2): 152-162.

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011b). "Language resources used in multilingual Question Answering Systems". *Online Information Preview*. 35 (4): 543 – 557

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2011c). "Multilingual Question-Answering System in Biomedical Domain on the Web: An evaluation". *Lecture Notes in Computer Science (LNCS)* 6941: 83:88

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2012) "Language Resources for Translation in Multilingual Question Answering Systems." *Translation Journal*, 16 (2).

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2013a) "Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general". *Revista Española de Documentación Científica*, 36, (2), e009

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2013b) "Nuevas tendencias en Recuperación de Información: búsqueda de respuestas desde la perspectiva de la traducción". *VI Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación*, Las Palmas de Gran Canarias.

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2014a). "Overview of Translation Techniques in Cross-Language Question Answering during the Last Decade". *Encyclopedia of Information Science and Technology*, 3, 4747-4755.

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2014b). "Searching Health Information in Question-Answering Systems". *Handbook of Research on ICTs for Healthcare and Social Services: Developments and Applications*. IGI Global, 2013. 474-490. Web. 6 Feb. 2014.

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2014c). "Question-Answering and E-Health: An advanced perspective in Specialized Information Retrieval". *9ª Conferencia Ibérica de Sistemas y Tecnologías de Información*.

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2014d)." Satisfacción de los usuarios en la búsqueda multilingüe de respuestas como recursos de información terminológica". *TIC, trabajo colaborativo e interacción en Terminología y Traducción*. Ediciones Interlingua. Granada: Comares.

Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. (2015). "Question Answering Track Evaluation in TREC, CLEF and NTCIR". *New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing Volume 353*, 2015, pp 13-22

Ong, C.C.; Day, M.; Hsu, M. (2009). "The measurement of user satisfaction with question answering systems". *Information & Management*, 46(7): 397-403

Otero, P. G.; López, I. G. (2010). "Wikipedia as multilingual source of comparable corpora". En *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, LREC, 21-25.

P

Peinado Herencia, V. (2004). *Uso de los sintagmas alineados como recurso de traducción en un enfoque híbrido para la recuperación de información multilingüe*. Memoria de trabajo del Periodo de Investigación. Madrid: Universidad Nacional de Educación a Distancia.

Peñas, A.; Magnini, B.; Forner, P.; Sutcliffe, R.; Rodrigo, A.; Giampiccolo, D. (2012). "Question answering at the cross-language evaluation forum 2003---2010". *Language Resource Evaluation*, 46, 2 (June 2012), 177—217

Pérez-Coutiño, M.; Solorio, T.; Montes-y-Gómez, M.; López-López, A.; Villaseñor-Pineda, L. (2004). "Toward a Document Model for Question Answering Systems". *Proceedings of the Second International Atlantic Web Intelligence Conference*.

Pérez-Coutiño, M.; Solorio, T.; Montes y Gómez, M.; López López, A.; Villaseñor Pineda, L. (2004). "The Use of Lexical Context in Question Answering for Spanish". *Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*.

Pérez Hernández, C. (2002). "Explotación de los corpóra textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento". *Estudios de Lingüística del Español (EliEs)*, 18

Peters, C. (2009). "What Happened in CLEF 2009: Introduction to the Working Notes". En Borri, F., Nardi, A., and Peters, C. (eds.), *Working Notes for the CLEF 2009 Workshop*.

Petrecca, F. (2001). "La representación del léxico científico-técnico en los diccionarios". *Panace@*, 2(4), 2-4.

Phillips, A. (1960). "A question-answering routine". Memo. 16, *Artificial Intelligence Project*.

Pomerantz, J. (2005). "A linguistic analysis of question taxonomies". *Journal of the American Society for Information Science and Technology*, 56 (7): 715-728.

Potthast, M.; Stein, B.; Anderka, M. (2008). "A Wikipedia-based multilingual retrieval model". En C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven y R.W. White (Eds.), *30th European conference on IR research, ECIR 2008*, Glasgow, volume 4956 LNCS of Lecture Notes in Computer Science (pp. 522–530). Berlin: Springer.

Pustejovsky, J.; Gaizauskas, R.; Saurí, R.; Setzer, A.; Ingria, R. (2002). *Annotation Guideline to TimeML 1.0*, disponible en <http://time2002.org>.

R

Ramamonjisoa, D.R. (2003). "Finding Relevant Answers in Question Answering System Contest". *Proceedings of NTCIR — 4*.

Raved, D.R.; Qi, H.; Wu, H.; Fan, W. (2001). *Evaluating Web-based Question Answering Systems*. Technical Report, University of Michigan.

Recorder, M.J.; Cid, P. (2004). "La documentación en la Traducción especializada". En C. Gonzalo García y V. García Yedra (eds.), *Manual de Documentación y Terminología para la Traducción Especializada*. Madrid: Arco/Libro, S.L., PP. 73-88

Redondo Madrigal, M. (1995). "El uso del diccionario en la comprensión lectora de Lenguas Extranjeras". *LFE, Revista de lenguas para fines específicos*, 02

Rijsbergen, C.J. (1979). *Information Retrieval*. 2nd ed. Londres: Butterworths.

Rijsbergen, C.J. (1999). *Information Retrieval*. Glasgow: University. Disponible en: <http://www.dcs.gla.ac.uk/~iain/keith/>

Ribeiro-Neto, B.; Muntz, R.R. (1996). "A Belief Network Model for IR". *SIGIR 1996*.

Rodrigo, A.; Pérez-Iglesias, J.; Peñas, A.; Garrido, G.; Araujo, L. (2010). "A Question Answering System based on Information Retrieval and Validation", *Notebook Papers/LABs/Workshops (CLEF 2010)*.

Rodrigo Yuste, Á. (2010). *Evaluación de sistemas de búsqueda y validación de respuestas*. Tesis doctoral.

Roger, S.; Ferrández, A.; Peral, J.; Ferrández, S.; López-Moreno, P. (2007). "An inference mechanism for question answering". *Journal of computer science & technology*, 1, 21-27.

S

Sager, J.C. (1993). *Curso práctico sobre el procesamiento de la terminología*. Madrid: Fundación Germán Sánchez Ruipérez.

Sager, J.C. (2001). "Terminology compilation: Consequences and Aspects of Automation". En Wright y Budin (eds.), *Handbook of Terminology Management 2*, pp. 761-771.

Salton, G. (1970). "Automatic Processing of Foreign Language Documents". *Journal of American Society for Information Sciences*, 21:187-194.

Salton, G. (1971). *The SMART Retrieval System Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G. (1989). *Automatic Text Processing – The Transformation, Analysis and Retrieval of Information by Computer*. Addison Wesley Publishing Company.

Salton, G.; Mc Gill, M.J. (1983). *Introduction to Modern Information Retrieval*. Nueva York: McGraw-Hill Computer Series

Sánchez Ramos, M.M. (2007). *El uso de los diccionarios electrónicos y otros recursos de Internet como herramientas para la formación del traductor (inglés-español)*. Granada: Editorial Universidad de Granada

Saracevic, T. (1995). "Evaluation of evaluation in information retrieval". *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Special issue of SIGIR Forum, 138-146

Simmons, R.F. (1965). "Answering English Questions by Computer: A Survey". *Communications of the ACM*, 8(1):53–70.

Simmons, R.F., McConlogue, K.L. (1963). "Maximum-depth indexing for computer retrieval of english language data". *American Documentation*, 14(1), 68-73.

Sinclair, J. (1996). *EAGLES. Preliminary recommendations on Corpus Typology*. [En línea]. En <http://www.ilc.cnr.it/EAGLES/corpusstyp/corpusstyp.html>

Sing G.O., C. Ardil, W. Wong & Sahib, S. (2005). "Response Quality Evaluation in Heterogeneous Question Answering System: A Black-box Approach". *Proceedings of World Academy of Science, Engineering and Technology*, 9 (Lisbon). (2005)

Soergel, D. (1997). "Multilingual thesauri on cross-language text and speech retrieval". En D. Hull y D. Oard (eds.), *AAAI Symposium on Cross-Language Text and Speech Retrieval*.

Solorio, T.; Pérez-Coutiño, M.; Montes-Gómez, M.; Villaseñor-Pineda, L.; López-López, A. (2005). "Question classification in Spanish and Portuguese". *Conference on intelligent text processing and computational linguistics, CICLing 2005*. Lecture notes in computer science, 3406, 612-619.

Sorg, P.; Cimiano, P. (2008). "Cross-lingual Information Retrieval with Explicit Semantic Analysis" *Working Notes for the CLEF 2008 Workshop*.

Sperberg-McQueen, C.M.; Burnard, L. (2002). *Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines*. The TEI Consortium

Sultan, M. (2006). *Multiple Choice Question Answering*. Tesis doctoral. University of Sheffield.

T

Taylor, J.R. (1999). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, 128–129.

Terol, R.M., Martínez-Barco, P., Palomar, M. (2007). "A knowledge based method for the medical question answering problem". *Computers in biology and medicine*, 37(10), 1511-1521.

Troyano Jiménez, L. (2007). "Extracción de terminología médica especializada de diccionarios generales bilingües (inglés-español): metodología de trabajo". *Traducción y mediación cultural: reflexiones interdisciplinares*, 215-224. Atrio.

Tsur, O. (2003). *Definitional Question-Answering Using Trainable Text Classifiers*. Tesis doctoral. University of Amsterdam.

Turtle, H.; Croft, W.B. (1991). "Evaluation of an inference network-based retrieval model". *ACM Transactions on Information Systems (TOIS)*, 9(3), 187-222.

Tyers, F.M.; Pienaar, J. A. (2008). "Extracting bilingual word pairs from Wikipedia". *Collaboration: interoperability between people in the creation of language resources for less-resourced languages*, 19.

V

Vallez, M.; Pedraza-Jimenez, R. (2007). "El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines". *Hipertext.net* [en 250reci], 5. <http://www.hipertext.net>

Verona, M.V.; Motta, E. (2004). *AQUA, A Knowledge-Based Architecture for a Question Answering System*, Tech Report Kmi-o4-15, Knowledge media institute, Milton Keynes, England.

Vicedo, J.L.; Rodríguez, H.; Peñas, A.; Massot, M. (2003). "Los sistemas de Búsqueda de Respuestas desde una perspectiva actual". *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 31.

Villena Román, J. (1997). *Sistemas de Recuperación de Información*. [En línea]. Valladolid: Departamento Ingeniería Sistemas Telemáticos, Universidad. <<http://www.mat.upm.es/~jmg/doct00/RecupInfo.pdf>>

Voorhees, E.M. (1999). "The TREC 8 Question Answering Track Report". En Voorhees y Harman (eds.), *Proceedings of the 8th Text Retrieval Conference*, 107-130.

Voorhees, E.M. (2002). "Overview of the TREC 2002 Question Answering Track". *Proceedings of the Eleventh Text Retrieval Conference*.

Voorhees, E. (2003). "Overview of TREC 2003". En: *Proceedings of the 12th Text Retrieval Conference*

Voorhees, E.M.; Tice, D. (1999). *The TREC-8 question answering track evaluation*. En Greece: Athens

Vossen, P. (1998). "Introduction to EuroWordNet". *Computer and the Humanities, Special Issues on EuroWordNet* 32, 73-89.

W

Warren, D. (1981). "Efficient Processing of Interactive Relational Database Queries Expressed in Logic". *Proceedings Seventh International Conference on Very Large Data Bases*. Cannes: VLDB Endowment. V.7, pp. 272-283.

Wedgwood, J.A. (2005). "Medical question-answering framework". *AMIA 2005 Symposium proceedings*, 1.150.

Weizenbaum, J. (1966). "Eliza: A computer program for the study of natural language communication between man and machine". *Communications of the ACM*, 9 (1), 36-45.

Woods, W. (1973). "Progress in Natural Language Understanding – An Application to Lunar Geology". En *AFIPS Conference Proceedings*, 42, 441–450.

Woods, W.A.; Kaplan, R.M.; Nash-Webber, B. (1972). "The Lunar Sciences Natural Language Information System". En: *BBN Final Report 2378*. Cambridge: Bolt, Beranek and Newman.

Wright, S.E.; Wright, L.D. (1997). "Terminology Management for Technical Translation". En Wright y Budin (eds.), *Handbook of Terminology Management* 1, pp. 147-159.

Y

Yu, H.; Lee, M.; Kaufman, D.; Ely, J.; Osheroff, J.A.; Hripcsak, G.; Cimino, J. (2007). "Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians". *Journal of Biomedicine Informatics*, 4, 236-251.

Yu, H.; Kaufman, D. (2007). "A cognitive evaluation of four online search engines for answering definitional questions posed by physicians". *Pacific Symposium on Biocomputing*, 12, 328-339

Yu, K.; Tsujii, T. (2009). "Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity". *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 121-124

Z

Zhang, Y., Jones, G.J., Zhang, S., Wang, B., Guo, Y., Ma, Y. (2007). "An investigation of question translation for English-Chinese cross-language question answering." En: *CIICT 2007 - Proceedings of the China-Ireland International Conference on Information and Communications Technologies*, 28-29 Agosto 2007, Dublin, Irlanda.

Zhou, D.; Turan, M.; Brailsford, T. (2007). "Ambiguity and unknown term translation in CLIR", *Working Notes for the CLEF 2007 Workshop*.

Zhou, D.; Truran, M.; Brailsford, T.; Wade, V.; Ashman, H. (2012). "Translation techniques in cross-language information retrieval". *ACM Comput. Surv.* 45, 1, Article 1 (December 2012), 44 páginas.

Zweigenbaum, P. (2005). "Question answering in biomedicine". De Rijke and Webber (eds.), *Proceedings Workshop on Natural Language Processing for Question Answering*, EACL 2003, 1-4. ACL, Budapest.

Zweigenbaum, P. (2009). "Knowledge and reasoning for medical question-answering". *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions*, pp. 1-2

ANEXOS

ANEXO 1. COLECCIÓN DE PREGUNTAS DE DEFINICIÓN BIOMÉDICAS

1. What is abortion?
2. What Is Acne?
3. What is acupuncture?
4. What Is ADHD?
5. What Is Allergic Asthma?
6. What is alopecia areata?
7. What is altitude sickness?
8. What Is Amebiasis?
9. What is anesthesia?
10. What is ankle sprain?
11. What Is Anthrax?
12. What is aortic aneurysm?
13. What is aortic valve regurgitation?
14. What is aromatherapy?
15. What Is Asthma?
16. What Is Atherosclerosis?
17. What is ayurveda?
18. What Is Back Pain?
19. What Is Bipolar Disorder?
20. What is borderline personality disorder?
21. What is bradycardia?
22. What is campylobacteriosis?
23. What is capsaicin?
24. What is carbon monoxide poisoning?
25. What is carpal tunnel syndrome?
26. What is cataract?
27. What is celiac disease?
28. What is central venous catheter?
29. What is chelation therapy?
30. What is Chemotherapy?
31. What is chickenpox?
32. What is Chinese medicine?
33. What is chiropractic?
34. What is chlamydia?
35. What Is Chronic Pain?

36. What is circumcision?
37. What is Clostridium difficile colitis?
38. What is cochlear implant?
39. What Is Colorectal Cancer?
40. What is COPD?
41. What is Crohn's disease?
42. What Is Cystic Fibrosis?
43. What is deep vein thrombosis?
44. What is degenerative disc disease?
45. What is depression?
46. What is diabetic ketoacidosis?
47. What is diabetic retinopathy?
48. What is disseminated intravascular coagulation?
49. What is diverticulosis?
50. What is Dry Mouth?
51. What is dyslexia?
52. What is endometriosis?
53. What is ephedra?
54. What is epididymitis?
55. What is epilepsy?
56. What is esophageal spasm?
57. What Is Fibromyalgia?
58. What is flatfoot?
59. What is Flu?
60. What is FluMist?
61. What Is Gastric Bypass Surgery?
62. What is gastroenteritis?
63. What is gastroesophageal reflux?
64. What is gastroparesis?
65. What is genital herpes?
66. What is gestational diabetes?
67. What is Gingivitis?
68. What is gotrrhea?
69. What Is Guillain-Barre Syndrome?
70. What is gynecomastia?
71. What is healing touch?
72. What is heart attack?
73. What is heart failure?
74. What is hemophilia?
75. What is hepatitis C?
76. What is herniated disc?
77. What is high blood pressure?
78. What is homeopathy?
79. What is Huntington's disease?
80. What is hydrocele?
81. What is hydrotherapy?
82. What is hypospadias?
83. What is hypothermia?
84. What is hypothyroidism?

85. What is Inflammatory Bowel Disease?
86. What Is Insomnia?
87. What Is Irritable Bowel Syndrome?
88. What is kava?
89. What is kernicterus?
90. What Is Lactose Intolerance?
91. What is laryngitis?
92. What is latex allergy?
93. What is laxative?
94. What is lipoma?
95. What is listeriosis?
96. What Is Lung Cancer?
97. What is melasma?
98. What is Metabolic Syndrome?
99. What Is Multiple Sclerosis?
100. What is naturopathic medicine?
101. What is oophorectomy?
102. What is Orthodontics?
103. What is Osteoporosis?
104. What is overactive bladder?
105. What is patellofemoral pain syndrome?
106. What Is Perimenopause?
107. What is pheochromocytoma?
108. What is pityriasis rosea?
109. What is pleurisy?
110. What is pneumothorax?
111. What is prediabetes?
112. What is prematurity?
113. What is presbyopia?
114. What is primary biliary cirrhosis?
115. What is prostate cancer?
116. What is radiation therapy?
117. What is retinal detachment?
118. What Is Rheumatoid Arthritis?
119. What is Root Canal Treatment?
120. What is rosacea?
121. What is roseola?
122. What is rotavirus?
123. What is salmonellosis?
124. What is SARS?
125. What is saw palmetto?
126. What is scabies?
127. What is scarlet fever?
128. What Is Schizophrenia?
129. What is sciatica?
130. What is scoliosis?
131. What is shigellosis?
132. What is sickle cell disease?
133. What is sinusitis?
134. What is sleep apnea?

135. What is smallpox?
136. What is spermatocoele?
137. What is spondylolisthesis?
138. What is stem cell transplant?
139. What is stillbirth?
140. What is Tartar?
141. What is TENS?
142. What is tension headache?
143. What is testicular cancer?
144. What is TMJ?
145. What is tonsillitis?
146. What is Tourette Syndrome?
147. What is transient ischemic attack ?
148. What is trichomoniasis?
149. What is tuberculosis?
150. What is vaginitis?
151. What is ventricular tachycardia?
152. What is vesicoureteral reflux?
153. What is virtual colonoscopy?
154. What is whiplash?
155. What is an ingrown toenail?
156. What is binge eating disorder?

ANEXO 2. Colección de preguntas de TREC y CLEF

	Pregunta en inglés
1.	How deep is Crater Lake?
2.	How did Bob Marley die?
3.	How did Janice Joplin die?
4.	How did Jimi Hendrix die?
5.	How did Mahatma Gandhi die?
6.	How did Pasolini die?
7.	How did River Phoenix die?
8.	How did Vitas Gerulaitis die?
9.	How far away from the sun is Saturn?
10.	How far away is the moon?
11.	How far is Pluto from the sun?
12.	How far is Yaroslavl from Moscow?
13.	How fast is sound?
14.	How fast is the speed of light?
15.	How high is the K2?
16.	How is ozone formed?
17.	How long is the Okavango River?
18.	How many counties are in Indiana?
19.	How many feet in a mile?
20.	How many inhabitants does Vietnam have?
21.	How many liters in a gallon?
22.	How many moons does Jupiter have?
23.	How many ounces are in a gallon?
24.	How many people live in Bombay?
25.	How many people live in Chile?
26.	How many people live in the Falklands?
27.	How many people live in Tokyo?
28.	How many pounds in a ton?
29.	How old is Deng Xiaoping?
30.	How old was Burt Lancaster when he died?
31.	How old was Elvis Presley when he died?
32.	How old was Howard Caine when he died?
33.	How old was Joan of Arc when she died?

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

34.	How tall is Allen Iverson?
35.	How tall is Kilimanjaro?
36.	How tall is Mt. Everest?
37.	How tall is the Matterhorn?
38.	In what country is Essen located?
39.	In which country is Mount Kilimanjaro located?
40.	Izmir is located in what country?
41.	What actress starred in "The Lion in Winter"?
42.	What American composer wrote the music for "West Side Story"?
43.	What are Aborigines?
44.	What are amphibians?
45.	What are antacids?
46.	What are breast implants made of?
47.	What are chloroplasts?
48.	What are cigarettes made of?
49.	What are coral reefs?
50.	What are enzymes?
51.	What are ethics?
52.	What are geckos?
53.	What are invertebrates?
54.	What are pathogens?
55.	What are platelets?
56.	What are polymers?
57.	What are pomegranates?
58.	What are Quaaludes?
59.	What are semiconductors?
60.	What are solar cells?
61.	What are spider veins?
62.	What are sunspots?
63.	What are the Poconos?
64.	What are the twin cities?
65.	What are triglycerides?
66.	What are xerophytes?
67.	What continent is Argentina on?
68.	What continent is Bolivia on?
69.	What continent is Egypt on?
70.	What continent is India on?
71.	What continent is Scotland in?
72.	What country is Berlin in?
73.	What currency do they use in Brazil?
74.	What currency does Argentina use?

75.	What currency does Luxembourg use?
76.	What currency is used in Algeria?
77.	What day and month did John Lennon die?
78.	What does a luthier make?
79.	What does ciao mean?
80.	What does CNN stand for?
81.	What does CPR stand for?
82.	What does DNA stand for?
83.	What does EKG stand for?
84.	What does HTML stand for?
85.	What does I.V. stand for?
86.	What does laser stand for?
87.	What does NAFTA stand for?
88.	What does NASA stand for?
89.	What does NASDAQ stand for?
90.	What does NEO stand for?
91.	What does SIDS stand for?
92.	What does the abbreviation OAS stand for?
93.	What does the abbreviation SOS mean?
94.	What does the acronym NASA stand for?
95.	What does the word fortnight mean?
96.	What does USPS stand for?
97.	What is a biosphere?
98.	What is a carcinogen?
99.	What is a fuel cell?
100.	What is a meerkat?
101.	What is a micron?
102.	What is a mirror made out of?
103.	What is a nanometer?
104.	What is a nematode?
105.	What is a panic disorder?
106.	What is a parasite?
107.	What is a prism?
108.	What is a stratocaster?
109.	What is a thyroid?
110.	What is a tsunami?
111.	What is acetaminophen?
112.	What is acetic acid?
113.	What is acid rain?
114.	What is acupuncture?
115.	What is Alice Cooper's real name?
116.	What is amitriptyline?

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

117.	What is amoxicillin?
118.	What is an annuity?
119.	What is an atom?
120.	What is an earthquake?
121.	What is an eclipse?
122.	What is an ulcer?
123.	What is angiotensin?
124.	What is anise?
125.	What is anorexia nervosa?
126.	What is another name for nearsightedness?
127.	What is another name for vitamin B1?
128.	What is Archimedes famous for?
129.	What is Australia's national flower?
130.	What is autism?
131.	What is bandwidth?
132.	What is bangers and mash?
133.	What is Betsy Ross famous for?
134.	What is bio-diversity?
135.	What is bipolar disorder?
136.	What is caffeine?
137.	What is California's capital?
138.	What is California's state bird?
139.	What is California's state tree?
140.	What is Canada's most populous city?
141.	What is carbon dioxide?
142.	What is cerebral palsy?
143.	What is cholesterol?
144.	What is cribbage?
145.	What is cryogenics?
146.	What is cryptography?
147.	What is diabetes?
148.	What is dianetics?
149.	What is Dick Clark's birthday?
150.	What is die-casting?
151.	What is Dr. Ruth's last name?
152.	What is e-coli?
153.	What is epilepsy?
154.	What is fibromyalgia?
155.	What is foreclosure?
156.	What is fungus?
157.	What is genocide?

158.	What is Hawaii's state flower?
159.	What is hybridization?
160.	What is hypertension?
161.	What is influenza?
162.	What is Java?
163.	What is leukemia?
164.	What is mad cow disease?
165.	What is Mardi Gras?
166.	What is Maryland's state bird?
167.	What is metabolism?
168.	What is Mexico's population?
169.	What is mold?
170.	What is molybdenum?
171.	What is MSNBC?
172.	What is Muscular Dystrophy?
173.	What is myopia?
174.	What is nanotechnology?
175.	What is naproxen?
176.	What is natural gas composed of?
177.	What is nepotism?
178.	What is neurology?
179.	What is neuropathy?
180.	What is New York's state bird?
181.	What is nicotine?
182.	What is nuclear power?
183.	What is osteoporosis?
184.	What is ouzo?
185.	What is ozone depletion?
186.	What is pandoro?
187.	What is pastrami made of?
188.	What is pectin?
189.	What is peyote?
190.	What is phenylalanine?
191.	What is phosphorus?
192.	What is pilates?
193.	What is plastic made of?
194.	What is platinum?
195.	What is pneumonia?
196.	What is poliomyelitis?
197.	What is porphyria?
198.	What is propylene glycol?
199.	What is pulmonary fibrosis?

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

200.	What is quicksilver?
201.	What is relative humidity?
202.	What is sake?
203.	What is saltpeter?
204.	What is schizophrenia?
205.	What is semolina?
206.	What is severance pay?
207.	What is sodium chloride?
208.	What is solar wind?
209.	What is sonar?
210.	What is strep throat?
211.	What is supernova?
212.	What is Susan B. Anthony's birthday?
213.	What is TCI?
214.	What is Teflon?
215.	What is thalassemia?
216.	What is the abbreviation for Texas?
217.	What is the age of our solar system?
218.	What is the area of Venezuela?
219.	What is the atomic weight of silver?
220.	What is the Boston Strangler's name?
221.	What is the brightest star?
222.	What is the capital city of Algeria?
223.	What is the capital city of New Zealand?
224.	What is the capital of Alaska?
225.	What is the capital of Burkina Faso?
226.	What is the capital of California?
227.	What is the capital of Chechnya?
228.	What is the capital of Congo?
229.	What is the capital of Ethiopia?
230.	What is the capital of Italy?
231.	What is the capital of Kosovo?
232.	What is the capital of Madagascar?
233.	What is the capital of Mongolia?
234.	What is the capital of Zimbabwe?
235.	What is the chemical symbol for nitrogen?
236.	What is the chunnel?
237.	What is the coldest place on earth?
238.	What is the currency of Bolivia called?
239.	What is the currency used in China?
240.	What is the current population in Bombay, India?

241.	What is the distance from Jupiter to the sun?
242.	What is the distance in miles from the earth to the sun?
243.	What is the earth's diameter?
244.	What is the European Parliament?
245.	What is the fourth highest mountain in the world?
246.	What is the gestation period for a cat?
247.	What is the highest mountain in Germany?
248.	What is the highest mountain in the world?
249.	What is the Iraqi currency?
250.	What is the largest city in Germany?
251.	What is the largest city in the U.S.?
252.	What is the largest city in the world?
253.	What is the literacy rate in Cuba?
254.	What is the location of Lake Champlain?
255.	What is the longest English word?
256.	What is the longest river in Norway?
257.	What is the longest river in the United States?
258.	What is the lowest point on earth?
259.	What is the meaning of thalassemia?
260.	What is the melting point of copper?
261.	What is the melting point of gold?
262.	What is the Milky Way?
263.	What is the most populated country in the world?
264.	What is the Moulin Rouge?
265.	What is the name of a Greek god?
266.	What is the name of Bill Clinton's wife?
267.	What is the name of the airport in Amsterdam?
268.	What is the name of the highest mountain in Africa?
269.	What is the name of the leader of Ireland?
270.	What is the North Korean national anthem?
271.	What is the Ohio state bird?
272.	What is the oldest city in the United States?
273.	What is the pH scale?
274.	What is the population of Australia?
275.	What is the population of China?
276.	What is the population of Japan?
277.	What is the population of Kansas?
278.	What is the population of Maryland?
279.	What is the population of Mexico?
280.	What is the population of Nigeria?
281.	What is the population of Ohio?
282.	What is the population of Seattle?

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

283.	What is the population of the Bahamas?
284.	What is the population of the United States?
285.	What is the population of Venezuela?
286.	What is the rainiest place on Earth?
287.	What is the real name of "Dr. Death"?
288.	What is the scientific name for elephant?
289.	What is the second highest mountain peak in the world?
290.	What is the size of Argentina?
291.	What is the speed of light?
292.	What is the state flower of Michigan?
293.	What is the statue of liberty made of?
294.	What is the wingspan of a condor?
295.	What is the world's highest peak?
296.	What is the world's population?
297.	What is the WTO?
298.	What is titanium?
299.	What is TMJ?
300.	What is tyvek?
301.	What is UEFA?
302.	What is UNICEF?
303.	What is Ursa Major?
304.	What is Valentine's Day?
305.	What is vertigo?
306.	What is viscosity?
307.	What is Wimbledon?
308.	What is witch hazel?
309.	What made Jane Goodall famous?
310.	What province is Edmonton located in?
311.	What province is Montreal in?
312.	What state does MO stand for?
313.	What was the birthplace of Edgar Allen Poe?
314.	What was the purpose of the Manhattan project?
315.	What was W.C. Fields' real name?
316.	What year did Alaska become a state?
317.	What year did Canada join the United Nations?
318.	What year did Hitler die?
319.	What year did Montana become a state?
320.	What year did South Dakota become a state?
321.	What year did the Vietnam War end?
322.	What year was the first automobile manufactured?
323.	What year was the PC invented?

324.	What's the farthest planet from the sun?
325.	What's the longest river in the world?
326.	What's the name of the Tokyo Stock Exchange?
327.	What's the population of Biloxi, Mississippi?
328.	What's the population of Mississippi?
329.	When did Aldous Huxley write, "Brave New World"?
330.	When did Alexander Graham Bell invent the telephone?
331.	When did beethoven die?
332.	When did Bob Marley die?
333.	When did Elvis Presley die?
334.	When did Hawaii become a state?
335.	When did Idaho become a state?
336.	When did Jaco Pastorius die?
337.	When did Lenin die?
338.	When did Muhammad live?
339.	When did Nixon die?
340.	When did North Carolina enter the union?
341.	When did Princess Diana and Prince Charles get married?
342.	When did the American Civil War end?
343.	When did the Bay of Pigs invasion take place?
344.	When did World War I start?
345.	When is Bastille Day?
346.	When is Boxing Day?
347.	When is Father's Day?
348.	When is Mexico's independence?
349.	When is Snoop Dog's birthday?
350.	When is St. Patrick's Day?
351.	When is the summer solstice?
352.	When was Abraham Lincoln born?
353.	When was Babe Ruth born?
354.	When was Beethoven born?
355.	When was Davy Crockett born?
356.	When was Dick Clark born?
357.	When was John D. Rockefeller born?
358.	When was Lyndon B. Johnson born?
359.	When was Nostradamus born?
360.	When was Queen Victoria born?
361.	When was Rosa Parks born?
362.	When was the Declaration of Independence signed?
363.	When was the internal combustion engine invented?
364.	When was the NFL established?
365.	When was the telegraph invented?

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

366.	When was Thomas Jefferson born?
367.	When was Ulysses S. Grant born?
368.	Where are the National Archives?
369.	Where are the Rocky Mountains?
370.	Where did Dylan Thomas die?
371.	Where did Howard Hughes die?
372.	Where did the Battle of the Bulge take place?
373.	Where did the ukulele originate?
374.	Where did yoga originate?
375.	Where does chocolate come from?
376.	Where does dew come from?
377.	Where is Dartmouth College?
378.	Where is Basque country located?
379.	Where is Big Ben?
380.	Where is Burma?
381.	Where is Devil's Tower?
382.	Where is Georgetown University?
383.	Where is Glasgow?
384.	Where is Guam?
385.	Where is Halifax located?
386.	Where is Hebron located?
387.	Where is Hitler buried?
388.	Where is John Wayne airport?
389.	Where is Kings Canyon?
390.	Where is Logan Airport?
391.	Where is Los Vegas?
392.	Where is Melbourne?
393.	Where is Milan?
394.	Where is Ocho Rios?
395.	Where is Perth?
396.	Where is Prince Edward Island?
397.	Where is Qatar?
398.	Where is Requena located?
399.	Where is Rider College located?
400.	Where is Romania located?
401.	Where is Santa Lucia?
402.	Where is South Bend?
403.	Where is Svalbard archipelago?
404.	Where is the Danube?
405.	Where is the Eiffel Tower?
406.	Where is the Euphrates River?

407.	Where is the Grand Canyon?
408.	Where is the highest point on earth?
409.	Where is the Holland Tunnel?
410.	Where is the Isle of Man?
411.	Where is the Kalahari desert?
412.	Where is the Louvre?
413.	Where is the Mayo Clinic?
414.	Where is the Reichstag?
415.	Where is the Savannah River?
416.	Where is the Shawnee National Forest?
417.	Where is the Smithsonian Institute located?
418.	Where is the Statue of Liberty?
419.	Where is the Taj Mahal?
420.	Where is the Valley of the Kings?
421.	Where is Tornado Alley?
422.	Where is Trinidad?
423.	Where is Tufts University?
424.	Where is Venezuela?
425.	Where is Webster University?
426.	Where is Windsor Castle?
427.	Where was George Washington born?
428.	Where was Harry Truman born?
429.	Where was John Adams born?
430.	Where was Pythagoras born?
431.	Where was Ulysses S. Grant born?
432.	Where's Montenegro?
433.	Who directed "Nikita"?
434.	Who discovered America?
435.	Who discovered oxygen?
436.	Who discovered radium?
437.	Who invented paper?
438.	Who invented television?
439.	Who invented the calculator?
440.	Who invented the radio?
441.	Who invented the saxophone?
442.	Who invented the telephone?
443.	Who is Anubis?
444.	Who is Arabella Kiesbauer?
445.	Who is Barbara Jordan?
446.	Who is Bill Clinton?
447.	Who is Charles Lindbergh?
448.	Who is Colin Powell?

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

449.	Who is Desmond Tutu?
450.	Who is Duke Ellington?
451.	Who is Henry Butler?
452.	Who is John Paul II?
453.	Who is Langston Hughes?
454.	Who is Paul Simon?
455.	Who is Peter Weir?
456.	Who is Terrence Malick?
457.	Who is the author of "Jonathan Livingston Seagull"?
458.	Who is the leader of India?
459.	Who is the monarch of the United Kingdom?
460.	Who is the owner of CNN?
461.	Who is the president of Bolivia?
462.	Who is the president of South Korea?
463.	Who is the prime minister of Australia?
464.	Who is the Prime Minister of Canada?
465.	Who is the Prime Minister of Ireland?
466.	Who is Tom Cruise married to?
467.	Who is William Wordsworth?
468.	Who is Yves Saint Laurent?
469.	Who is Zebulon Pike?
470.	Who killed John F. Kennedy?
471.	Who killed Martin Luther King?
472.	Who painted Olympia?
473.	Who starred in "The Poseidon Adventure"?
474.	Who was Abraham Lincoln?
475.	Who was Buffalo Bill?
476.	Who was Charles Lindbergh's wife?
477.	Who was Galileo?
478.	Who was Jane Goodall?
479.	Who was Lacan?
480.	Who was Maria Theresa?
481.	Who was Monet?
482.	Who was Picasso?
483.	Who was president in 1913?
484.	Who was Quetzalcoatl?
485.	Who was Rosa Chacel?
486.	Who was the 21st U.S. President?
487.	Who was the 22nd President of the US?
488.	Who was the 23rd president of the United States?
489.	Who was the 33rd president of the United States?

490.	Who was the first woman in space?
491.	Who was the president of the United States in 1994?
492.	Who was the tallest U.S. president?
493.	Who was the third U.S. President?
494.	Who was Whitcomb Judson?
495.	Who won the Nobel Peace Prize in 1992?
496.	Who won the Nobel Prize for Literature in 1994?
497.	Who won the nobel prize in literature in 1988?
498.	Who won the Oscar for best actor in 1970?
499.	Who wrote "An Ideal Husband"?
500.	Who wrote the book, "Huckleberry Finn"?

ANEXO 3. CUESTIONARIO ARTÍCULO 8

UNIVERSIDAD DE GRANADA

FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

Edad: _____ Sexo: _____ Curso: _____



	Bajo	Medio	Alto
Grado de experiencia con herramientas de recuperación de información en la Web			
	Bajo	Medio	Alto
Grado de conocimientos médicos			

Responde a las siguientes preguntas marcando la casilla correspondiente a la puntuación que consideres oportuna:

0 = Totalmente en desacuerdo, 1= Muy en desacuerdo, 2 = En desacuerdo en ciertos aspectos, 3 = Indeciso, 4 = De acuerdo en ciertos aspectos, 5 = Muy de acuerdo, 6= Totalmente de acuerdo

	0	1	2	3	4	5	6
La <u>interacción</u> con el sistema de búsqueda de respuestas es clara y comprensible							
Resulta sencillo <u>aprender</u> a utilizar el sistema							
Es fácil llegar a estar <u>completamente familiarizado</u> con el uso del sistema							
Encuentro fácil utilizar el sistema para que realice <u>lo que yo quiero</u> que haga							
Encuentro fácil <u>utilizar</u> el sistema							
Utilizar el sistema podría mejorar mi <u>eficacia</u> en el desarrollo de las tareas académicas							
El sistema podría ser <u>útil</u> en el desarrollo de mis tareas							
El uso del sistema podría mejorar el <u>rendimiento</u> en mi trabajo							
El uso del sistema en mis tareas podría incrementar mi <u>productividad</u>							
El uso del sistema podría hacer más <u>fácil</u> mi trabajo							
Este sistema de búsqueda de respuestas es <u>fiable</u>							
La <u>ayuda</u> ofrecida por el sistema es suficiente							

La tecnología utilizada por el sistema y la información que ofrece están <u>actualizados</u>							
La información facilitada por el sistema es fácil de <u>comprender</u>							
La información facilitada por el sistema es <u>relevante</u>							
La información facilitada por el sistema es <u>completa</u>							
La información facilitada por este sistema está <u>personalizada</u>							

Tabla cuestionario.

GLOSARIO

ALFA DE CRONBACH	Coeficiente que sirve para medir la fiabilidad de una escala de medida, y cuya denominación Alfa fue realizada por Cronbach en 1951.
ALINEACIÓN	Se basa en un modelo estadístico basado en la longitud en caracteres de las oraciones en uno y otro corpus. Se basa en la idea de que las oraciones largas (nº caracteres) en una lengua tienden a ser traducidas a oraciones largas en la otra lengua y que las oraciones cortas (nº caracteres) tienden a ser traducidas a oraciones cortas.
AMBIGÜEDAD	Hablamos de ambigüedad en RI cuando existe una palabra, sintagma u oración que es susceptible de dos o más significados o interpretaciones, es decir, la palabra, sintagma u oración se puede entender de más de una manera.
AVEP	Average Precisión se centra en las primeras respuestas relevantes. Se trata del promedio de precisiones computadas en el punto de cada una de las respuestas relevantes en la secuencia ordenada. Un sistema con buen funcionamiento situará las respuestas relevantes en las primeras posiciones de la lista. La AveP pretende medir el orden de las respuestas relevantes en la lista de resultados recuperados.
BASE DE DATOS	Conjunto de datos en soporte automatizado pertenecientes a un mismo campo y almacenados sistemáticamente para su posterior uso.
BASES DE DATOS RELACIONALES	Obras diseñadas para generar un conjunto de esquemas de relaciones que permitan almacenar la información con un mínimo de redundancia, pero que a la vez faciliten la RI.
BASES DE DATOS TERMINOLÓGICAS	Obras que contienen referencias bibliográficas y acceso a las bases de datos bibliográficas, pero contienen además diversos contextos de definición, ejemplos de uso y también contextos de los

	equivalentes en lengua extranjera, así como otra información complementaria a la entrada terminológica.
BÚSQUEDA DE RESPUESTAS	Área de la disciplina de la RI que se centra en la recuperación de información concreta en un dominio general o específico.
CLIR	Sistemas de RI que son capaces de operar en una colección de documentos multilingües dada una consulta determinada, se componen de una colección multilingüe de documentos y recuperan información relevante sin importar que lengua sea usada.
CLEF	Este foro comenzó a celebrarse a partir del año 2000, cuando la evaluación multilingüe empieza a tener auge dentro del ámbito de la RI. Es el foro europeo más destacado y está financiado con fondos de la Unión Europea. Sus dos objetivos más importantes son promover el desarrollo de infraestructuras para la puesta en marcha y evaluación de los sistemas de RI que trabajan con lenguas europeas en contextos monolingües y multilingües; y crear prototipos con información que pueda ser útil y usada por los desarrolladores del mercado.
COBERTURA	Proporción de documentos relevantes conocidos previamente por el usuario y pertenecientes a la colección que se ha recuperado.
CORPUS	Recurso lingüístico formado por una colección de fragmentos de documentos escritos en una determinada lengua que ha sido seleccionada y ordenada siguiendo unos criterios lingüísticos específicos para ser usada como muestra de un idioma o un lenguaje de especialidad.
DICCIONARIO	Recopilación de palabras, locuciones, giros y sintagmas de una lengua o, dentro de ella, los términos de una ciencia, técnica, artes, especialidad, etcétera, generalmente dispuesto en orden alfabético.
DICCIONARIO AUTOMATIZADO	Dícese de los diccionarios electrónicos que permiten ser incorporados sencillamente a otros sistemas electrónicos.
EUROWORDNET	Recurso lingüístico en donde se establece una serie de relaciones entre un conjunto de sinónimos o <i>synset</i> (<i>synonym sets</i>) dentro de una misma lengua, y se relacionan entre las diferentes lenguas disponibles por medio de un índice interlingüístico.
EFICACIA	La eficacia (<i>effectiveness</i> en inglés) de un SRI alude a la

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

	capacidad de dicho sistema para recuperar información pertinente en relación con cierta consulta desechando al tiempo la inútil.
EXACTITUD	Medida de evaluación que relaciona el sesgo de una estimación. Cuanto menor es el sesgo, más exacta es una estimación. Cuando se expresa la exactitud de un resultado, se expresa mediante el error absoluto que es la diferencia entre el valor experimental y el valor verdadero.
EXHAUSTIVIDAD	Medida de evaluación tradicional, aunque menos utilizada. Es la proporción de material relevante recuperado del total de los documentos que son relevantes en la base de datos, independientemente de que éstos se recuperen o no.
FHS	En First Hit Success (FHS) se asigna valor 1 si la primera respuesta ofrecida es correcta, y valor 0 si no lo es (por lo que sólo considera la respuesta que aparece en primer lugar en la lista de resultados).
GRAMÁTICA COMPUTACIONAL	Descripción formatilizada del conocimiento lingüístico, es decir, son herramientas que contienen cualquier tipo de información lingüística y que se puede recuperar con ciertos formalismos.
HERRAMIENTA LINGÜÍSTICA	Conjunto de software o programas que permiten la manipulación o gestión de recursos lingüísticos.
HONQA	Se trata de un SBR que recupera información en inglés, francés e italiano, desarrollado por la <i>Health On the Net Foundation</i> . Las fuentes de información utilizadas son muy variadas, aunque, por lo general, se trata de portales médicos generales o especializados en una determinada enfermedad. Este sistema se caracteriza por tener una interfaz muy sencilla donde plantear la pregunta deseada, si bien ofrece la posibilidad de una búsqueda avanzada que permite especificar un idioma determinado, la fuente de información de donde se extraerán las respuestas así como el número máximo de respuestas que se desea obtener.
INDEXACIÓN CONCEPTUAL	Proceso que consiste en traducir los documentos a una representación conceptual independiente al idioma.
INTER-LINGUAL INDEX (IDI)	Índice interlingüístico que relaciona las diferentes lenguas de EuroWordNet.
INTERACTIVIDAD	Cualidad que permite un intercambio activo entre el ordenador o

	el software y la persona que lo está utilizando.
INTERFAZ	Zona de contacto o conexión entre dos componentes de hardware; entre dos aplicaciones, o entre un usuario y una aplicación. Apariencia externa de una aplicación informática.
LAB	Se denomina <i>lab</i> al conjunto de tareas específicas que determinan los organizadores de los principales congresos en RI para el testeo y evaluación de los sistemas desarrollados. Depende del foro también se denomina <i>task</i> .
LENGUAJE SQL	El lenguaje de consulta estructurado o SQL (por sus siglas en inglés Structured Query Language) es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en ellas. Una de sus características es el manejo del álgebra y el cálculo relacional que permiten efectuar consultas con el fin de recuperar de forma sencilla información bases de datos, así como hacer cambios en ellas.
MAP	Mean Average Precision mide el promedio medio de precisión de un conjunto de preguntas en donde las respuestas son ordenadas según relevancia.
MEDQA	SBR que actualmente no se encuentra disponible pero que tuvo una amplia repercusión entre usuarios e investigadores. Contaba con una interfaz muy sencilla que resultaba muy amigable, pero a la vez presentaba una sofisticada arquitectura que le permitía elaborar y generar respuestas adecuadas a las preguntas planteadas. Aunque uno de sus principales problemas era que, a diferencia de otros SBR, resultaba algo lento al proporcionar las respuestas, utilizaba una amplia variedad de fuentes de información, entre las que se encontraban <i>Wikipedia</i> , <i>Medline</i> o <i>Medline Plus</i> .
MRR	Mean Reciprocal Rank (MRR) asigna el valor inverso de la posición en la que la respuesta correcta fue encontrada (1 si es la primera, $\frac{1}{2}$ si es la segunda, $\frac{1}{3}$ si es la tercera, y así sucesivamente), o cero si la respuesta correcta no fue encontrada. Según esta medida, solamente hay una respuesta correcta dentro de la lista de resultados ofrecidos por el sistema, y el valor final es el promedio de los valores obtenidos para cada pregunta. MRR asigna un valor alto a las respuestas que están en las posiciones más altas de la clasificación.
NTCIR	NTCIR (NII Test Collection for IR Systems) es un foro centrado en temas de RI con gran repercusión y centrado principalmente

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

	en idiomas asiáticos.
ONTOLOGÍA	Especificación formal y explícita de conceptualización compartida.
QUALIM	SBR financiado por Microsoft y desarrollado por el investigador Michael Kaisser, de la Universidad de Edimburgo, como parte de la investigación para su tesis doctoral. Aunque se define como una demo, se trata de un sistema que cuenta con un funcionamiento aceptable y recupera tanto información textual –para lo que utiliza únicamente la enciclopedia <i>Wikipedia</i> – como gráfica –extraída del buscador de imágenes de Google–.
RECUPERACIÓN DE INFORMACIÓN	La RI trata de modelar, diseñar e implementar sistemas que sean capaces de proporcionar acceso basado en contenidos, rápido y efectivo. La meta de un SRI es estimar la relevancia de elementos de información a la necesidad de información de un usuario, expresada en una consulta (Croft <i>et al.</i> , 2010)
RECURSOS LINGÜÍSTICOS	Colección de datos en forma textual, ya sea hablada o escrita, o con información sobre lenguas y en donde el objetivo de la tecnología sea el procesamiento del material lingüístico.
PRECISIÓN	Medida de evaluación tradicional. Es la proporción de documentos recuperados relevantes del total de documentos recuperados.
PREGUNTAS DE DEFINICIÓN	Preguntas dirigidas a responder preguntas simples con la fórmula <i>What is...?</i> que dependen de diferentes factores
PREGUNTAS DE LISTA	Preguntas que solicitan un cierto número de respuestas de un mismo tipo y suelen plantearse de forma imperativa (<i>tell me...</i> , <i>name all of London's airports</i> o <i>List 5 pharmaceutical companies that manufacture antibiotics</i>).
PREGUNTAS FACTUALES	Preguntas que principalmente gestionan los SBR en las que se solicitan datos referidos a: persona, tiempo, localización, organización, medida, recuento, objeto, entre otros.
PROCESAMIENTO DEL LENGUAJE NATURAL	Campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. No trata de la comunicación por medio de lenguajes naturales de una forma abstracta, sino de diseñar mecanismos para comunicarse que sean eficaces computacionalmente —que se puedan realizar por medio de programas que ejecuten o simulen la comunicación—.

SEMOTE	SEMOTE, que actualmente no se encuentra operativo, es un SBR que permitía a los usuarios plantear preguntas generales o específicas sobre diferentes dominios temáticos. Se caracterizaba por utilizar una amplia variedad de recursos –desde páginas web dedicadas a temas específicos hasta portales web más elaborados– para extraer las repuestas a las preguntas planteadas. Además, los resultados ofrecidos solían ser bastante exhaustivos.
SISTEMA LKB (<i>LINGUISTIC KNOWLEDGE BUILDING</i>)	Sistema que permite la representación, acceso y manipulación de las unidades léxicas activas en el mismo.
SISTEMAS DE BÚSQUEDA DE RESPUESTAS	Sistemas específicos de RI que ofrecen respuestas precisas y comprensibles a preguntas factuales, de definición o de lista, en lugar de presentar al usuario una lista de documentos relacionados con la búsqueda, de modo que el usuario no ha de leer documentos completos para obtener la información requerida.
SBR MULTILINGÜES	SBR donde el idioma en el que se plantea la pregunta puede ser escritos en varios idiomas a partir de una consulta expresada en un idioma determinado.
SBR TRANSLINGÜE	SBR donde el idioma en el que se plantea la pregunta puede ser recuperación de documentos escritos en un único idioma diferente del utilizado para expresar la consulta.
START	SBR desarrollado por el <i>Massachusetts Institute of Technology</i> que permite a los usuarios plantear preguntas sobre temas muy diversos, ya sean especializados o no. Cuenta con una sencilla interfaz y sus tiempos de respuesta son considerablemente aceptables. Las fuentes de información de las que extraen las respuestas son muy variadas, entre las que se encuentran páginas o sitios web generales como <i>Wikipedia</i> , diccionarios de uso general, <i>Internet Public Library</i> , <i>WorldBook</i> , <i>The World Factbook 2008</i> , entre otros, así como otras páginas o sitios dedicados a una determinada ámbito temático como diccionarios y enciclopedias especializados, etc.
TASK	Se denomina <i>task</i> al conjunto de tareas específicas que determinan los organizadores de los principales congresos en RI para el testeo y evaluación de los sistemas desarrollados. Depende del foro también se denomina <i>lab</i> .
TESAURO	Tipo de lenguaje documental formado por un vocabulario controlado dotado de relaciones que son las que les distinguen de

Recursos y Herramientas lingüísticos para los Sistemas de Búsquedas de Respuestas monolingües y multilingües: Análisis y evaluación

	otros instrumentos terminológicos y que le proporcionan la categoría del lenguaje.
TRACK	Se denominan <i>track</i> a las diferentes sesiones o apartados teáticos en los que se organizan las conferencias en RI y que se van actualizando cada año según los temas de investigación en auge.
TRADUCCIÓN COMPRENSADA	Proceso que consiste en relacionar todas las posibles correspondencias de un término.
TRADUCTOR AUTOMÁTICO	Sistema informático que puede llevar a cabo traducciones de una lengua a otra con o sin la ayuda de la intervención humana.
TRANSLITERACIÓN	Proceso de representar los signos de un sistema de escritura con los signos de otro, de tal modo que el lector pueda entender la grafía original de una palabra, aunque desconozca el idioma original.
TREC	Conocido como el primer congreso en RI, TREC, que se celebra anualmente, realiza un papel muy importante al ofrecer un paradigma de referencia para la investigación en esta área.
TRR	Medida para evaluar la existencia de varias respuestas correctas ofrecidas por un sistema ante una misma pregunta y asigna un valor a cada respuesta de acuerdo con su posición en la lista de resultados recuperados. Si en una lista de respuestas aparece varias veces repetida la correcta, el usuario puede considerarla como más fiable. En estos casos no es suficiente tener en cuenta únicamente la primera respuesta correcta en las evaluaciones, por lo que TRR las tiene en cuenta a todas. Así, si la primera y la tercera respuesta de una lista de resultados son correctas para una pregunta el valor de TRR será $1/1+ 1/3$.