Hybrid Machine Translation Oriented to Cross-Language Information Retrieval: English-Spanish Error Analysis

Chapter · April 2019

DOI: 10.1007/978-3-030-16181-1_18

CITATION
1

READS
114

3 authors:

Juncal Gutiérrez-Artacho
University of Granada
68 PUBLICATIONS **246** CITATIONS
SEE PROFILE

María Dolores Olvera-Lobo
University of Granada
174 PUBLICATIONS **848** CITATIONS
SEE PROFILE

Irene Rivera Trigueros
University of Granada
24 PUBLICATIONS **22** CITATIONS
SEE PROFILE

This is a pre-copyedited version of a contribution published in: Rocha Á., Adeli H., Reis L., Costanzo S. (eds) New Knowledge in Information Systems and Technologies. WorldCIST'19 2019. Advances in Intelligent Systems and Computing, vol 930 published by Springer. The definitive authenticated version is available online via https://doi.org/10.1007/978-3-030-16181-1_18

Hybrid Machine Translation oriented to Cross-Language Information Retrieval: English-Spanish error analysis

Juncal Gutiérrez-Artacho[1][0000-0002-0275-600X], María-Dolores Olvera-Lobo[2,3][0000-0002-0489-7674] and Irene Rivera-Trigueros[1] (✉)[0000-0003-4877-4083]

[1] University of Granada, Department of Translation and Interpreting, Faculty of Translation and Interpreting, C/ Buensuceso, 11, 18003, Granada, Spain
juncalgutierrez@ugr.es; irenerivera@ugr.es
[2] University of Granada, Department of Information and Communication, Colegio Máximo de Cartuja, Campus Cartuja s/n, 18071, Granada, Spain
[3] CSIC, Unidad Asociada Grupo SCImago, Madrid, Spain
molvera@ugr.es

Abstract. The main objective of this study focuses on analysing the automatic translation of questions (intended as query inputs to a Cross-Language Information Retrieval System) and on the creation of a taxonomy of translation errors present in hybrid machine translation (HMT) systems.

An analysis of translations by HMT systems was carried out. From these, there is a proposal of a type 1, 2 or 3 error taxonomy weighted according to their level of importance. Results indicate that post-editing is an essential task in the automatic translation process.

Keywords: Cross-Language Information Retrieval; Hybrid Machine Translation Systems; Translation Errors; Post-Editing

1 Introduction

Cross-language information retrieval (CLIR) is centred upon the search for documents, reconciling queries and documents which are written in different languages [1]. In CLIR systems, translating queries is the most frequent option since they are shorter texts than the documents and their translations have limited computational costs [2–4]. One of the most used resources for undertaking translation processes in the field of CLIR is machine translation (MT) [5–8]. Although the automatic translations offered by these systems lack the level of excellence of human translations [9, 10], they are useful within the translation process. In this regard, it is interesting to

analyse the functioning of MT and the errors detected in the translation of questions from the point of view of their potential as CLIR tools.

The translation market can be defined as global, decentralised, specialised, dynamic, virtual and demanding [11]. Globalisation and the eagerness of businesses to expand into international markets has meant an increase in MT, as in many cases it is impossible to satisfy the demand for translations with human translators alone. In addition what is sought is the maximum cost reduction possible [12]. This situation has caused the profile of the translator to change, as ever more companies in the linguistic sphere are expanding their traditional offer of translation services to include services related to MT and post-editing [13, 14].

For the evaluation of MT systems and contribute to their improvement, error classification plays a key role. The existence of taxonomies including inaccuracies and the most common errors will facilitate their identification at the post-editing stage [15, 16]. The aim of this study focuses on analysing the automatic translation of questions (intended as query inputs to a CLIR system) and in the creation of a taxonomy of translation errors present in MT systems for Spanish (SP) and English (EN).

2 Machine translation and post-editing

One of the current trends in MT −along with Neural and Adaptive MT− is the combination of different types of architectures, giving rise to hybrid technologies [17, 18]. These systems combine the advantages of rule-based MT (RBMT) and statistical MT (SMT). Hybrid machine translation (HMT) systems attempt to solve the problems detected in these two technologies with the objective of producing better quality translations [19–21].

The identification and classification of translation errors is essential for the assessment of the effectiveness of MT systems. A number of different proposals have been implemented according to the motivation of the research, the languages used, or the fact that human or machine translations are being assessed [15].

Laurian [22] distinguished three main types of error: a) errors in isolated words, b) errors in the expression of relationships and c) errors in the structure or presentation of the information. Furthermore, other typologies [23] are based on the quantitative distribution of the errors found during research undertaking and include categories related to morphology, syntax, lexicon, punctuation, style, textual coherence, textual pragmatism and literal translations of the original text. A more detailed proposal [24] establishes a typology organised around four categories of main errors−lexicon, syntax, grammar and errors due to deficiencies in the original text−which, are divided into subcategories. Some classifications [15, 25–27] develop exhaustive hierarchies structured into different levels and depending on the linguistic elements affected.

An interesting aspect in the post-editing of MT systems is the cognitive effort required to correct lexical, grammatical or style errors [28]. Indeed, post-editing work is accelerated and can be carried out much more efficiently when there is guidance to facilitate the task. The guidelines depend on different factors and vary according to the desired quality or post-editing type employed [16]. In 1985, Wagner [29] offered

some post-editing recommendations, which constitutes, still today, a reference as they can be applied to different types of post-editing. These guidelines are still in force and have laid the groundwork for other authors to expand on them [16, 30–32].

3 Methodology

The MT systems used in the study needed to be free of charge, contain Spanish and English amongst the languages available and apply hybrid technology. Systran and ProMT were the only two HMT systems that fulfilled all the requirements[1]. In 2009, Systran introduced the first MT hybrid engine onto the market. For its part, ProMT presented in 2012 the ProMT DeepHybrid system.

The corpus used is a collection questions proposed by CLEF (Cross-Language Evaluation Forum). These collections are used in this type of forum to carry out the assessment of IRS (Information Retrieval Systems) and their techniques, allowing for comparative studies [33–38]. Two collections of questions about European legislation from the ResPubliQA (2009 and 2010) track, related to the Europarl corpus were used, which includes European Parliament acts in various languages [39]. The corpus, comprised, of a sample of 100 questions, was translated from EN-SP and vice-versa by ProMT and Systran. This gave the result of 400 translations. There was an analysis of all errors produced by the HMT systems. The proposed error taxonomy takes other existing classifications as a base [7, 15, 22–27]. To establish the weighting assigned to each error a sample of 200 translations was taken. There was an identification of type 1 errors (minor); type 2 (medium) and type 3 (serious). This process was well defined in order to avoid ambiguity when performing the evaluation. Finally, the most frequent errors were determined in order to assess HMT systems.

4 Results

4.1 Error taxonomy

The taxonomy covers five large groups: a) orthography, b) lexicon, c) grammar, d) semantics and e) discourse. Each of these groups presents various levels (Figure 1):

Orthography
This section includes punctuation errors, confusion between upper and lower case and spelling errors. For example, when translating *registered designations of origin* Systran translates it as *denominaciones de origen registradoas* (spelling error), instead of *registradas*, which would be the correct word.

Lexicon

[1] Nowadays, Systran has already implemented Neural Machine Translation in its MT systems.

This group includes errors referring both to omissions and additions, which can be unnecessary, if they affect functional words like prepositions or articles, or essential ones, if they affect the content of the translation. Errors have also been detected in the translation of abbreviations, initials and proper or institutionalized nouns. In this category are also those words or expressions that the MT systems has not translated.

Grammar

This group includes morphology errors, which affect word order, and errors specific to interrogative sentences. The errors related to morphology are those connected to i) changes in the grammar category; ii) errors in verbs, either tense or person; iii) confusion between the verbs *ser* and *estar*; iv) errors in the translation of English modal verbs, reflexive verbs or pronoun verbs; v) concordance errors, both in the verbal and noun phrase; vi) errors in the contraction of the article; vii) errors in the use of the passive voice, and viii) confusion when determining the subject of the sentence.

Semantics

This category includes errors of meaning, either because there has been a confusion of acceptance with a homographic word, or due to a disambiguation problem. This group also contains errors in the translation of collocations, locutions and set phrases, as some of these are not identified as such and are literally translated.

Discourse

At the level of discourse, there has been an identification of style errors produced either by the use of an inappropriate register or because errors due to orthotypographic conventions. There is also the inclusion of the errors related to linguistic diversity, in this case, between British and American English, and errors created by the translation of words or expressions that should be conserved in their original language.

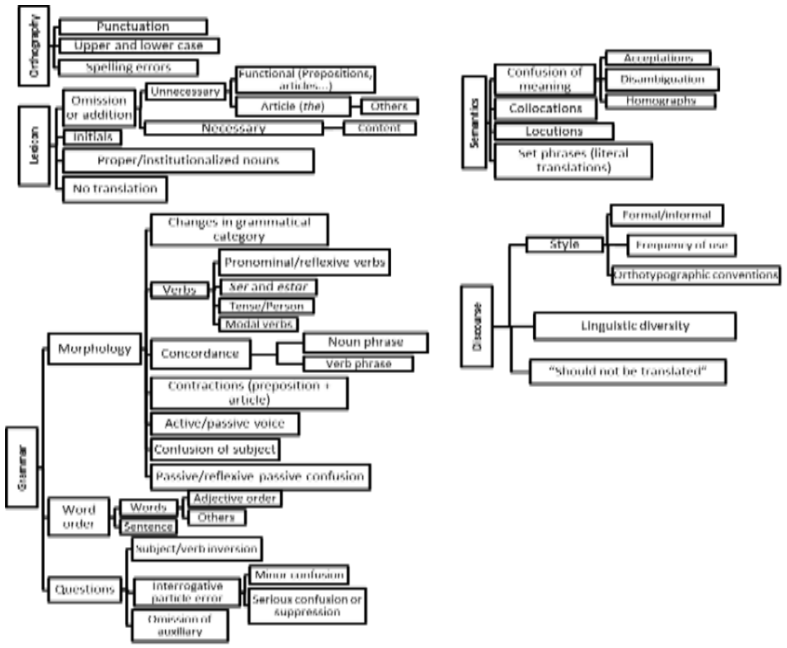

Figure. 1. Taxonomy of translation errors in EN-SP, SP-EN HMT systems (Own authorship)

Once this taxonomy was established, errors were grouped according to their level of importance (Table 1). Type 1 errors are minor, they correspond to errors that do not alter the meaning of the question. These errors are usually minor syntax errors which are not related to the content in itself. Type 2 errors are considered of medium importance, as they are errors, of either syntax or content, which modify the meaning of the question, although without making it unintelligible. In this case, errors in collocations are frequent. Type 3 errors include syntax or content errors that modify the meaning of the question in a way that makes it unintelligible.

Table 1. Types of translation errors in English-Spanish, Spanish-English HMT systems

Type 1 errors	Orthography	• Upper and lower case
	Lexicon	• Omissions or unnecessary additions
	Grammar	• Modal verb error • Confusion of passive and reflexive voice • Contraction of preposition or article • Error in word order • Minor confusion in interrogative particle
	Discourse	• Errors in linguistic diversity • Style errors
Type 2 errors	Orthography	• Punctuation errors
	Grammar	• Changes in grammatical category • Confusion of pronominal and reflexive verbs

		• Errors in verb tense and/or person • Verb or noun phrase concordance errors • Serious confusion or suppression of interrogative particle
	Semantics	• Collocation errors • Expression errors
Type 3 errors	Orthography	• Spelling errors
	Lexicon	• Omissions or unnecessary additions • Error in the translation of proper or institutionalized nouns • Error in the translation of initials • No translation
	Grammar	• Confusion of verbs *ser* and *estar* • Confusion of active and passive voice • Confusion of subject • Error in sentence order • Subject/verb inversion • Omission of auxiliary verb
	Semantics	• Confusion of meaning errors (disambiguation, acceptation or homographs) • Literal translation of set phrases
	Discourse	• "Should not be translated"

The most common type 3 errors were those related to the order of the elements in the sentence (20.97%) and those errors caused by the confusion of meaning, either due to an acceptation error or confusion with a homograph (31.18%), or due to a disambiguation error (18.82%) (Table 2).

Table 2. Type 3 errors in the analysed HMT

			Systran	**ProMT**	**Total**
Type 3 errors	Orthography	Spelling errors	3	1	4 (2.15%)
	Lexicon	Omission	1	1	2 (1.08%)
		Proper/institutionalized nouns	5	11	16 (8.60%)
		Initials	2	3	5 (2.69%)
		No translation	3	5	8 (4.30%)
		Addition	1	2	3 (1.61%)
	Grammar	Subject/verb inversion	7	1	8 (4.30%)
		Verb *ser*/*estar*	3	1	4 (2.15%)
		Sentence order	29	10	39 (20.97%)
		Omission of auxiliary verb	1	-	1 (0.54%)
	Semantics	Acceptation/homographs	23	35	58 (31.18%)
		Disambiguation	19	16	35 (18.82%)
	Discourse	"Should not be translated"	1	2	3 (1.61%)
	Total		98	88	186

For medium errors (Table 3), there is a prominence of collocation errors (37.36%) and concordance errors, both in the verb phrase (18.68%) and the noun phrase (10.99%).

Table 3. Type 2 errors in the analysed HTM

			Systran	ProMT	Total
Type 2 errors	Grammar	Changes in grammatical category	3	5	8 (8.79%)
		Pronominal/reflexive verb	2	3	5 (5.49%)
		Verb tense	-	6	6 (6.59%)
		NP Concordance	4	6	10 (10.99%)
		VP Concordance	9	8	17 (18.68%)
		Serious / omission interrogative particle	4	2	6 (6.59%)
	Semantics	Collocation	19	15	34 (37.36%)
		Locution	4	1	5 (5.49%)
	Total		45	46	91

In the minor errors (Table 4) there is a prominence of functional additions or omissions. (17.11% and 7.24%, respectively), errors in word order (15.13%), errors in the interrogative particle (9.21%) and errors in the use of upper case (14.47%). However, regarding the latter, it must be pointed out that in a corpus of questions on European legislation there were numerous names of bodies, institutions, committees, etc., therefore, this type of error was expected.

Table 4. Type 1 errors in the analysed HMT

			Systran	ProMT	Total
Type 1 errors	Orthography	Upper or lower case	15	7	22 (14.47%)
	Lexicon	Functional omission	6	5	11 (7.24%)
		Functional addition	16	10	26 (17.11%)
	Grammar	Word order (others)	15	8	23 (15.13%)
		Passive/reflexive passive	12	1	13 (8.55%)
		Modal verb	2	6	8 (5.26%)
		Interrogative particle	11	3	14 (9.21%)
		Preposition + article contraction	2	2	4 (2.63%)
	Discourse	Linguistic diversity	1	2	3 (1.97%)
		Style errors	21	7	28 (18.42%)
	Total		101	51	152

A total of 422 errors were identified according to Table 5, of which 34.4% were minor (type 1), 20.1% were medium (type 2) and 45.5% were considered as serious (type 3). ProMT is the MT system that obtained the best results, although without large differences, returning 43.3% of the total errors compared to the 56.6% of Systran.

Table 5. Types of translation errors in the analysed HMTs

Type 1 ProMT	50 (11.9%)	Type 1 Systran	95 (22.5%)	Total Type 1	145 (34.4%)
Type 2 ProMT	44 (10.4%)	Type 2 Systran	41 (9.7%)	Total Type 2	85 (20.1%)
Type 3 ProMT	89 (21.1%)	Type 3 Systran	103 (24.4%)	Total Type 3	192 (45.5%)
Total ProMT	**183 (43.4%)**	**Total Systran**	**239 (56.6%)**	**Total**	**422**

If the languages involved are considered according to Table 6, when translating SP-EN it is observed how most errors are type 3, followed by type 1 errors, whereas there is a reduced percentage of Type 2. In SP-ES translation, there is a predominance of Type 3 errors and there is also an increase in type 2 errors, this could be since translating to Spanish can produce concordance errors (type 2 errors). In terms of total errors, the HMTs perform better overall when translating SP-EN.

Table 6. Errors according to translation direction

Total Errors	**Total SP-EN**	184 (43.60%)	**Type 1 SP-EN**	76 (41.3%)	**Type 2 SP-EN**	16 (8.7%)	**Type 3 SP-EN**	92 (50%)
	Total EN-SP	238 (56.39%)	**Type 1 EN-SP**	69 (29%)	**Type 2 EN-SP**	69 (29%)	**Type 3 EN-SP**	100 (42%)

5 Conclusions

In machine translations numerous types of errors appear that depend on the grammar of the languages involved, the topic of the translations or their complexity, amongst other factors. MT systems, although constituting support tools, require post-editing, as a human task linked to professional translators, which plays a fundamental role in the translation process.

The main contribution of this study, put forward from the perspective of CLIR, focuses on the establishment of a taxonomy of errors specific to the MT of questions, a type of input frequently employed in their CLIR queries. In addition, there has also been an in-depth analysis regarding existing classifications, concerning errors related to style, register, frequency of use and errors related to orthotypographic conventions. Regarding errors related to verbs, there has also been an identification of new cases such as those caused by the confusion of pronoun or reflexive verbs, the verbs *ser* −used to talk about permanent or lasting attributes− and *estar* −used to indicate temporary locations and states, and in the translation of modal verbs. Referring to errors caused by functional omissions or additions, already identified in other typologies, there is a distinction of those errors produced by the incorrect translation of the arti-

cle, *the*. Finally, in order-related errors, with regards to the already existing classifications, those related to adjective order are added.

Lastly, it is worth mentioning that, although the quality of MT is still deficient, as proved in previous work [21], the demand for this type of translation tool is generalised and growing, especially in the multilingual context of the Internet. Therefore, we should focus our efforts in their improvement.

References

1. Zhou, D., Truran, M., Brailsford, T., Wade, V., Ashman, H.: Translation Techniques in Cross-language Information Retrieval. ACM Comput. Surv. 45, 1:1--1:44 (2012).
2. Banchs, R.E., Costa-Jussà, M.R.: Cross-Language Document Retrieval by Using Nonlinear Semantic Mapping. Appl. Artif. Intell. 27, 781–802 (2013).
3. Sharma, V.K., Mittal, N.: Cross Lingual Information Retrieval (CLIR): Review of Tools, Challenges and Translation Approaches. In: Satapathy S., Mandal J., Udgata S., Bhateja V. (eds) Information Systems Design and Intelligent Applications. Advances in Intelligent Systems and Computing, vol 433. Springer, New Delhi (2016).
4. Hull, D.A., Grefenstette, G.: Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '96. pp. 49–57. ACM Press, New York, New York, USA (1996).
5. Olvera-Lobo, M., Gutierrez-Artacho, J.: Language resources used in multi-lingual question-answering systems. Online Inf. Rev. 35, 543–557 (2011).
6. García-Santiago, L., Olvera-Lobo, M.-D.: Analysis of automatic translation of questions for question-answering systems. Inf. Res. 15, (2010).
7. Olvera-Lobo, M.D., Garcia-Santiago, L.: Analysis of errors in the automatic translation of questions for translingual QA systems. J. Doc. 66, 434–455 (2010).
8. Madankar, M., Chandak, M.B., Chavhan, N.: Information Retrieval System and Machine Translation: A Review. Procedia Comput. Sci. 78, 845–850 (2016).
9. Allen, J.: Post-editing. In: Somers, H.L. (ed.) Computers and Translation: a translators guide. pp. 297–317. John Benjamins, Amsterdam/Philadelphia (2003).
10. Koponen, M.: Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. J. Spec. Transl. 131–148 (2016).
11. Olvera-Lobo, M.D., Castro-Prieto, M.R., Quero-Gervilla, E., Munoz-Martin, R., Munoz-Raya, E., Murillo-Melero, M., Robinson, B., Senso-Ruiz, A., Vargas-Quesada, B., Dominguez-Lopez, C., int, A., Granada, U.: Translator training and modern market demands. Perspect. Transl. 13, 132–142 (2005).
12. Lagarda, A.L., Ortiz-Martinez, D., Alabau, V., Casacuberta, F.: Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. SPEECH Lang. 32, 109–134 (2015).
13. Temizöz, Ö.: Postediting machine translation output: subject-matter experts versus professional translators. Perspectives (Montclair). 24, 646–665 (2016).
14. Torres-Hostench, Olga, Presas, Marisa y Cid-Leal, P. (coords): El uso de traducción automática y posedición en las empresas de servicios lingüísticos españolas: Informe

de investigación ProjecTA 2015. , Bellaterra (2016).
15. Costa, Â., Ling, W., Luís, T., Correia, R., Coheur, L.: A linguistically motivated taxonomy for Machine Translation error analysis. Mach. Transl. 29, 127–161 (2015).
16. Mesa-Lao, B.: Introduction to post-editing – The CasMaCat GUI 1 . Introduction : Why post-editing MT outputs ? (2013).
17. Costa-Jussà, M.R., Fonollosa, J.A.R.: Latest trends in hybrid machine translation and its applications. Comput. Speech Lang. 32, 3–10 (2015).
18. Labaka, G., España-Bonet, C., Màrquez, L., Sarasola, K.: A hybrid machine translation architecture guided by syntax. Mach. Transl. 28, 91–125 (2014).
19. Hunsicker, S., Yu, C., Federmann, C.: Machine Learning for Hybrid Machine Translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation. pp. 312–316. , Montreal (2012).
20. Tambouratzis, G., Athena, I., Centre, R., Amaroussiou, P.: Comparing CRF and template-matching in phrasing tasks within a Hybrid MT system. In: Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra). pp. 7–14 (2014).
21. Gutiérrez-Artacho, J., Olvera-Lobo, M.-D., Rivera-Trigueros, I.: Human Post-editing in Hybrid Machine Translation Systems: Automatic and Manual Analysis and Evaluation. In: Rocha, A., Adeli, H., Reis, L.., and Costanzo, S. (eds.) nds and Advances in Information Systems and Technologies. WorldCIST'18 2018. Advances in Intelligent Systems and Computing, vol 745. pp. 254–263. Springer, Cham (2018).
22. Laurian, A.M.: Machine Translation: What type of post-editing on what type of documents for what type of users. In: Proceedings of the 10th Internacional Conference on Computational Linguistics and 22nd annual meeeting on Association for Computational Linguistics. pp. 236–238 (1984).
23. Krings, H.: Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes. Kent State University Press (2001).
24. Schäfer, F.: MT post-editing: How to shed light on the "unknown task " - Experices made at SAP. Jt. Conf. 8th Int. Work. Eur. Assoc. Mach. Transl. 4th Control. Lang. Appl. Work. 133–140 (2003).
25. Farreús, M., Costa-Jussà, M.R., Morse, M.P.: Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. J. Am. Soc. Inf. Sci. Technol. 63, 174–184 (2012).
26. Sivakama, S., Prema, V., Savitha, G.: Engineering and Scientific International Journal (ESIJ) A Comparative Study of Occurrence of Errors in Machine Translation in a Multilingual Environment. Eng. Sci. Int. J. . 3, (2016).
27. Vilar, D., Xu, J., D'Haro, L., Ney, H.: Error analysis of statistical machine translation output. In: Proceedings of LREC (2006).
28. Nunes Vieira, L.: Cognitive Effort in Post-Editing of Machine Translation : Evidence from Eye Movements , Subjective Ratings , and Think-Aloud Protocols, http://hdl.handle.net/10443/3130, (2016).
29. Wagner, E.: Post-editing systran - a challenge for commission translators. Terminol. Trad. 3, (1985).
30. O' Brien, S.: Researching and Teaching Post-Editing, http://www.mt-archive.info/MTS-2009-OBrien-ppt.pdf, (2009).
31. TAUS: MT Post-editing Guidelines, https://www.taus.net/academy/best-

practices/postedit-best-practices/machine-translation-post-editing-guidelines.
32. Hu, K., Cadwell, P.: A Comparative Study of Post-editing Guidelines. Balt. J. Mod. Comput. 4, 346–353 (2016).
33. Olvera-Lobo, M.-D., Gutiérrez-Artacho, J.: Evaluación de los sistemas QA de dominio abierto frente a los de dominio especializado en el ámbito biomédico. In: I Congreso Español de Recuperación de Información (CERI 2010). pp. 161–169. , Madrid (2010).
34. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Question-answering systems as efficient sources of terminological information: an evaluation. Heal. Inf. Libr. J. 27, 268–276 (2010).
35. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Multilingual Question-Answering System in Biomedical Domain on the Web: An Evaluation. Lect. Notes Comput. Sci. 6941, 83–88 (2011).
36. Olvera-Lobo, M.-D., Gutierrez-Artacho, J.: Performance Analysis in Web-based Question Answering Systems. Rev. Esp. Doc. Cient. 36, (2013).
37. Olvera-Lobo, M.-D., Gutierrez-Artacho, J.: Question Answering Track Evaluation in TREC, CLEF and NTCIR. In: Rocha, A and Correia, AM and Costanzo, S and Reis, LP (ed.) New contributions in information systems and technologies, Vol 1, Pt 1. pp. 13–22 (2015).
38. Gutiérrez Artacho, J.: Recursos y herramientas lingüísticos para los sistemas de búsqueda de respuestas monolingües y multilingües, (2015).
39. Koehn, P.: Europarl : A Parallel Corpus for Statistical Machine Translation. MT Summit. 11, 79--86 (2005).