Snomed2Vec: representation of SNOMED CT terms with Word2Vec

Ignacio Martínez Soriano
MedLab Media Group S.L. &
Hospital University "Rafael Méndez"
Lorca (Murcia), Spain
ignacio.martinez@carm.es

Juan Luis Castro Peña
dept. of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain
castro@decsai.ugr.es

Jesualdo T. Fernandez Breis
dept. of Computer Science and Systems
University of Murcia
Murcia, Spain
jfernand@um.es

Ignacio San Roman
Chief Artificial Intelligence Officer
MedLab Media Group S.L.
Madrid, Spain
i.sanroman@medlabmg.com

Adrian Alonso Barriuso
Research Technical Lead
MedLab Media Group S.L.
Madrid, Spain
a.alonso@medlabmg.com

David Guevara Baraza
Nurse Expert Information System
Hospital University "Rafael Méndez"
Lorca (Murcia), Spain
david.guevara@carm.es

Abstract*— Hospital Information Systems (H.I.S) use Electronic Health Record to store heterogeneous data from the patients. One important goal in this kind of systems is that the information must be, normalized and codify with a clinical terminology to represent exactly the healthcare meaning. Usually this process need human experts to identify and map the correct concept, this is a slow and tedious task. One of the most widespread clinical terminologies with more projection is Snomed-CT. This is an ontology multilingual clinical terminology that represent the clinical concepts with a unique code. We introduce in this paper Snomed2Vec, new approach of semantic search tool to find the most similar concepts using Snomed-CT. This is an ontology based named entity recognition system using word embedding, that suggest what is the most similar concept, that appear in a text. To evaluate the tool we suggest two kind of validations, one against a corpus gold with diagnostic from clinical reports, and a social validation, with a public free web access. We publish an access web to the academic world to use, test and validate the tool. The results of validation shows that this process help to the specialist to the election of choose the correct concepts from Snomed-CT. The paper illustrates 1) how create the initial big corpus of texts, to train the word2vec models, 2) how we use this vector space model to create our final Snomed2Vec vector space model, 3) The use of the cosine similarity distance, to obtain the most similar concepts, grouping by the hierarchies from Snomed-CT. We publish to the academic world: https://github.com/NachusS/Snomed2Vec access to the public web tool, and the notebook, for develop and test this paper.

Keywords— word2vec, snomed CT, semantic similarity, word embedding, ontology matching, named entity recognition.

I. INTRODUCTION

Electronic Health Records (EHR) include heterogeneous information from different sources. The majority of the Hospital Information System, use a representation of the clinical information unstructured as free text. The access to the knowledge inside theses records is very hard to the management care patients, medical research or decision support systems. So we need clinical tools to identify the meaning of the text, codify the clinical concepts with a single standard code, to permit semantic interoperability and improve the access to the clinical knowledge. This kind of tools need a clinical terminology to do this process.

Systematized Nomenclature of Medicine-Clinical Terms "Snomed CT" [1] is an ontology [2] multilingual clinical terminology to identify healthcare meaning from the clinical texts, it's representing clinical concepts with a unique code [3], allowing to identify concepts, synonyms and relationships with other clinical concepts. The entities has organized like a taxonomy of terms, and a framework of rules to define every term with one meaning. The taxonomy structure has parent-child relationships between the terms.

Usually to find a specific clinical concept in the ontology, you navigate through its hierarchy or use a syntactic search, to find the correct entity into the description terms. These process to find a specific concept it is hard, slow and inadequate, for the time you spend and the problem in the clinical language to choose the correct meaning for a case. Sometimes it is necessary a semantic search to find the correct meaning of the concept, when you need to know the abbreviations, acronyms in the language, for designed words that depends of the specific clinical specialization. This process need a human specialist in the same knowledge field to recognize, what is the correct meaning of a clinical text. Usually the human expert that codify the clinical text has general knowledge in the medicine field, and they find problems to choose the correct concept in the clinical ontology to normalize and codify a text.

To do this is necessary use a named entity recognition (NER) an information extraction system that get the named entities from a free text, One of this kind of tools is the ontology-based named entity recognition, the process to identify a entity inside the ontology classes and axioms.

There are some clinical named entity recognition frameworks, for automatic semantic Tag (cTakes [35], MetaMap [33], MedLEE [32], KnowledgeMap [34]). Ones apply rules-based system on a comprehensive clinical vocabulary dictionary-based systems (gazetters) and Machine learning algorithms. Initially the main language to use these tools was English [36], but they are already releasing new versions to be able to adapt them to Spanish, as "MetaMap". For Spanish NER we have IxaMed[37], STMC[18], DNER[17]. The main goal of all of these tools are focus in the automatic semantic tag and name entity recognition with a unique code, in our use case we need a kind of recommendation system that suggest similar concept from different classes of a ontology. With our new approach Snomed2Vec, ontology-based NER with word embedding we can do it.

We present Snomed2Vec, a tool to improve the ontology-based named entity recognition process, that suggest what is the most similar concepts from Snomed-CT grouping by its top level hierarchy. So the system suggest to the human

expert user, a set of most similar concepts named in the different "classes" hierarchies of Snomed-CT.

It is a novel usefulness approach for Snomed-CT, because can show the relations from a concept with different concepts from other branches hierarchies, using a similarity measure.

To develop the tool, we apply different stages:

1) *First stage, get a big corpus text and preprocessing.*
2) *Second stage, Create the vectors spaces models.*
3) *Implement the similarity measure method.*

The structure of this paper is: In section II we overview the situation of ontology-based named entity recognition technologies in the medical domain. In section III we describe the logic model of Snomed-CT ontology and the python library that we use to get vector space models, genism [26]. In section IV we show our implementation of our ontology-based named entity recognition word embedding, Snomed2Vec. In section V we introduce the results of our evaluation and examples of use case. In section VI we present the conclusions and different approaches for future work.

Availability: https://github.com/NachusS/Snomed2Vec access to the academic public, the web tool, and the notebook for develop and test this paper.

II. BACKGROUND AND RELATED WORK

Exists different approach of Ontology-based named entity recognition:

A. Ontology Matching

Ontology Matching try to solve a problem of semantic heterogeneity [9]. Lexical information, including names and labels describing entities are valuable to matching systems. A simple method is calculating the surface similarity between the strings. A review of the string-based metrics have been evaluated [10]. Edit distance was adopted by many matching systems as RiMOM [11], ASMOV [12] and agreementMaker [13], to measure the similarity between two words. And approach to use word embedding for the matching of the entities, to the WordNet ontology, we can found at [14].

B. Name-based Techniques

The aim of this techniques is to identify the most similar string in a text search process. As in [7] we are going to search the most similar term, from the all descriptions of Snomed-CT. There are several techniques to do this, but the distance of the cosine and the Levenshtein distance are the most characteristics. The cosine distance similarity, is measured by identifying the number of common words between two strings and their relation to the word of each string, while Levenshtein is an editing distance, the measure of similarity is the changes that we need to do in a world to be equal like other.

C. Word embedding Ontology Matching

Previous works [4] showed a novel solution named OPA2Vec (Ontologies Plus Annotations to Vectors) that extend Onto2Vec [5]. The goal of these papers are, the use of formal content of ontologies and the meta-data expressed as annotation axioms to generate feature vectors for any named entity in an ontology able to represents biological entities in ontologies can be represented using word embedding[8] model Word2Vec. OPA2Vec mix formal ontology and annotation axioms applying a word2vec pre-trained model English PubMed database [6], on the paper's abstracts and body-text, from our collected data. We use the same philosophy process, adapting Snomed-CT ontology to create a big corpus with the description terms and hierarchies to develop a vector space model.

III. SEMANTIC TOOLS

A. Snomed-CT Ontology

Snomed-CT, "Systematized Nomenclature of Medicine Clinical Terms", is the most complete clinical terminology in the world [15]. It is multilingual terminology and provides a collection of medical terms definitions, codes and synonyms used in clinical documents and reports. Snomed-CT covers a broad spectrum of clinical concepts that appear in a patient's medical record, it is composed of most of the clinical concepts used in health documents. It allows to interrelate different concepts and express different levels of detail, using expressions that contain one or more concepts.

It has a representation in different languages and dialects, using a set of reference, allowing the mapping between different health classification systems.

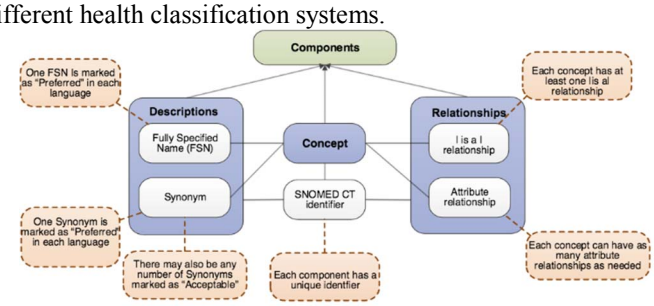

Fig. 1. Snomed-CT Logic Model Structure

The structure of the logic model component (see Fig. 1), is representing the core content of the terminology.

Snomed-CT Component is organized, in concepts, descriptions and relationships. A **concept** is the way to represent a clinical meaning, with a unique numerical code. **Descriptions**, representing the human readable term of a concept. **Relationships** represents an association between concepts. The meaning of the concept has a hierarchy structure, form general to more detailed.

Concepts have several descriptions associated with them. The same concept contains at least two types of descriptions, one is called Fully Specified Name (**FSN**), with which a concept can be univocally identified, and another with the same meaning that represents a **synonym**, according to the language or dialect used.

There is another value, in the "**description type**" that identifies, if that description is a preferred one, according to the language and it is known like "**preferred term**".

The descriptions are composed of "**terms**" that are strings of characters, generating words or phrases that the human being can understand.

A **relationship** represents an association between two concepts. These relationships define characteristics of the concept. The links between concepts, contains a meaning. Snomed-CT is an ontology, so "relationships" can be represented as a triplet (Object, Attribute, Value). The type of relationship that identifies and define a concept is "**Is-a**".

Concepts Model of Snomed CT is defined, using a combination of editorial rules and formal logic. The structure of the model has a hierarchy from the root concept, to the descendent with a "is-a" relationship. This tree structure has a supertype root concept, and the rest are subtypes of this.

There are a "**Top level Concepts**", defines like the direct subtypes of the root concept, and define the name and class of the branch.

There are 19 top level Concepts with its branches. The principals are:

- |**Clinical finding**| represents the result of a clinical observation, it used to represent diagnoses.
- |**Procedure**| represents activities carried out in the provision of health services.
- |**Substance**| represents general substances, the chemical of pharmaceutical/biological products.
- |**Body structure**|, |Observable entity|, |Physical object|, |Event|, etc…[16]

B. Word Embedding with Word2Vec

Word Embedding is a kind of statistical language model where a sequence of words are represented by a probability distribution, where words from a vocabulary are mapped to vectors of real number. Creating a vector space model, for the text is analyzing. The vectorial space created, are represented as a dense real-valued low dimensional matrix **M** of size **VxD**, where **V** is the vocabulary size and **D** is the predefined embedding dimension of the vector. Bengio et al.[21] and T. Mikolov et al. [22] proposed different neural networks to train the word embedding, where the probability of a word given by a previous word was estimating by the cross-entropy criterion.

Word2Vec is a semantic learning framework with a shallow neural network to learn the representations of the words in the text, the vector representations of words with similar context tend to be close to each other in the vector space.

It's available in two architectures, **CBoW** (Continous Bag of Words), Fig. 2, which uses a context to predict a target word, and the **skip-gram** model that try to predict the surrounding words given the current word.

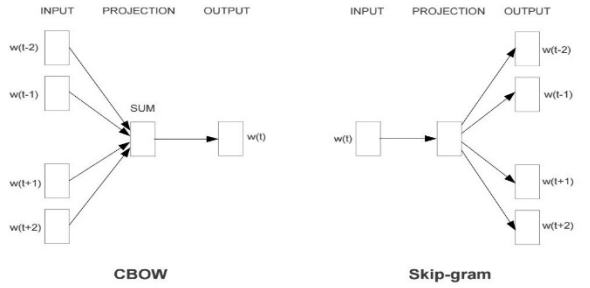

Fig. 2. Continous Bag of Word (CBoW) and Skip-Gram Models.

Word2Vec uses a trick, we don't train a simple neural network with a single hidden layer to perform a certain task. The goal is just to learn the weights of the hidden layer. These weights of the hidden layer it will be the values of the "word vectors" that we're trying to learn.

C. Big Data Corpus

To train our Snomed2Vec Model with Word2Vec, we need a big data corpus. We define two approach:

1. One representation using a local domain with the clinical emergency discharge reports of the Hospital University "Rafael Méndez" from Lorca (Murcia) Spain. We use 636439 clinical reports, form 2009-2017, after normalized the text, (lowering the words and eliminate punctuation sign) we have a Data set prepared to train our Word2Vec Model. The main special feature of this kind of text emergency clinical reports is that has made, with short phrases, many abbreviation words. This kind of corpus to train the model Word2Vec, improve the accuracy like we see at [17][18], for the facility to disambiguate abbreviates, and identify typewrite mistakes.
2. The other representation model is with a general domain, like Wikipedia. We dump the articles of the Spanish Wikipedia until December 2018, and normalized the text, like the other dataset corpus. To extract the text from the dump we use WikiExtractor [19], First we tried use a pretrained corpus of a Spanish Wikipedia with Word2Vec [20]. But for our approaches are better than all the words were normalized.
3. To get our final Big data Corpus to train we use the Spanish descriptions terms of Snomed-CT, join in every corpus, so we have to final Corpus, one representing a local domain with (clinical discharge emergency reports + descriptions terms of Snomed-CT concepts) , and other one, representing a general domain (Spanish Wikipedia dump + descriptions terms of Snomed-CT)

IV. IMPLEMENTATION

To develop our Snomed2Vec approach we did the next stages:

A. Big Data Corpus to vector space model

1) Preprocessing Text, for Initial Corpus:

To create our Snomed2Vec, model we need a big corpus text to train the word2vec framework. For our two approach, one local domain and other general domain, we use:

Premise for all the corpus, we normalized the text, lowing words and remove some punctuation sign. We remove only some stopwords, because we need every all the words to identify the negative expressions and the short words like abbreviations and acronyms in the text. To do this we use NLTK NLP [23] library for python language.

Creation of the big text corpus:

- DataSet01, all the discharge emergency reports at the Hospital from 2009-2017 (636439 reports). We use in every row a report in one line.
- DataSet02, the Spanish Wikipedia dump, with the same structure, for every row, we put the title, abstract and article in a continuous line.
- DataSet03, we use the descriptions term of the Snomed-CT concept in every row.

All the DataSets are the text tokenized, to prepare the input for the Word2Vec algorithm.

The final DataSets that we use is a combination of the three corpus. Final_Corpus01=(DataSet01+DataSet03), Final_Corpus02=(DataSet02+DataSet03).

In addition of the Final_Corpus to get the initials word space model, we need a specific structure to get the text Snomed2Vec Corpus. We define form Snomed-CT the next corpus: Snomed2Vec_Corpus: |**Id-Concept | Description Tokens | Top Level Hierarchy |** Fig.3. To get the top level hierarchy of every Concept from Snomed-CT we use, a transitive closure [24][25] a comprehensive view of all the

supertype ancestors of a concept derived by traversing all the |is a| relationships between that concept and the root concept. "*A the **transitive closure** of a binary relation R on a set X is the smallest relation on X that contains R and is transitive*".

idConcept	corpusTerm	jerarquia
102002	[hemoglobina, okaloosa]	Sustancia
102002	[hb, 48, cd7, leu, arg]	Sustancia
103007	[virus, fibroma, ardillas]	Organismo
104001	[escision, lesion, rotula]	Procedimiento
104001	[escision, local, lesion, o, tejido, rotula]	Procedimiento

Fig. 3. Initial Snomed2Vec Corpus.

B. *Snomed2Vec Model with Word2Vec Skip-Gram*

The Word2Vec framework we use to create the space word vector model is gemsim [26] python library.

The model we choose to train the corpus texts, it is **Skip-Gram** model, and the parameters based on the bibliography [7][17][18] we choose, *model:Word2Vec(corpus, size=300, window=8, workers=4, min_count=1, sample=0.05, sg=1, iter=5, hs=0).*

Parameters explication:

*Corpus:*the tokenized and normalized text. *Size* of the vector word, *window:* the slide window size of the word target context, *min_count*: the number of word frequency that we discard, in our case, we choose only 1, because we want that all the words from the Snomed descriptions terms has a vector, *sg:* Skip-Gram model.

With this definitions skip-gram word2vec framework we created two space word vectors models, one from Final_Corpus01, to represent a local domain representation, and other one from Final_Corpus02, to represent a general domain representation of the space word vector model.

With these two training space model, we can create our final space vector model for Snomed2Vec.

To represent the vector description term for every Snomed-CT concept, we apply the next design:

If we denoted $\boldsymbol{v(w)}$ as the vector of a word $\boldsymbol{w}$ in the Model $\boldsymbol{M}$. We extend the model from the words of Snomed description term(d), by it **sum of the vectors** words in (d),

Given a description sentence $d = w_1 w_2 \dots w_n$

We apply the vectors word model to the Snomed description text:

$$\forall description(d): w_1 \dots w_n \Rightarrow v(d) = \sum_{i=1}^{n} v(w_i)$$

With this we created the final Snomed2Vec Model, with this structure: **|Id-Concept | Description Tokens | Top Level Hierarchy | Description Term Vector |** (Fig. 4.)

```
SnomedWork.head()
```

	idConcept	jerarquia	corpusTerm	vecSnom
0	102002	Sustancia	[hemoglobina, okaloosa]	[-0.660218, 0.112245, 0.067039, 0.441094, -0.50456, -0.724666, -0.107986, 0.313558, 0.198024, 1....
1	102002	Sustancia	[hb, 48, cd7, leu, arg]	[-0.317944, 1.87942, 1.76004, 1.72051, 0.75724, -1.13105, -0.499457, -1.84984, -0.570444, -0.630...
2	103007	Organismo	[virus, fibroma, ardillas]	[-0.14364, -0.006139, 0.080456, 0.936581, -1.16385, -0.552844, 0.081489, -0.584596, -0.185695, 0...
3	104001	Procedimiento	[escision, lesion, rotula]	[-1.0588, -0.371483, -1.84789, 1.65618, -0.235278, 0.101576, -1.08105, 0.863104, -0.374193, 1.04...

Fig. 4. Final Snomed2Vec structure Model.

You can see an example of the flow process generation Snomed2Vec Model-01 in Fig.5 and Fig.6.

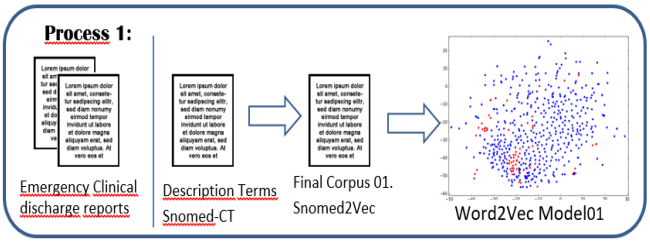

Fig. 5. Created Word2Vec Model with final Corpus.

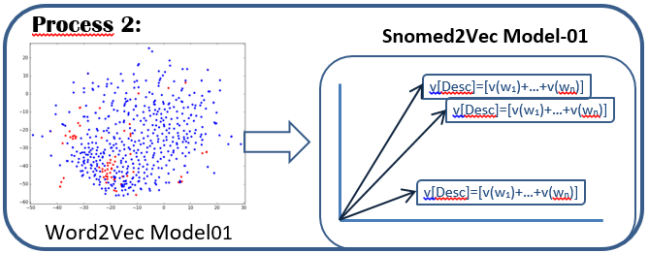

Fig. 6. Applying the Word2Vec to the Descriptions term, we get Snomed2Vec Final Model.

C. *Similares Method (cosine similarity)*

To identify the similarity between two concepts we use the cosine distance of the vectors from the model.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

We adapt the method to get the cosine similarity using the vector space model **Snomed2Vec**, we get the most similar vector from this new space, we choose the implementation of gensim in Word2Vec, and create a new method **"similares"** to get the most similar concept from a text query, using the vector Space Model Snomed2Vec.

The novel approach, is that we use a method to get the cosine distance similarity from two vector over two different space vector model:

1. One Space Vector Model, created with the words of Final_Corpus01 or over Final_corpus02. Vectors of every word in the vocabulary.
2. Other Space Vector Model "Snomed2Vec", created with the sum of the vectors words of the descriptions terms, for every concept we have a vector $v(d)$ that represent the description text of the concept.

When we get the vector of a query text, $v(q)$, and we apply the cosine distance between this $v(q)$ over all the vectors of the Snomed2Vec space model, getting a number **n** of most_similar vectors. We use the same implementation of genism library, to get this method more efficient possible.

We have implemented three method:

- **vectorSnomed** (queryText, model, size):
  - *Inputs:* queryText(q), model the space of vector model to check, size= the size of the vector.
  - *Outputs*: $(q) = w_1 \dots w_n \Rightarrow v(q) = \sum_{i=1}^{n} v(w_i)$
- **Mas_similar**(txt, model, topn):
  - *Inputs:* txt, to search the more similaries, model = vector space to check, topn, number of more similaries words with its distance to get.
  - *Output*: a list of ***n*** items more closed in the space model from the vector(txt), words of the vocabulary and the cosine distance, more closed to 1.0 its more similar each other.

- **Similares** (term, jer='all', nMax=3):
  - *Inputs:* **term** = a query text, a word or a phrase, **jer =** it is the hierarchy top level of Snomed, to get the most similar concepts from that hierarchy, **nMax**, you can get the number of concepts you want to get, depending of the hierarchy.
  - Inside parameter: there is a threshold **topn**=1000, to get the 1000 closest concepts from Snomed depend of the query text.
  - *Output:* ***results***: a list with the the ***nMax*** more similar concepts form Snomed-CT grouping by its hierarchy. ***Jerarquias*** a counter list of **topn** more closed of the query term, grouping by its hierarchy

D. Visualization results.

To visualize a high-dimensional dataset, like word vector produce by word2vec model, we use [27] t-Distributed Stochastic Neighbor Embedding (t-SNE), is a technique for dimensionality reduction that is particularly well suited for the visualization of high dimensional datasets by giving each data point a location in a two or three-dimensional map. This is a variation of Stochastic Neighbor Embedding [28] and produces better visualizations by reducing the tendency to crowd points together in the center of the map.

We use t-SNE scikit-learn[29] implementation to represent visually the insight relation between the subsets of concepts by ***results*** of ***similares()*** method, and the query text.

There are two frameworks for visualization of high dimensional datasets, Tensorflow projector [30] and word2vec explorer [31] very interesting to use.

Example Fig. 7.

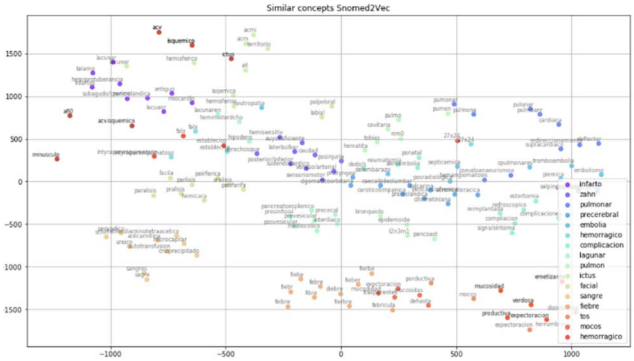

Fig. 7. Visual Distribution of the concepts vector Snomed2Vec

V. CASE STUDY. EVALUATION

a) Use cases:

We show the next uses cases, to show how the Snomed2Vec can disambiguate acronyms and abbreviations, and get the more closed concepts from differents hierarchies.

Prueba para desambiguar abreviaturas:

```
# itu = Insuficiencia tracto urinario.
sent = 'itu'
result, jerarquias = similares(sent, jer='all',nMax=3)
```

```
result
```

```
[('197850006|cistitis tricomoniasica|Hallazgo clínico', 0.7546818256378174),
 ('8776008|cistitis amebiasis|Hallazgo clínico', 0.7383185029029846),
 ('8776008|cistitis amebiana|Hallazgo clínico', 0.7371204495429993),
 ('346575006|farmaco miscelaneo cistitis|Producto farmacéutico / biológico',
  0.6103503108024597),
```

```
sent = 'tce'
result, jerarquias = similares(sent, jer='all',nMax=3)
result
```

```
[('62564004|conmocion perdida conocimiento|Hallazgo clínico',
  0.7439171075820923),
 ('82271004|traumatismo craneoencefalico|Hallazgo clínico',
  0.7432726621627808),
```

```
sent = 'artritis postraumatica 5 dedo de pie derecho'
result, jerarquias = similares(sent, jer='all',nMax=2)
result
```

```
[('1074921000119104|artritis pie derecho|Hallazgo clínico', 0.91978520154953),
 ('1074301000119109|artritis traumatica pie derecho|Hallazgo clínico',
```

```
#catarro vias altas
sent = 'cva'
result, jerarquias = similares(sent, jer='all',nMax=2)
result
```

```
[('135882008|catarro febril|Hallazgo clínico', 0.7946891188621521),
 ('63129006|catarro|Estructura corporal', 0.7820351123809814),
 ('95885008|faringitis micoplasmatica|Hallazgo clínico', 0.7787644863128662),
```

```
# Eleccion jerarquia = Procedimiento
sent = 'edema de mano tras yeso'
result, jerarquias = similares(sent, jer='Procedimiento',nMax=10)
result
```

```
[('239683008|aplicacion inmovilizador yeso antebrazo muñeca articulada|Procedimiento',
  0.8045514822006226),
 ('46223001|aplicacion ferula estatica dedo mano|Procedimiento',
  0.80344927309436),
```

b) Evaluations.

Our Snomed2Vec framework, a new approach for Semantic search match to the Ontology Snomed-CT was validated in a study case.

We suggest a social academic validation in order to improve the Snomed2Vec framework. To do that, we publish the notebooks, code of Snomed2Vec and a public access web to text, improve and evaluate our tool. We publish access and repositories https://github.com/NachusS/Snomed2Vec

Other validation was the use of a corpus gold data, of Spanish emergency diagnostic from the clinical reports, that prepare two expert in codification. They use the browser search at ihtsdotools.org to check.

We have used precision, recall and F-measures to analyze the performance of the tool: Precision: P = TP/(TP+FP), Recall: R =TP/(TP+FN), where TP = true positive, TN=true negative, FN =False negative and FP= false positive To identify the $F_{Measure}$, we use the same representation like [7].

$$F_{Measure} = (1+\beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$$

The f-Measure score can be interpreted as a weighted average of precision and recall. We establish the same score that [7], β=0.7 in order to put more emphasis on precision than recall. We consider that the precision is more important for an automatic mapping tool:

TABLE I. TABLE MEASURES

Concept	Our Approach
Precision	0.8097
Recall	0.7469
$F_{Measure}$	0.7879

VI. CONCLUSION AND OUTLOOK

In this paper we have proposed a ontology-based named entity recognition to get the most similarity concepts from the Ontology, Snomed-CT. Our approach is to use word embedding to represent the descriptions terms of the Snomed Concepts, like a vector space model. We use Word2Vec framework to train two kind of big corpus, one with a local domain (emergencies discharges clinical reports) and other with a general domain (Spanish Wikipedia dump articles). In

this new model Snomed2Vec, we have representing all the Snomed concepts with its descriptions, top level hierarchy and the vector of the description term. We implement methods to apply the cosine distance, and get the collections of most similar vectors, between a vector of a query text, and all the vectors from the space model. Adapting the function implemented by genism library, most_similar() to our new Snomed2Vec Space model. With this process we suggest to the human specialist, a set of most similar concept, in order to generate a help for the codification of clinical reports.

In future This framework, can be very useful, to get the automatic binding to get the semantic clinical code, in the normalized clinical history like ISO 13606 or HL7 CDA Ver3.

In the Hospital Information System, all the diagnostic, procedures of the clinical reports, must be codify with specifics clinical terminology like ICD-10-MC, LOINC or Snomed-CT, with our help tools approach Snomed2Vec, the specialist in clinical codification can choose the correct concept in Snomed-CT, and map to the other terminologies. Other new approach is integrate this tools to the process of acquisition data in the Hospital Information System Forms, and help to a semiautomatic codification in live time, for the physicians.

REFERENCES

[1] Côté RA, Robboy S. Progress in Medical Information Management: Systematized Nomenclature of Medicine (SNOMED). JAMA. 243(8):756–762; 1980.

[2] El-Sappagh, Shaker H. Ali, Francesco Franda, Farman Ali and Kyung Sup Kwak. "SNOMED CT standard ontology based on the ontology for general medical science." BMC Med. Inf. & Decision Making (2018).

[3] IHTSDO (International Health Terminology Standards Development Organization), SNOMED CT Compositional Grammar Specification and Guide. https://confluence.ihtsdotools.org/display/DOCSCG/Compositional+Grammar+-+Specification+and+Guide. Last seen 10 Feb. 2019

[4] F. Z. Smaili, X. Gao, and R. Hoehndorf, "OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity based prediction," Bioinformatics, Nov 2018.

[5] Fatima Zohra Smaili, Xin Gao, Robert Hoehndorf; Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations, Bioinformatics, Volume 34, Issue 13, 1 July 2018.

[6] http://bio2vec.net/data/pubmed_model/ Last access: 10 February 2019.

[7] Allones J. L., . Martinez D., Taboada M. Automated Mapping of Clinical Terms into SNOMED-CT. An Application to Codify Procedures in Pathology. Journal of Medical Systems, 2014.

[8] Zhang, Yuanzhe, Xuepeng Wang, Siwei Lai, Shizhu He, Kang Liu, Jian Zhao and Xueqiang Lv. "Ontology Matching with Word Embeddings." CCL (2014).

[9] Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering pp. 158–176 (2013).

[10] Stoilos, G., Stamou, G., Kollias, S.: A string metric for ontology alignment. In: The Semantic Web–ISWC 2005, pp. 624–637. (2005).

[11] Li, J., Tang, J., Li, Y., Luo, Q.: Rimom: A dynamic multistrategy ontology alignment framework. Knowledge and Data Engineering, IEEE Transactions on 21(8), 1218–1232 (2009).

[12] Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. Web Semantics: Science, Services and Agents on theWorldWideWeb 7(3), 235–251.(2009).

[13] Cruz, I.F., Antonelli, F.P., Stroe, C.: Agreementmaker: efficient matching for large real-world schemas and ontologies. Proceedings of the VLDB Endowment 2(2), 1586–1589 (2009).

[14] Zhang Y. et al. Ontology Matching with Word Embeddings. In: Sun M., Liu Y., Zhao J. (eds) Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. NLP-NABD 2014.

[15] Donnelly K. Snomed-ct: The advanced terminology and coding system for ehealth. Stud Health Technol Inform. 2006.

[16] https://confluence.ihtsdotools.org/display/DOCSTART/6.+SNOMED+CT+Concept+Model. Last access: 10 February 2019.

[17] Martinez I., Castro JL: DNER Clinical (named entity recognition) from free clinical text to Snomed-CT concept. WSEAS Transactions on Computers, 2017.

[18] Martinez I., Castro JL: STMC: Semantic Tag Medical Concept Using Word2Vec Representa. 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). Karsltad (Sweden).2018.

[19] WikiExtractor. http://medialab.di.unipi.it/wiki/Wikipedia_Extractor. Last access: 11 Feb 2019.

[20] Cristian Cardellino: Spanish Billion Words Corpus and Embeddings (March 2016), https://crscardellino.github.io/SBWCE/

[21] Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilisticlanguage model. The Journal of Machine Learning Research. 2003.

[22] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of Workshop at the International Conference on Learning Representations, 2013.

[23] https://www.nltk.org/ Last access: 12 Febrary 2019.

[24] https://confluence.ihtsdotools.org/display/DOCTSG/7.5.2+Transitive+closure+implementation. Last access: 12 Febrary 2019.

[25] https://confluence.ihtsdotools.org/display/DOC/Technical+Resources. Last access: 12 Febrary 2019.

[26] Řehuřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks,45–50.

[27] L.J.P. van der Maaten and G.E. Hinton.Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

[28] G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. In Advances in Neural Information Processing Systems, volume 15, pages 833–840, Cambridge, MA, USA, 2002.

[29] Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR 12.

[30] http://projector.tensorflow.org/ Last access: 11 Febrero 2019.

[31] https://github.com/dominiek/word2vec-explorer/ Last access: 2019

[32] Friedman C, "Towards a comprehensive medical language processing system: methods and issues," Proceedings of the AMIA annual fall symposium: American Medical Informatics Association, 1997.

[33] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. May 2010;

[34] Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A, 3rd. The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annu Symp Proc. 2003:195-199.

[35] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al., "Mayo clinical Text Analysis and Knowledge ExtractionSystem (cTAKES): architecture, component evaluation and applications," Journal of the American Medical Informatics Association, 2010;17:507-13.

[36] Menasalvas E., Rodriguez-Gonzalez A., Costumero R., Ambit H., Gonzalo C. (2016) Clinical Narrative Analytics Challenges. In: Flores V. et al. (eds) Rough Sets. IJCRS 2016.vol 9920. Springer.

[37] K. Gojenola, M.Oronoz, A. Pérez, A. Casillas. IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts", Proceedings of the 8th International Workshop on Semantic Evaluation , Dublin, Ireland, 2014