Journal of King Saud University – Computer and Information Sciences (2017) xxx, xxx–xxx

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Arabic medical entity tagging using distant learning in a Multilingual Framework

Viviana Cotik [a], **Horacio Rodríguez** [b,*], **Jorge Vivaldi** [c]

[a] *Universidad de Buenos Aires, Buenos Aires, Argentina*
[b] *Polytechnical University of Catalonia, Barcelona, Spain*
[c] *Universitat Pompeu Fabra, Roc Boronat 132, Barcelona, Spain*

Received 22 April 2016; revised 1 August 2016; accepted 13 October 2016

KEYWORDS

Semantic Tagging;
Multilingual;
Medical domain;
Arabic Natural Language Processing

Abstract A semantic tagger aiming to detect relevant entities in Arabic medical documents and tagging them with their appropriate semantic class is presented. The system takes profit of a Multilingual Framework covering four languages (Arabic, English, French, and Spanish), in a way that resources available for each language can be used to improve the results of the others, this is specially important for less resourced languages as Arabic. The approach has been evaluated against Wikipedia pages of the four languages belonging to the medical domain. The core of the system is the definition of a base tagset consisting of the three most represented classes in *SNOMED-CT* taxonomy and the learning of a binary classifier for each semantic category in the tagset and each language, using a distant learning approach over three widely used knowledge resources, namely *Wikipedia*, *Dbpedia*, and *SNOMED-CT*.

1. Introduction and motivation

Semantic Tagging, a *Natural Language Processing (NLP)* task that has attracted recently the interest of many *NLP* researchers, can be defined as the task of assigning to some linguistic units occurring within a document a unique semantic tag chosen from a predefined tagset.

There is a wide agreement on approaching some *NLP* tasks and applications at the semantic level, but there is also agreement on considering that the current state of the technology does not allow a full accurate semantic parsing of text of an unrestricted domain. So, most systems restrict themselves to partial semantic interpretation, at lexical level (*Semantic Tagging*), or clause level (Semantic Role Labelling). *Semantic Tagging* is, so, a crucial task, *per se*, or as a necessary component of Semantic Interpretation Systems.

After this introduction, the organization of the article is as follows: In Section 2 we sketch the most basics characteristics as well as the state of the art of the *Semantic Tagging* approaches. Section 3 presents the methodology followed. The experimental framework is described in Section 4. Results are shown and discussed in Section 5. Finally, Section 6 presents our conclusions and further work proposals.

2. Related work

Semantic Tagging is a difficult task whose key elements are the following:

* Corresponding author.
E-mail addresses: vcotik@dc.uba.ar (V. Cotik), horacio@cs.upc.edu (H. Rodríguez), jorge.vivaldi@upf.edu (J. Vivaldi).

http://dx.doi.org/10.1016/j.jksuci.2016.10.004

(i) *the genre* of the document to be processed. The terminology used in documents belonging to different domains differs heavily, but even within a specific domain, terminology, general wording and sub-languages present very different characteristics. Consider, for instance, within the medical domain, genres like scientific literature, drug description outlets, medical report discharges, clinical proofs results, social media comments about diseases, and drugs and their efficiency. The characteristics of the wording used in these genres are highly diverse. We focus in this article on *WP* pages, and the evaluation is made on this kind of document.

(ii) *the linguistic units* to be tagged. There are two commonly followed approaches. Those that tag the entities occurring in the text and those that tag the mentions of these entities. Frequently, entities are represented by co-reference chains of mentions. Consider the following example: "*Asthma* is thought to be caused by... *Its* diagnosis is usually based on ... *The disease* is clinically classified ...". In these sentences there is an entity (*asthma*) referred three times, and, thus, forms a co-reference chain of three mentions. In our work, units to be tagged are terminological string found in *WP*. So, the linguistic units are phrases filtered by termhood conditions, i.e. only POS sequences corresponding to valid terms are allowed. These sequences are language dependent and correspond to basic (non recursive) noun phrases headed by a noun.

(iii) *the tagset*. Frequently the tagset is really a set of categories with no explicit relations between them. A crucial point is its granularity (or size). The spectrum of tagset sizes is immense. Fine-grained tagsets can consist of thousands (as is the case of *WordNet* synsets) or even millions (as is the case of *WP* pages) of tags. Coarse-grained tagsets contains only a few tags. In our case, we have used a tagset consisting of only three tags. Details of the selection are given in Section 3.2.

Regarding the resources used for the task, curated resources (terminologies, lexica, ontologies, etc.), such as Classification of Diseases and Related Health Problems (*ICD-9*, *ICD-10*),[1] Medical Subject Headings (*MeSH*).[2] nomenclature of drugs and their brands in *DrugBank*[3] and *Gray 's Anatomy*,[4] etc, are available and are widely used for this task. However, using these resources is not straightforward. Some of the terms allow multiple variations[5] (not all of them are collected in the resources) while others (specially the most frequent) are highly ambiguous.[6] Besides, recognizing and classifying the mentions in documents is highly challenging.

An important source of information for tasks similar to ours are the proceedings of the *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text* (Uzuner et al., 2010). Within this contest, Yeganova et al. (2012) uses two rather simple statistical approaches obtaining good results. Halgrim et al. (2011) and Uzuner et al. (2010) apply a cascade of classifiers for extracting medication information from discharge summaries. Our previous work, Vivaldi and Rodríguez (2015), is similar to that presented here but limited to English. Another source of information is the *DDI Extraction 2013* (task 9 of Semeval-2013, Segura-Bedmar et al., 2014). Focusing on a narrower domain, Drug-Drug interaction, the shared task included two challenges: (i) Recognition and Classification of Pharmacological substances, and (ii) Extraction and classification of Drug-Drug interactions. The former is clearly a case of *Semantic Tagging*, in this case reduced to looking for mentions of drugs within biomedical texts, but with a finer granularity of the tagset, It included *drug*, *brand*, *group* (group of drug names) and *drug-n* (active substances not yet approved for human use).

Regarding the techniques involved, many approaches have been proposed for dealing with *Semantic Tagging*, such as rule-based methods and supervised *Machine Learning* (*ML*). A common limitation is the dependence on a narrow domain/-genre/tagset/language making its adaptation to other settings highly difficult (and costly). We faced the adaptation issue by:

- Using a multilingual setting in which the process in one language can help the process in other (usually less resourced) languages.
- Using a set of wide-coverage domain-free resources for learning and using a low cost learning method, *distant learning*. Specifically we include as resources: *SNOMED-CT*,[7] that is restricted to the medical domain, and two widely used domain-independent encyclopaedical ones: *WP* pages (including data obtained from Infoboxes) and categories, and *DBP*.[8]

Related to *Semantic Tagging*, the first faced problem and the one that has attracted more attention is *Word Sense Disambiguation*. In Agirre and Edmonds (2006) and Navigli (2009) we can find two excellent surveys on this issue. A more recent survey, covering many *Semantic Tagging* techniques and comparing them, can be found in Gerber et al. (2011). A unified framework including *Word Sense Disambiguation* and *Entity Linking* is presented in Moro et al. (2014). Wikifiers[9] proceed mostly in two steps: candidate detection and classification/ranking. See Roth et al. (2014) for a recent, excellent and comprehensive analysis. Closely related to wikification is the task of *Entity Linking*. This task has got an explosive development starting with the *Entity Linking* challenge within the *TAC KBP* framework,[10] from 2010. Overviews of the contests are the main sources of information: Ji et al. (2010), Ji et al. (2011), James Mayfield and Artiles (2012), and Mayfield et al. (2013).

English is, by far, the most supported language for biomedical resources. The National Library of Medicine (NLM)

[1] http://www.who.int/classifications/icd/en/.
[2] http://www.ncbi.nlm.nih.gov/mesh.
[3] http://www.drugbank.ca/.
[4] http://www.bartleby.com/107/.
[5] See for example the English sets ["abdomen", "venter", "stomach", "belly"], ["fever", "pyrexia", "febris"] or the set of acronyms ["ADE", "ADR", "DAR"] all sharing the meaning of "adverse drug reaction" but using different wordings.
[6] Acronyms is a major source of ambiguity as for example: "MI" in English can be a synonym of "myocardical infartion", "mitral insufficiency" or "mental Illness" while in Spanish it may refer to "metabolic index", "mesenteric ischemia" or "menstruation induction".
[7] https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html.
[8] http://wiki.dbpedia.org/.
[9] Wikifiers are programs that provide texts with content enrichment by displaying information from *WP*.
[10] http://www.nist.gov/tac/2014/KBP/.

https://www.nlm.nih.gov/ maintains the Unified Medical Language System (UMLS)[11] that groups an important set of resources to facilitate computer systems to "understand" the meaning of the language of biomedicine and health. Only a small fraction of such resources are available for languages other than English. A relevant aspect of information extraction in medicine is the recognition and identification of biomedical entities (like *disease*, *genes*, *proteins* ...). Several named entity recognition (NER) techniques have been proposed to recognize such entities based on their morphosyntactical pattern and context. NER can be used to recognize previously known names and also new names, but cannot be directly used to relate these names to specific biomedical entities found in external databases. For this identification task, a dictionary approach is necessary. A problem is that existing dictionaries often are incomplete and different variations may be found in the literature; therefore it is necessary to minimize this issue as much as possible. Regarding Arabic NER, a good reference is Benajiba et al. (2010). There are a number of tools that take profit of the UMLS resources. Some of the more relevant are *Metamap* (Aronson and Lang, 2010) and *Whatizit* (Rebholz-Schuhmann et al., 2008). Cotik et al. (2015) uses RadLex to detect concepts in radiology reports written in Spanish.

3. Methodology

3.1. Outline

This paper, as most of the systems showed in Section 2 proposes a *ML* solution to a tagging task. Therefore, it requires two main steps: training and annotation (see Fig. 1). The main drawback of this type of solutions is the dependency on annotated documents, which usually are hard to obtain. Our main target in this research is to train a classifier minimizing the impact of this issue and keeping good results.

For such a purpose we use as learning examples -within the *distant learning* paradigm- a set of seed words obtained with a minimal human supervision. As mentioned in Section 1, we use domain-independent Knowledge Sources as *WP*, and *DBP*. The reasons for this choice are: (i) they provide good interlingual linking and (ii) although domain-independent, they provide a nice coverage of the medical domain, including links to codified datasets. Besides, we include *SNOMED-CT* because of (i) its rich coverage of a high variety of medical entities and (ii) a well-founded taxonomic class organization.

The overall architecture of our system is shown in Fig. 1.

The process of *Semantic Tagging* is carried out by a module shown in the bottom of the figure. The process consists of the performance of a set of binary classifiers (one for each class in the tagset and for each language) followed by meta-classifiers (one for each language) that combines the results of the binary classifiers.

The training of the binary classifiers is performed using a *distant learning* approach from the three *Knowledge Sources*.

For English all the processes are easier because of the direct availability of all the *Knowledge Sources*. For other languages the process is more complex due to the limitation of *Knowledge Sources* (some languages, specially Arabic, lack some of the resources or have smaller coverage). In such cases, We performed cross-lingual mappings. The results of different learning processes clearly depend on the size and quality of the training material.

Besides the initial assignment of seed terms corresponding to *WP* categories to *SNOMED-CT* classes no manual intervention was needed. It is worth noting that, only seed terms that have associated *WP* pages are considered. The results, so, are sets of *WP* pages to be used for learning the classifiers.

Fig. 2 depicts an overall view of the learning components (occurring in the top of Fig. 1). As can be seen, the system proceeds in three steps: (i) building the base tagset and the set of relevant *WP* categories, this process is further detailed in Sections 3.2 and 3.3, (ii) selecting the seedterms for learning, expanded in Fig. 3 and explained in detail in Section 3.4, and (iii) Learning the binary classifiers and the metaclassifiers. The classification component of the system (bottom of Fig. 1) is expanded in Fig. 4.

3.2. Selection of the tagset

Our tagset consists of the three most populated categories from the 19 top categories in the class structure of *SNOMED-CT*. In the rest of the article we refer to these categories as *BP* (Body Part), *DRUG*, and *DISEASE*. We have used *SNOMED-CT* for English (although there exist, too, a partial version for Spanish and a proprietary version for French). Using BioPortal SPARQL endpoint[12] we have extracted the top categories, and from them the set of terms under each one. For all the languages we have collected the set of translations (using *DBP*) and we have filtered out the terms not existing in the corresponding *WP* (as page or as category). We have selected for our experiments the three categories having a higher coverage considering all the languages.

3.3. Defining the initial set of relevant WP categories

Although our *distant learning* approach for obtaining additional training material is based on three *Knowledge Sources*, (*WP*, *DBP*, and *SNOMED-CT*) using, when needed, their interlingual capabilities, a previous step, limited to the English *WP* has to be carried out and its results are used for processing the other *Knowledge Source*.

Following the approach described in Vivaldi and Rodríguez (2010), that automatically extracts scored lists of terms from both *WP* pages titles and *WP* categories titles, we got the set of the most reliable *WP* categories.[13] This resulted on a set of 239 *WP* categories. We manually assigned to such categories a unique *SNOMED-CT* class. Let us denote $Cats^{wp.en}_{class}$ $\left(Cats^{wp.en}_{BP}, Cats^{wp.en}_{DRUG}, \text{and } Cats^{wp.en}_{DISEASE}\right)$ these sets. We take profit of the graph structure of *WP*. *WP* consists basically of two graphs, the page graph and the category graph. In the former the nodes are *WP* pages while in the later the nodes are *WP* categories. Edges consist of *WP* links. We consider

[11] https://www.nlm.nih.gov/research/umls/.

[12] http://sparql.bioontology.org/.

[13] See Vivaldi and Rodríguez (2010) for details about the way of obtaining such categories from *WP* resources. The system provides terms corresponding to both *WP* pages and categories, but we use here only the later.

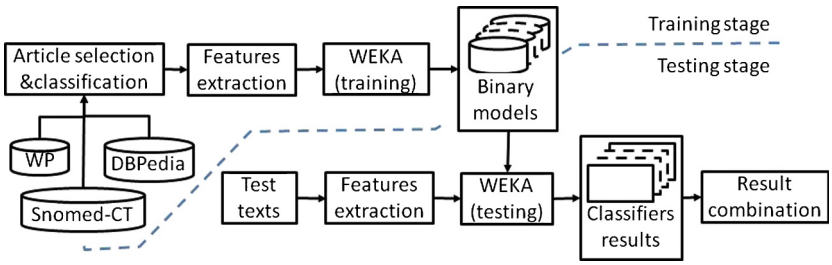

Figure 1 Training and testing pipelines.

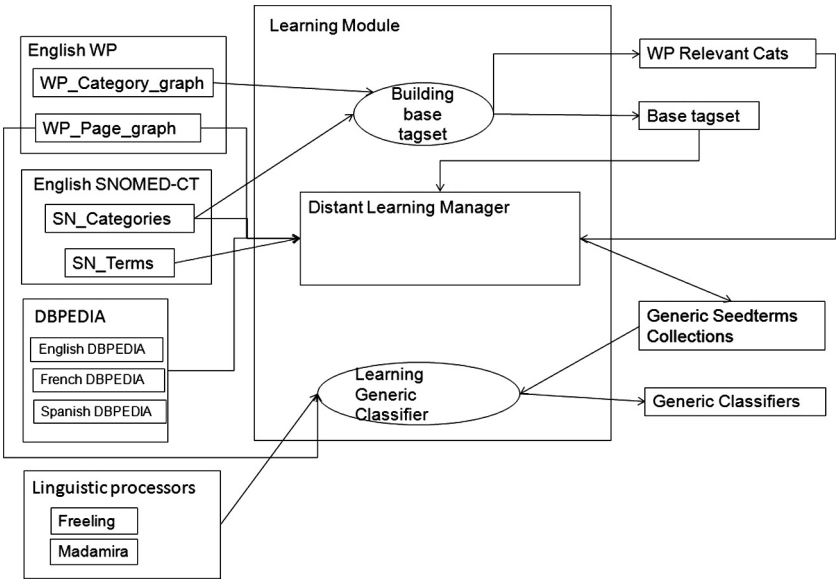

Figure 2 Learning module.

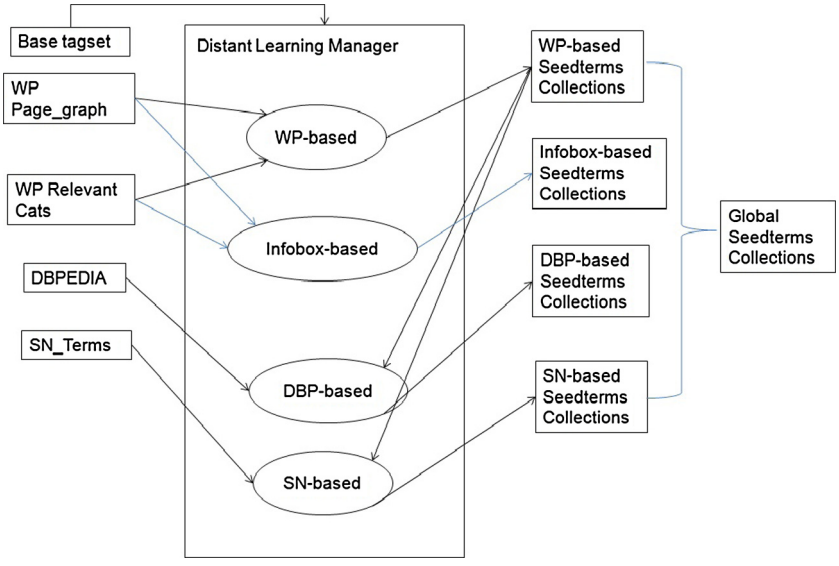

Figure 3 Distant learning manager.

three types of links: page-category (categories to which one page belongs), category-page (pages corresponding to a given category) and category-category (super and sub categories of a given one). We compute the score of a page from the scores of the categories it belongs to, and the score of a category from the scores of the pages belonging to it). In this way, using an iterative procedure, good pages reinforce their categories and good categories their pages.

Please cite this article in press as: Cotik, V. et al., Arabic medical entity tagging using distant learning in a Multilingual Framework . Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2016.10.004

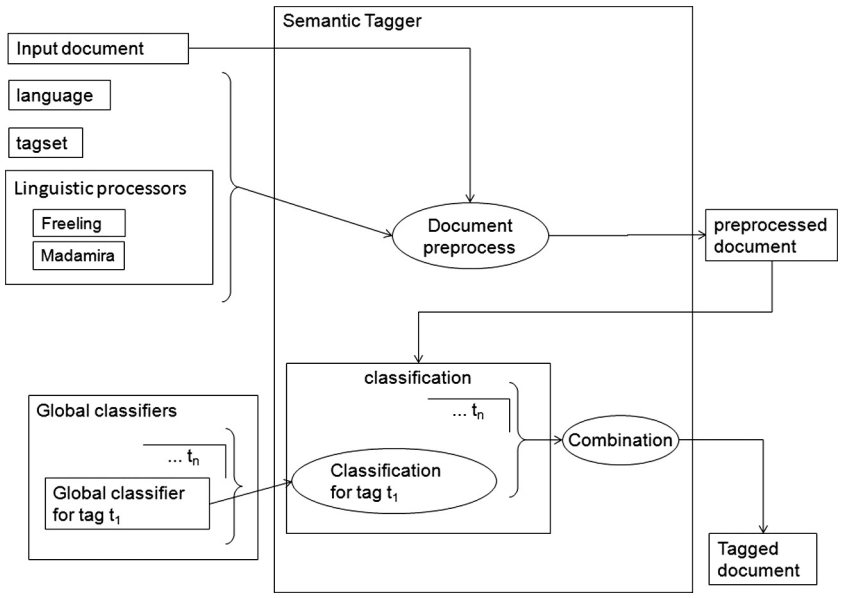

Fig. 4 Semantic tagging component.

3.4. Obtaining the seed terms for each tag and language

To obtain the seed terms needed for learning the classifiers, we proceed in four ways, using our three *Knowledge Sources* (see He et al. (2011) and Yeganova et al. (2012) for analysis of these and other resources used for similar purposes). The process is shown in Fig. 3. The following *Knowledge Sources* have been used:

- *WP*, although being a general purpose resource, it densely covers the medical domain. English *WP* contains terminological units from multiple medical thesauri and ontologies, as pointed out above. The current full coverage of the four *WP* used in our research are[14]: 5,093,100 pages for the English *WP*, 410,657 pages for the Arabic *WP*, 1,730,505 pages for the French *WP*, and 1,198,094 pages for the Spanish *WP*.
- *DBP* is one of the central linked data dataset in Linked Open Data (*LOD*). It currently contains more than 3.5 million things, and 1 billion *RDF* (Resource Description Framework) triples with a nice coverage of the medical domain. Unfortunately there is not *DBP* for Arabic and, so, for this language instances from this source have to be collected indirectly (through the existence of Arabic labels, i.e. labels consisting of a string suffixed with @ar, attached to *DBP* resources in the available *DBPs*).
- *SNOMED-CT*, with more than 350,000 concepts, 950,000 English descriptions (concept names) and 1,300,000 relationships is the largest single vocabulary ever integrated into *UMLS*.[15] The basic *SNOMED-CT* source, and the only used in our work, is in English.

Our system is based on Vivaldi and Rodríguez (2010) and Vivaldi and Rodríguez (2015) extending the later reference, that was restricted to English, for working in a multilingual setting. Our aim is to collect medical terminologies for the semantic classes and languages involved. The process is as follows:

1. As said in Section 3.3, for each semantic class a set $Cats^{wp.en}_{class}$ of relevant categories of English *WP* has been collected.
2. From each of these $Cats^{wp.en}_{class}$ sets we obtain the set of *WP* pages and we remove the pages corresponding to more than one set. Werefer to these sets as $Pages^{wp.en}_{class}$ ($Pages^{wp.en}_{BP}$, $Pages^{wp.en}_{DRUG}$, and $Pages^{wp.en}_{DISEASE}$). These three sets are our first collection of domain terms.
3. From each of the three $Pages^{wp.en}_{class}$ collections we collect all the *WP* infoboxes and infobox slots.[16] We manually selected the pairs ⟨infobox, slot⟩: specific for the corresponding class. We then collected all the pages owning any of these pairs, resulting on the second collection of domain terms, $Pages^{infobox.en}_{class}$.
4. The third collection of domain terms was obtained from *SNOMED-CT*. We selected the set of terms under the three top classes of *SNOMED-CT* class structure, *Clinical Finding/Disorder*, *Body_structure*, and *Pharmaceutical/biological product* that can be mapped into our own classes. From the set retrieved from *SNOMED-CT* only the terms existing in *WP* have been collected, resulting in $Pages^{sn.en}_{class}$.[17]

[14] see updated statistics at https://stats.wikimedia.org/EN/Sitemap.htm.
[15] http://www.nlm.nih.gov/research/umls/.

[16] Although *WP* pages consist basically of free text, some pages include, too, structured information. The most popular way of composing and including this kind of information is using predefined templates attached to some categories. These structures are named infoboxes and their items infobox slots. For instance, the disease infobox contains slots for ICD-9 and ICD-10 codes, MeSH entries, UMLS CUI, etc.
[17] It is worth noting that although terms coming from *SNOMED-CT* not existing in *WP* are filtered out, some of the remaining terms could be new (not detected previously by the other methods) because the way of selection is different.

5. The last source of classified medical terms (and the most productive) is *DBP*. For accessing *DBP* data we used the *DBpedia Sparql endpoint*[18] that allows an easy way of building the queries and an efficient way of accessing the data. Using as seed terms the members of $Pages^{wp.en}_{class}$ we collected the most useful predicates (balancing their coverage and specificity) and obtained the set of subjects in the *rdf* triples involving such predicates. In this way we collected our fourth set of medical terms, $Pages^{dbp.en}_{class}$. See Vivaldi and Rodríguez (2015) for details of this process.
6. We then get the union of the four datasets procured in previous steps. We have, so, the three sets $Pages^{all.en}_{class}$. For each page we computed a *purity score*, i.e. a score ranging in [0, 1] measuring the confidence of the page belonging to its corresponding *SNOMED-CT* class. Specifically we define a purity measure of a page as the inverse of the number of semantic tags to which their categories belong. So, as in our work we use three semantic tags, the purity ranges from 1/3 to 1. If all the WP categories are mapped to a unique semantic tag the purity is 1.
7. From the sets $Pages^{all.en}_{class}$ and the use of *English DBpedia* labels we obtained the corresponding translated terms, if existing, to Arabic, French, and Spanish, $Pages^{all.ar}_{class}$, $Pages^{all.fr}_{class}$, and $Pages^{all.sp}_{class}$.
8. Using the *English DBP*, *French DBP*, and *Spanish DBP* labels, we enriched the corresponding sets in the other languages, including Arabic. The way of enriching a set of terms for a target language comes from the presence in a *DBP* resource for another source language of a label for the target. As there is no *DBP* for Arabic, this language cannot contribute to enriching the others but only takes profit of the other languages' enrichment.
9. We iterate the two last steps until no more terms are found. The final figures for the four languages and three semantic classes can be found in Table 1.

Not all the methods for selecting the seedwords perform equally for the different semantic tags and languages. It is worth noting that more than a half of the seedwords used for learning have been selected by the *DBP* source. For instance, for Arabic and for the tag *BP* about 60% of the seedterms (1077) come from DBP (657 against 420). As the seedwords sizes for Arabic are smaller, not using *DBP* seeds can result on small datasets (taking into account, too, that some of the *WP* pages are filtered out, and that only pages with purity 1 are used). So although probably for English we have enough training material without using *DBP*, for the other languages, specially for Arabic, *DBP* data have to be used.

3.5. Learning the classifiers

Following Huang and Riloff (2010), for each semantic class and language we generate training instances by automatically labelling each mention of a seed term with its designated semantic class. The core idea of our approach is that for each seed term *t* of a class *tag*, all the mentions of *t* in its *WP* page can be considered positive examples for learning the class *tag*. For each mention we create feature vectors for the classifiers, the seeds themselves are hidden and only contextual features are used to represent each training instance. Proceeding in this way the classifier is forced to generalize with limited overfitting.

We created a suite of binary contextual classifiers, one for each semantic class and language. The classifiers are learned using, as in Huang and Riloff (2010), SVM models using Weka toolkit (Hall et al., 2009). Each classifier makes a scored decision as to whether a term belongs or not to its semantic class. Examples for learning correspond to the mentions of the seed terms in the corresponding *WP* pages. Let $x_1, x_2, \ldots, x_n$ be the seed terms for the semantic class *tag*. For each x_i we obtain its *WP* page and we extract all the mentions of seed terms occurring in the page. Positive examples correspond to mentions of seed terms corresponding to semantic class *tag* while negative examples correspond to seed terms from other semantic classes. Frequently, a positive example occurs within the text of the page but often many other positive and negative examples occur as well. We have analyzed the average distribution of positive and negative terms for all the languages and semantic tags. The results are depicted in Table 2. As can be seen, for most languages and classes the number of examples (positive and negative) for training is high and well balanced. Features are simply words occurring in the local context of mentions.

The corpus of each semantic class and language is divided into training and test sections. For processing the full corpus we use a linguistic processor to identify content words in each sentence and create feature vectors that represent each constituent in the sentence. For each example, the feature vector captures a context window of *n* words to its left and right[19] without surpassing sentence limits. The linguistic processing includes sentence splitting, tokenizing, POS tagging, and Named Entity Recognition. For English, French, and Spanish *Freeling* toolbox[20] (Padró et al., 2012) has been used to perform this task. For Arabic we have used *Madamira*[21] (Pasha et al., 2014).

For evaluation we used *WP categories - SNOMED-CT classes* manually annotated mappings as gold standard as explained in Section 3.3. We considered for each semantic class *tag* a gold standard set including all the *WP* pages with purity 1, i.e. those pages unambiguously mapped to *tag*. The accuracy of the corresponding classifier is measured against this gold standard set.

We proceed with the sets of seed terms (one set for each semantic class and language) collected as described in Section 3.4, i.e. $Pages^{wp.l}_{tag}$, for $tag \in \{BP, DRUG, DISEASE\}$ and $l \in \{ar, en, fr, en\}$. Some of the *WP* pages corresponding to the selected terms are removed due to: (i) having less than 100 words, (ii) difficulties in extracting useful plain text (pages consisting mainly of itemized lists, formulas, links, and so) and (iii) having purity lower than 1.

The whole set of seed terms for every category *tag* (see Table 1) was split in two sections: training and test. Each section has been limited to 500 WP pages. The whole set of training documents was used, regardless the origin of its members, although, obviously, most of the mentions of a seed term occur within the documents associated to its origin. For learning a binary classifier for a semantic class *tag*, all the mentions of

[18] http://dbpedia.org/sparql.

[19] In the experiments reported here *n* was set to 3.
[20] http://nlp.lsi.upc.edu/freeling/.
[21] http://nlp.ldeo.columbia.edu/madamira/.

Table 1 Seedwords datasets sizes.

Medical category	English	Arabic	French	Spanish
BODY PART	6464	1077	2183	2663
DISEASE	14,033	1771	3957	3994
DRUG	13,520	1085	2651	2347

Table 2 Distribution of positive and negative seedwords. Pairs consist of the ratio of positive and negative counts per page.

Medical category	English	Arabic	French	Spanish	All
BODY PART	(1.95, 2.63)	(1.11, 1.32)	(0.73, 0.77)	(0.80, 0.86)	(1.36, 1.73)
DISEASE	(6.75, 4.10)	(4.07, 1.84)	(2.01, 0.87)	(2.10, 1.12)	(4.75, 2.74)
DRUG	(2.15, 3.07)	(0.78, 2.49)	(0.83, 1.33)	(1.34, 1.78)	(1.75, 2.19)
All	(4.00, 3.41)	(2.36, 1.88)	(1.34, 0.99)	(1.51, 1.21)	(2.94, 2.45)

Table 3 Results (F1) obtained for English with different sources and combination methods.

Origin of the seed terms	Best result	Using a meta-classifier
Wikipedia	87,4	89,6
SNOMED	87,4	88,8
DBpedia	94,0	94,9
Overlall	94,0	94,9

all the seed terms of *tag* within all the training documents are triggers for a positive example while all the mentions of all the seed terms of *tag'*, for $tag \neq tag'$, are triggers of negative examples. Following Huang and Riloff (2010), each example is represented as a n-dimension binary vector where dimensions correspond to lemmas of content words occurring in the context of each trigger. Contexts correspond to windows (limited to size 3) of words surrounding the mentions without surpassing sentence limits.

4. Experimental framework

We have applied the method described here to the three semantic categories and four languages. Let *l* be the language and *tag* the semantic category. For each seed term *tag* in $Pages_{tag}^{wp.l}$ we obtain its corresponding *WP* page and, after cleaning, POS tagging, and sentence segmenting, we extracted all the mentions (compliant with termhood condition). For each mention the vector of features is built and the three learned binary classifiers corresponding to *l* are applied to it. If none of the classifiers classifies the instance as belonging to the corresponding semantic class no answer is returned. If only one of the classifiers classifies positively the instance, the corresponding class is returned. Otherwise a combination step has to be carried out. For combining the results of the binary classifiers two methods have been implemented:

- *Best Result*. This method returns the class of the best scored individual result of the binary classifiers.
- *Meta-classifier*. A SVM multiclass classifier is trained using as features the results of the basic binary classifiers together with the context data already used in the basic classifiers. The resulting class is returned.

5. Results

Table 3 depicts the global results obtained for English when applying both combination methods for the three knowledge sources extracted.[22] As it can be seen, using the meta-classifier slightly outperforms the best score method. Using *DB* as source of seed words consistently outperforms the other sources. For the other languages we have used for learning the union of the seedwords coming from all the resources.

The global results are presented in Table 4. As can be observed, there is a severe drop in accuracy for languages other than English. The reasons could be due to:

- The differences in size of the training material as shown in Table 1.
- The differences in *WP* coverage pointed out is Section 3.4.
- The degradation of data quality resulting from the cross-lingual mappings that never can be considered error free.
- The differences in accuracy of the linguistic processors for the different languages involved.

It is worth noting that Arabic results slightly outperforms Spanish ones despite the lower size of the training material and *WP* coverage. This is due to the excellent performance of Arabic classifier for *DISEASE* class probably due to the quality of these annotations in *WP* medical pages.

It is difficult to compare our results with other state-of-the-art systems performing the same task because of the lack of gold standard datasets and the differences on used tagsets and languages. To our knowledge, *WP* pages have not been

[22] Reported values are an average over the results for each *SNOMED-CT* class. Actual values, for the case of *SNOMED-CT* only seed terms, range among 73.0–94.8 (precision) and 67.1–93.6 (recall).

Table 4 Global results.

Semantic class	English	Arabic	French	Spanish
BP	0.93	0.35	0.93	0.24
DISEASE	0.95	0.78	0.51	0.64
DRUG	0.71	0.33	0.51	0.54
All	0.94	0.54	0.75	0.53

used previously as gold standard for this task. A shallow comparison could be carried out for English with the Concept Extraction task of the *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text*, Uzuner et al. (2010) and with the *DDI Extraction 2013* (task 9 of Semeval-2013, Segura-Bedmar et al. (2014), both sketched in Section 2. This informal comparison is just for seeing whether our results can be placed within the state-of-the-art ranges for similar tasks. In the case of *2010 i2b2/VA* the results of the three best scored systems range from 0.78 to 0.85. Our results (0.94) clearly outperform these ones. In the case of *DDI* for entity tagging, the figures of the three best scored systems range from.51 to 0.83 for DrugBank data and from 0.37 to 0.56 for Medline data, closer to our genre. In this case the comparison should be performed with our results on *DRUG* classifier (0.71). Once again our results seem to outperform the ones obtained in this contest. Although to be fair, and lacking a direct comparison, we simply can say that our results can be considered state-of-the-art. For French also a shallow comparison could be made with the CLEF eHealth2015[23] contest, task 1b (Clinical Named Entity Recognition) (Neveol et al., 2015). In this case the results range from 0.70 to 0.76 while ours were of 0.75. To be fair we should point out that the task in this case was clearly more challenging than our (it consisted on 10 *UMLS* categories detection and classification task). Unfortunately no way of evaluation even shallow can be made for Arabic and Spanish (the results of Cotik et al. (2015) are not comparable since the task is slightly different).

6. Conclusions and further work

We have presented a system that automatically detects and tags medical terms that correspond to *WP* pages found in *WP* pages. The tagset used, consisting of three categories, is derived from *SNOMED-CT* taxonomy. The system has been applied to four languages including Arabic. The results, although not directly comparable with other approaches, seem to reach at least state-of-the-art accuracy (compared with best systems in related contests). A relevant benefit of this approach is that the effort for obtaining positive/negative examples for training has been reduced to a minimum.

Some of the tools used in this experimentation are for general purpose. Their performance may not be appropriate for some medical terms (ex. *1,3-difluoro-2-propanol* or *8-cyclopentyl-1,3-dipropylxanthine*, among others) due to the intrinsic complexity of such terms and the difficulty in processing such terms with standard NLP tools. We plan to introduce some improvement in our tools or use already existing/available specialized tools, such as *Metamap* (Aronson and Lang, 2010).

Several lines of research will be followed in the next future.

- The main limitation of our system is that training and testing data of the *ML* algorithm is only based on *WP* pages. The use of data provided by the Concept Extraction task of the 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text will be considered for training and test sets. Our aim is to build a system robust enough to be applied to more challenging genres, as Electronic Health Reports.
- As our results are based on three knowledge sources, an obvious way of possible improvement is the combination and/or the specialization of the resources for learning more accurate classifiers. Specially extending the capabilities of *DBP* seems to be a good research direction.
- Using a finer grained tagset and including more challenging categories (as symptoms, clinical findings, procedures, impairments, ...).
- Moving from semantic tagging of medical entities to semantic tagging of relations between such entities is a highly exciting objective, in the line of recent challenges in the medical domain (and beyond).

Acknowledgements

This work was partially supported by the TUNER project (Spanish Ministerio de Economía y Competitividad, TIN2015-65308-C5-5-R) and the GRAPH-MED project (Spanish Ministerio de Economía y Competitividad, TIN2016-77820-C3-3-R).

References

Agirre, E., Edmonds, P., 2006. Word Sense Disambiguation: Algorithms and Applications. AAAI Workshop, Nancy Ide and Chris Welty.

Aronson, A.R., Lang, F.-M. 2010. An overview of MetaMap: historical perspective and recent advances. In: JAMIA, vol. 17, pp. 229–236.

Benajiba, Y, Zitouni, I., Diab, M., Rosso, P., 2010. Arabic named entity recognition: using features extracted from noisy data. In: Proceedings of 48th Annual Meeting of the Association for Computational Linguistics, ACL-2010, Uppsala, Sweden, July 11–16, pp. 281–285.

Cotik, V., Filippo, D., Castano, J. 2015. An approach for automatic classification of radiology reports in spanish 634–638.

Gerber, A., Gao, L., Hunte, J., 2011. A scoping study of (who, what, when, where) semantic tagging services. In: Research Report, eResearch Lab, The University of Queensland.

Halgrim, S., Xia, F., Solti, I., Cadag, E., Uzuner, O., 2011. A cascade of classifiers for extracting medication information from discharge summaries. J. Biomed. Semantics.

[23] https://sites.google.com/site/clefehealth2015/.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., 2009. The weka data mining software: an update. In: SIGKDD Explorations.

He, J., de Rijke, M., Sevenster, M., van Ommering, R., Qian, Y., 2011. Generating links to background knowledge: a case study using narrative radiology reports. In: Proceedings of CIKM'11, Glasgow, Scotland, UK.

Huang, R., Riloff, E., 2010. Inducing domain-specific semantic class taggers from(almost) nothing. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pp. 275–285.

James Mayfield, H.T.D., Artiles, Javier, 2012. Overview of the TAC 2012 knowledge base population track, Text Analysis Conference (TAC).

Ji, H., Grishman, R., Dang, H.T., Griffitt, K., Ellis, J., 2010. Overview of the TAC 2010 knowledge base population track, Text Analysis Conference (TAC).

Ji, H., Grishman, R., Dang, H.T., 2011. Overview of the TAC 2011 Knowledge Base Population Track, Text Analysis Conference (TAC).

Mayfield, J., Ellis, J., Getmana, J., Mott, J., Li, X., Griffitt, K., Strassel, S.M., Wright, J., 2013. Overview of the kbp 2013 entity linking track. In: Text Analysis Conference (TAC).

Moro, A., Roganato, A., Navigli, R., 2014. Entity linking meets word sense disambiguation: a unified approach. Trans. ACL, 231–244.

Navigli, R., 2009. Word sense disambiguation: a survey. ACM Comput. 41.

Neveol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., Zweigenbaum, P., 2015. Clef ehealth evaluation lab 2015 task 1b: clinical named entity recognition. In: Proceedings of CLEF.

Padró, L., Stanilovsky, E., 2012. Freeling 3.0: towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA, Istanbul, Turkey.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A.E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.M., 2014. Madamira: a fast, comprehensive tool for morphological analysis and disambiguation of arabic. In: Proceedings of LREC.

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Jimeno, A., 2008. Text processing through Web services: calling Whatizit. Bioinformatics Appl. Note 4/2, 296–298.

Roth, D., Ji, H., Chang, M.-W., Cassidy, T., 2014. Wikification and beyond: the challenges of entity and concept grounding. In: Tutorial at the 52nd Annual Meeting of the Association for Computational Linguistics.

Segura-Bedmar, I., Martnez, P., Zazo, M.H., 2014. Lessons learnt from the ddi extraction-2013 shared task. J. Biomed. Infor. ISSN: 1532–0464.

Uzuner, Özlem, South, B.R., Shen, S., DuVall, S.L., 2010. i2b2/va challenge on concepts, assertions, and relations in clinical text. In: J. Am. Med. Inform. Assoc., 18, pp. 552–556.

Uzuner, Özlem, Solti, Imre, Cadag, Eithon, 2010. Extracting medication information from clinical text. J. Am. Med. Inf. Assoc. 17, 514–518.

Vivaldi, J, Rodríguez, H., 2010. Using wikipedia for term extraction in the biomedical domain: first experience. In: Procesamiento del Lenguaje Natural, vol. 45, pp. 251–254.

Vivaldi, J., Rodríguez, H., 2015. Medical entities tagging using distant learning, Computational Linguistics and Intelligent Text Processing – 16th International Conference, CICLing 2015, Cairo, Egypt, April 14–20, 2015, Proceedings, Part II, pp. 631–642. http://dx.doi.org/10.1007/978-3-319-18117-247.

Yeganova, L., Kim, W., Comeau, D., Wilbur, W.J., 2012. Finding biomedical categories in medline. In: BMC Biomedical Semantics.

Please cite this article in press as: Cotik, V. et al., Arabic medical entity tagging using distant learning in a Multilingual Framework . Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2016.10.004