Semantic tagging of French medical entities using distant learning

Viviana Cotik[1], Horacio Rodríguez[2], and Jorge Vivaldi[3]

[1] Universidad de Buenos Aires, Buenos Aires, Argentina, vcotik@dc.uba.ar,
[2] University of Catalonia, Barcelona, Spain, horacio@lsi.upc.edu
[3] Universitat Pompeu Fabra, Roc Boronat 132, Barcelona, Spain, jorge.vivaldi@upf.edu

Abstract. In this paper we present a semantic tagger aiming to detect relevant entities in French medical documents and tagging them with their appropriate semantic class. These experiments has been carried out in the framework of CLEF2015 eHealth contest that proposes a tagset of ten classes from *UMLS* taxonomy. The system presented uses a set of binary classifiers, and a combination mechanisms for combining the results of the classifiers. Learning the classifiers is performed using two widely used knowledge source, one domain restricted and the other is a domain independent resource.

Keywords: Machine Learning, SNOMED CT, UMLS, Wikipedia, semantic tagger, binary classifiers, distant learning

1 Introduction

Recently, we [1] developed a semantic tagger for the medical domain performing on pages of the English Wikipedia[4] (*WP*) previously selected as belonging to the medical domain, using a distant learning approach. Our aim in this paper is exploring whether the approach can be applied to other language (French), other genre (EMEA and Medline documents) and other tagset. We performed these experiments within the framework of CLEF2015 eHealth contest [2], more specifically in Task 1b, Clinical Named ENtity Recognition [3].

Semantic Tagging (*ST*) can be defined as the task of assigning to some linguistic units of a text a unique tag from a semantic tagset. It can be divided in two subtasks: detection and tagging. The first one is similar to term detection and Named Entity Recognition (*NER*), while the latter is closely related to Named Entity Classification (*NEC*).

Other *Natural Language Processing* (*NLP*) tasks related to Semantic Tagging are *Word Sense Disambiguation* (*WSD*), aiming to tag each word in a document with its correct sense from a senses repository, and *Entity Linking* (*EL*), aiming to map mentions in a document to entries in a Knowledge Base.

The key elements of *Semantic Tagging* task are:

[4] http://en.wikipedia.org

i) *the document*, or document genre, to be processed. In this paper we focus on the medical domain and the genre of documents are those included in CLEF2015 contest, namely EMEA and Medline documents in French (see [4] for a description of the corpus).

ii) *the linguistic units* to be tagged. There are two commonly followed approaches, those that tag the entities occurring in the text, i.e. Entity Linking, as [5], and those that tag mentions of these entities, as [6]. Frequently, entities are represented by co-reference chains of mentions. Consider the following example (from the article "Asthma" in Wikipedia). "*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 the first approach, the entity (the whole set of three mentions) will be tagged as a disease, in the second one, which we follow in this work, each mention is detected and tagged independently, so only the first and last mentions are tagged as diseases. In this work, units to be tagged are terminological strings found in the source documents.

iii) *the tagset*. A crucial point is its granularity (or size). The spectrum of tagset sizes is immense. In one extreme of the spectrum, fine-grained tagsets can consist of thousands (as is the case of *WSD* systems that use WordNet[5] synsets as tags), or even millions (as is the case of wikifiers that use Wikipedia titles as tags). In the other extreme we can found coarse-grained tagsets. In the medical domain, for instance, in the *i2b2/VA* challenge [7] the tagset consisted on three tags: *Medical Problem*, *Treatment*, and *Medical Test*. In the *Semeval-2013 task 9* [8] focusing on drug-drug interaction (*DDI*), the tagset included *drug*, *brand*, *group* (group of drug names), and *drug-n* (active substance not approved for human use). Besides these task specific tagsets, subsets of Category sets in the most widely used medical resources (*MeSH®, SNOMED-CT*[6]*, UMLS®*) are frequently used as tagsets. In this research we used a subset of the top UMLS categories, namely, *Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology*, and *Procedures*.

Our approach consists of learning a binary classifier for each of the categories[7], whose results are combined using a simple voting schema. The cases to be classified are, according the contest instructions, the mentions in the document corresponding to term candidates, to refer to any of the concepts in the tagset. No co-reference resolution is attempted and, so, co-referring mentions could be tagged differently.

Most of the approaches to Semantic Tagging for small-sized tagsets, as our, use supervised Machine Learning (*ML*) techniques. The main problem found when applying these techniques is the lack of enough annotated corpora for

[5] http://wordnet.princeton.edu/

[6] http://ihtsdo.org/snomed-ct/

[7] In fact only 9 classifiers are learned, for the *Geographic Areas* category a conventional NERC is used.

learning. In our system we overcome this problem following a distant learning approach. Distant learning is a paradigm for relation extraction, initially proposed by [9], which uses supervised learning but with supervision not provided by manual annotation but obtained from the occurrence of positive training instances in a knowledge source or reference corpus. In [1] SNOMED CT, Wikipedia, and DBPEDIA[8] have been used as knowledge sources while in the research reported here only the first one has been used.

After this introduction, the organization of the article is as follows: In section 2 we sketch the state of the art of Semantic Tagging approaches. Section 3 presents the methodology followed in our previous work while Section 4 discusses the modifications performed for dealing with the current task. The experimental framework is described in section 5. Results are shown and discussed in section 6. Finally section 7 presents our conclusions and further work proposals.

2 Related Work

English is, by far, the most supported language for biomedical resources and tools. The National Library of Medicine[9] (NLM®) maintains the Unified Medical Language System[10] (UMLS®) that groups an important set of resources to facilitate the development of computer systems to "understand" the meaning of the language of biomedicine and health. It is worthnoting that only a small fraction of such resources exist for other languages.

A relevant aspect of information extraction is the recognition and identification of biomedical entities (like *disease*, *genes*, *proteins* ...). Several Named Entity Recognition techniques have been proposed to recognize such entities based on their morphology 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.

There is a number of tools that take profit of the UMLS resources. Some the more relevant are:

- *Metamap* [10] is a pipeline that provides a mapping among concepts found in biomedical research English texts and those found in the UMLS Metathesaurus ®. For obtaining such link the input text undergoes a lexical/syntactic analysis and a number of mapping strategies. Metamap is highly configurable (it has data, output and processing options) and is being widely used since 1994 by many researchers for indexing biomedical literature.
- *Whatizit* [11] is also a pipeline for identifying biomedical entities. It includes a number of processes where each one is specialized in a type of task (*chemical entities*, *diseases*, *drugs*...). Each module processes and annotates text

[8] http://wiki.dbpedia.org/
[9] http://www.nlm.nih.gov/
[10] http://www.nlm.nih.gov/research/umls/

connecting to a publicly available specific databases (e.g. UniProtKb/Swiss-Prot, gene ontology, DrugBank. . . .).
– Semantrix [11] is a private company that has developed the Ontotext Semantic Biomedical Tagger. It is an information extraction system designed to process biomedical texts using a number of biomedical databases.

Keeping on the medical domain, an important source of information are the proceedings of the *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text*[12] [7]. The challenge included three sub-tasks, the first one, *Concept Extraction*, namely patient medical problems, treatments, and medical tests, corresponding to Semantic Tagging[13]. Almost all the participants followed a ML supervised approach. Regarding the first task, the one related to our system, final results (evaluated using F1 metric) range from 0.788 to 0.852 for exact matching and from 0.884 to 0.924 for the lenient inexact matching.

A more recent and also interesting source of information is the *DDI Extraction 2013* (task 9 of Semeval-2013) [8]. Focusing on a narrower domain, Drug-Drug interaction, the shared task included two challenges: i) Recognition and Classification of Pharmacological substances, and ii) Extraction 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. Regarding the first task, the overall results (using F1) range from 0.492 to 0.8. As *DDI* corpus was compiled from two very different sources, *DrugBank* definitions and *Medline* abstracts, the results are quite different depending on the source of the documents, for *DrugBank*, the results range from 0.508 to 0.827, while for *Medline*, clearly more challenging, the results range from 0.37 to 0.53.

3 Methodology followed in our previous work

3.1 Outline

As we mentioned above, the system presented here is heavily based on [1]. In this Section we sketch the previous system (see details in the reference). The system proposes a machine learning solution to a tagging task. Therefore, it requires two main steps: training and annotation (see Figure 1). The main drawback of this type of solutions is the dependency on annotated documents, which usually are hard to obtain. Our main target was to train classifiers minimizing the impact of this issue and keeping good results. For such a purpose we use, within the distant learning paradigm, as learning examples, a set of seed words obtained with a minimal human supervision. We used as semantic classes the top level categories of the SNOMED CT hierarchy. More specifically its six more frequent classes.

[11] http://semantrix.com.au
[12] Other i2b2/VA contests deal with other relevant medical text processing problems as co-reference detection or identification of medications, doses, forms of administration, etc.
[13] The other two tasks were *Assertion classification* and *Relation classification*.

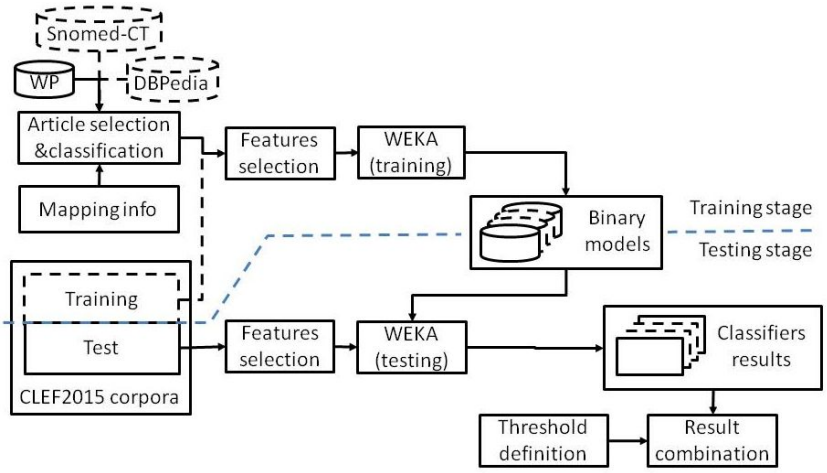

Fig. 1. Train and testing pipelines

We obtain an instance-based classifier (upper section in Figure 1) for each semantic class using seed words extracted from three widely used knowledge sources (section 3.2). The only form of human supervision is, as described below, the assignment of about two hundred Wikipedia categories to their appropriate SNOMED CT semantic class. Later (lower section in Figure 1) such models are used to classify new instances.

3.2 Features extraction

To obtain the seed terms needed for learning the classifiers, we proceed in three ways, using two different general purpose knowledge sources, Wikipedia and DBPEDIA, and one, SNOMED CT, specific for the medical domain (see [12] and [13] for analysis of these and other resources). From these sources, only Wikipedia has been used in the work presented here.

Wikipedia, although being a general purpose resource, densely covers the medical domain; it contains terminological units from multiple medical thesauri and ontologies, such as Classification of Diseases and Related Health Problems (ICD-9, ICD-10), Medical Subject Headings (MeSH), and Gray's Anatomy, etc. We describe here the main characteristics of the method followed to obtain the seed terms from Wikipedia, for the other sources [1] should be consulted.

First we got the set of the most reliable Wikipedia categories[14]. This resulted on a set of 237 Wikipedia categories. We manually assigned to such categories a unique SNOMED CT class from the set of 6 most frequent ones. For each of these categories we obtained the full set of associated pages. For each page, we

[14] See [14] for details about the way of obtaining such categories from Wikipedia resources

calculate a *purity factor*, i.e. a score (ranging in [0,1]), of the appropriateness of such page to a given SNOMED CT class[15]. For such classes only the pages having a purity of 1 are kept.

The seed terms have been obtained with low human supervision. As can be noticed by the way of collecting the seed terms, above, terms have associated Wikipedia pages. The results, so, are sets of Wikipedia pages to be used for learning the classifiers.

Following [15], we generate training instances by automatically labelling each instance of a seed term with its designated semantic class. When we create feature vectors for the classifier, 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.

3.3 ML machinery

We created a suite of binary contextual classifiers, one for each semantic class. The classifiers are learned using, as in [15], Support Vector Machine (SVM) models using Weka toolkit [16]. Each classifier makes a weighted 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 Wikipedia pages. Let $x_1, x_2, \ldots, x_n$ the seed terms for the semantic class t and knowledge source k, i.e. $x_i \in R_t^k$. Note that in this work only the source $k = wp$ is used. For each x_i we obtain its Wikipedia 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 t 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. Features are simply words occurring in the local context of mentions.

The above mentioned procedure applies for regular Wikipedia pages but our mechanism foresee also to use training corpus provided by the organizers. In this case the occurrence of a given tagged term is a positive example for the class that has been tagged but negative for the remaining classes.

For processing the full corpus we use an in-house general purpose sentence segmenter and POS tagger to identify non empty 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[16] without surpassing sentence limits.

For evaluation we used Wikipedia categories - SNOMED CT classes mappings as gold standard. We considered for each semantic class t a gold standard

[15] A purity 1 means that all the Wikipedia categories attached to the page are mapped (directly or indirectly) into the same SNOMED CT class, lower values of the purity may mean that the assignment of Wikipedia categories to SNOMED CT classes is not unique or not exists.

[16] In the experiments reported here n was set to 3.

set including all the Wikipedia pages with purity 1, i.e. those pages unambiguously mapped to t. The accuracy of the corresponding classifier is measured against this gold standard set.

4 Current Methodology

Although our aim is applying the previous approach to the current setting there significant differences that have to be faced:

1) The tagset is greater and comes from a different source (*UMLS* instead of *SNOMED CT*).
2) The language is French instead of English.
3) The genre of documents (EMEA and Medline) is very different from Wikipedia pages.

So, we performed the following changes over our previous system:

First, the way of collecting seedterms described in section 3.2 is modified as follows: We manually mapped the *UMLS* tagset into the set of SNOMED CT top categories (to the full 19 categories set, not to the 6 most frequent categories as in the previous system) and, further to English Wikipedia categories. We filtered out the English Wikipedia categories lacking French counterpart. For some *UMLS* categories as *LIVB* the mapping was not biunivocal, for other cases second level *SN* categories needed to be considered.

After filtering out categories not containing French interwiki links, for some of the *UMLS* classes a rather small set of Wikipedia classes remained, so we decided to extend the set by considering the French entity mentions occurring in the training set, we collected in this way 73 additional categories. We selected for each *UMLS* class the set of mentions tagged with the corresponding tag in the training collection existing as page or category in the French Wikipedia. In the case of pages we obtained the corresponding Wikipedia categories. Once collected a set of candidate French Wikipedia categories, we discarded those not having English counterpart and we manually revised the resulting set for accepting or rejecting each candidate and for assigning it to the correct UMLS tag. Then, for each UMLS tag we iterated over all their English Wikipedia categories for collecting all the pages having a purity 1 and having a French counterpart. In this way we obtained the initial sets of French seed words for each *UMLS* class.

Further processing, described in section 3.3, is basically the same. The only difference is that for French we have used for processing documents, in learning and test phases, the *Freeling* toolbox[17] has been used.

5 Experimental framework

First, we proceed to collect the seed terms for each semantic class t[18] and each knowledge source k. In our experiments we focussed on the Wikipedia-based ap-

[17] http://nlp.lsi.upc.edu/freeling/

[18] As said previously only 9

proach. The results for obtaining French terms starting from English Wikipedia categories are shown in Table 1 and Table 2.

Table 2 shows the number of French Wikipedia pages (i.e. terms) according their *UMLS* class with independence of their purity figure. For obtaining such pages we started using English data as shown in Table 1. This Table shows the global figures of the extraction process from the very beginning to the figure which represents the total number of French Wikipedia pages available for training.

Table 1 shows the global figures of the extraction process. In all the cases we found for the method based Wikipedia the number of terms (row 2), the French extension from the training data (row 3), the length of the initial categories set (row 4). Rows 5 and 6 show the number of pages from the English and French Wikipedia. The reason to discard some Wikipedia articles are: i) only pages with a length greater that 100 words are accepted, ii) some pages has been discarded due to difficulties in extracting useful plain text (pages consisting mainly of itemized lists, formulas, links, and so) and iii) Only Wikipedia pages with a purity 1 have been selected. In Table 2 we show the number of accepted terms splitted according the semantic class to which they belong.

Table 1. Terms effectively used for training

	WP only
initial WP categories	237
additional categories	73
Total WP categories	310
Total WP pages (EN)	16,972
Total available WP pages (FR)	3,564

6 Results

As mentioned above the learning phase has been followed using Wikipedia categories / *UMLS* classes mappings as golden standard and Wikipedia pages as input documents. For each seed term we obtained its corresponding Wikipedia page and, after cleaning, POS tagging, and sentence segmenting, we extracted all the mentions. For carrying out this linguistic process we used the *Freeling* suite (see [17] for details). For each mention the vector of features is built and the 9 learned binary classifiers are applied to it[19]. If none of the classifiers clearly classify the instance as belonging to the corresponding semantic class no answer is returned. If only one of the classifiers classifies positively the instance, the

[19] As quoted above the GEOG tag have been extracted using a conventional *NERC* based on using French DBPEDIA as a gazetteer

Table 2. Terms available to be used for training according to its SNOMED class

UMLS class	WP only
Disorder	876
Procedures	624
Physiology	485
Anatomy	1,587
Live Beings	27
Chemicals and Drugs	392
Phenomena	0
Devices	0
Objects	0
Total	2,402

corresponding *UMLS* tag 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*. As results of binary classifiers are scored, this method simply returns the class of the best scored individual result. It takes into account two threshold values: i) minimum class score and ii) minimum delta to the next better class score.
- *Meta-classifier*. A SVM multiclass classifier is trained using as features the results of the basic binary classifiers together the context data already used in the basic classifiers. The resulting class is returned.

For the experiments presented in this paper, only the first combination method has been tested. Several number of tests has been done by i) changing the number of WP articles including in training and ii) changing the threshold values mentioned above.

Table 3 depicts the global results as reported by the organization of CLEF2015. Unfortunately the material officially delivered included some severe issues regarding offset calculation. This is the main reason of the poor results reported. After detecting such issues the organisation of the contest proposed to fix the issue and resubmit a run. So we plan, once fixed the bug, to incorporate to the paper in the final release a new table showing our final results.

The results shown in Table 3 are really poor and far from the results obtained from our previous version performing on English Wikipedia pages, where we obtained accuracies of 87.4 for Wikipedia-based and Snomed-based approaches and 94.3 for DBpedia-based one. They require some explanation.

They can be justified from one side with some issues in our program to produce the results in a stand-off format as required by the organization. Table 5 shows very clearly that the terminological density of our Wikipedia corpus is several times lower that the training corpora provided by the organization.

Table 3. Results as reported by the organization of the CLEF2015's

–	entities exact match						entities inexact match					
run	TP	FP	FN	Precision	Recall	F1	TP	FP	FN	Precision	Recall	F1
EMEA	0	2260	1067	0	0	0	83	2177	938	0,0367	0,0813	0,0506
MEDLINE	82	2895	888	0,0275	0,0845	0,0416	354	2623	672	0,1189	0,345	0,1769

Table 4. Results locally calculated

run	TP	Tagged	Right selected	Semantically right
EMEA	2260	1090	421 (38,6%)	156 (6,9%)
MEDLINE	2895	640	286 (44,7%)	118 (3,9%)
MEDLINE*	2895	1008	450 (44,6%)	189 (6,3%)

Table 5. Terminological density in WP and CLEF corpora

	#Terms	#Sentences	Density [terms/sentence]
CLEF	4669	1692	2,76
WP	874	3158	0,50

Table 6. Medical entities as tagged in file 4176905.txt

Full sentence	Modifications des protéines sériques et du liquide synovial au cours de la polyarthrite rhumatoide .
Entities	protéines sériques, protéines, sériques, liquide synovial, liquide, synovial, polyarthrite rhumatoide, polyarthrite, rhumatoide

From the other side, such density is obtained by a tagging that embed several terms in a single sentence. An example of this situation is shown in file 4176905.txt. The sentence and terms tagged are shown in Table 6. There is no doubt that the tagging is correct but it is not clear that such concrete sentence contains 9 terms instead of 3 as most term extractors will do. This fact partially explain the low number of strings tagged by our system (see in Table 4 columns TP *versus* Tagged)

Obviously, the fact that in the current experiments learning is done from Wikipedia and test is performed over very different genres of documents, EMEA and Medline, while in our previous system the genre of training and test documents was the same is a drawback. The different coverage of French and English

Wikipedia and the lower accuracy of Freeling when performing on French texts are important factors, too.

Another minor issue is that text seems to include some kind of segmentation (see for example: *l' enfant* or *d ' activation plaquettaire induite par l ' héparine* among many others). The words by themselves are not important but such segmentation may cause errors in the POS tagging stage and this fact may be a real problem.

Nevertheless, Table 4 [20] shows the results using only the strings as comparison element (that is, without taking into consideration the offset values). The column "Right selected" refers only to the number of strings correctly selected while the column "Semantically right" refers to the UMLS tag assigned to such strings. Obviously a string may be correctly selected but the class assigned to it may be wrong. The analysis of this table confirms that the results are poor.

Leaving aside *GEOG*, that is detected using a specific mechanism, the best performing classes are *ANAT* and *LIVB* with a precision greater 50% probably due to the fact that they are the two classes more frequent in our training corpus.

In order to improve the results, we perform some tests using both WP and CLEF in the training stage. The results obtained are shown in Table 7. It shows an improvement in the performance but also shows a problem in the string selection. Examining the results in more details it reveals that if we consider only right selected strings the precision is about 50%. Table 8 gives a more detailed view about the results; it shows for a given true class it indicates which has been the estimated classes. The best result has been obtained for the class *DISO* that reaches a precision higher than 70%..

Table 7. Results locally calculated using both WP and CLEF as training corpus

run	TP	Tagged	Right selected	Semantically right
EMEA	2260	1088	394 (36,2%)	205 (9,1%)
MEDLINE	2895	1356	542 (66,7%)	281 (9,4%)

7 Conclusions and further work

We have presented a system that automatically detects and tags medical terms in general medical documents. The tagset used is derived from *UMLS* taxonomy. The results of the system, as discussed in previous section are poor and far from the obtained in our previous system, performing on English Wikipedia pages.

[20] This table shows two results for MEDLINE documents. The first is the result actually delivered for the contest. In this result a number of documents has been lost. The issue has been corrected and its result is indicated with an '*'. Please note that the figures correspond with those resulted for EMEA documents.

Table 8. Error analysis

Right class	Class proposed by the classifiers									
	DISO	PHEN	PROC	PHYS	ANAT	LIVB	CHEM	DEVI	OBJC	GEOG
DISO	238	2	49	14	30	26	43	4	1	0
PHEN	2	0	3	0	1	0	2	0	1	0
PROC	34	0	61	3	3	10	9	0	0	0
PHYS	14	1	9	5	1	4	15	1	3	0
ANAT	12	2	10	6	28	5	15	1	1	0
LIVB	15	0	14	4	5	71	0	0	0	0
CHEM	15	0	11	7	9	13	76	1	1	0
DEVI	2	0	1	2	1	0	3	2	0	0
OBJC	1	0	1	0	0	0	0	0	5	0
GEOG	0	1	1	0	1	0	4	0	0	0
Precision	71.47	0.00	38.13	12.20	35.44	55.04	45.51	22.22	41.67	0.0

An initial error analysis has detected a program issue in the way of computing off-sets of the detected mentions, also the extremely high difference in the density of mentions in the corpus used for learning (French Wikipedia pages) and for testing (French *EMEA* and *Medline* documents) seems to point to a high disagreement between training and test. A third issue is related to limitations in the performance of *Freeling*, specially in the basic tokenisation task.

The framework developed allows to perform additional experimenting changing several design parameters like the number of terms used for training, context width, features definition, etc. Some tests will be performed to optimize such parameters.

Several lines of research and a pending work will be followed in the next future (beyond fixing the issues reported above).

- As our results are based on one of the three knowledge sources used in our previous work, an obvious way of possible improvement is the use of the other two resources (SNOMED CT and DBPEDIA)
- A combination and/or the specialization of the resources for learning more accurate classifiers. The application of the DBPEDIA based approach, clearly the most productive one, to all the classes merits a deeper investigation.
- A careful combination of learning from the learning dataset and from additional material should be experimented
- Table 8 shows that three of the classes produced no results at all and another one only detected one term. In these cases the corresponding classifiers have a extremely low accuracy, probably due to few training examples. So acquiring additional examples for these cases could result on improvements.
- 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).

– Improving the selection of medical entities by using POS pattern learning, adapting our term extractor to the tagging policy of medical entities in Quaero corpus and improving adaptation of Freeling to French medical texts.

8 Acknowledgements

This work was partially supported by the SKATER project (Spanish Ministerio de Economía y Competitividad, TIN2012-38584-C06-01 and TIN2012-38584-C06-05).

References

1. Vivaldi, J., Rodríguez, H.: Medical entities tagging using distant learning. In: CICLing 2015, Part II, LNCS. Volume 9042. (2015) 631–642
2. Goeuriot, L., Kelly, L., Hanna Suominen, L.H., Névéol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2015. clef 2015 - 6th conference and labs of the evaluation forum. Lecture Notes in Computer Science (LNCS), Springer (2015)
3. Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., Zweigenbaum, P.: CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In: CLEF 2015 Online Working Notes, CEUR-WS (2015)
4. Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The Quaero french medical corpus: A ressource for medical entity recognition and normalization. In: Proceedings of the Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing - BioTxtM2014. (2014) 29–30
5. Ling, X., Singh, S., Weld, D.S.: Design challenges for entity linking. TACL **3** (2015) 315–328
6. Gattani, A., Lamba, D.S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., Doan, A.: Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. PVLDB **6**(11) (2013) 1126–1137
7. Özlem Uzuner, South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. In: Journal of the American Medical Informatics Association. Volume 18. (2011) 552–556
8. Segura-Bedmar, I., Martínez, P., Zazo, M.H.: Lessons learnt from the DDI extraction-2013 shared task. In: Journal of Biomedical Informatics, Elsevier, ISSN: 1532-0464. (January 2014)
9. Mintz, M., Bills, S., Snow, R., , Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the ACL. (2009) 1003–1011
10. Aronson, A.R., Lang, F.M.: An overview of Metamap: historical perspective and recent advances. In: JAMIA. Volume 17. (November 2010) 229–236
11. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Jimeno, A.: Text processing through web services: calling Whatizit. In: Bioinformatics Applications Note. Volume 4. (November 2008) 296–298
12. He, J., de Rijke, M., Sevenster, M., van Ommering, R., Qian, Y. In: Generating Links to Background Knowledge: A Case Study Using Narrative Radiology Reports, Glasgow, Scotland, UK. (October 2011)

13. Yeganova, L., Kim, W., Comeau, D., Wilbur, W.J.: Finding biomedical categories in medline®. In: Journal of Biomedical Semantics. (2012)
14. Vivaldi, J., Rodríguez, H.: Using Wikipedia for term extraction in the biomedical domain: first experience. In: Procesamiento del Lenguaje Natural. Volume 45. (2010) 251–254
15. Huang, R., Riloff., E.: Inducing domain-specific semantic class taggers from(almost) nothing. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden (2010) 275–285
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. In: SIGKDD Explorations. (2009)
17. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the 8th international conference on Language Resources and Evaluation, European Language Resources Association (ELRA) (2012)