Semantic tagging and normalization of French medical entities

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

[1] Universidad de Buenos Aires, Buenos Aires, Argentina,
vcotik@dc.uba.ar
[2] Universitat Politècnica de Catalunya, UPC, Barcelona, Spain,
horacio@lsi.upc.edu
[3] Universitat Pompeu Fabra, UPF, Barcelona, Spain, jorge.vivaldi@upf.edu

Abstract. In this paper we present two tools for facing task 2 in CLEF eHealth 2016. The first one is a semantic tagger aiming to detect relevant entities in French medical documents, tagging them with their appropriate semantic class and normalizing them with the Semantic Groups codes defined in the *UMLS*. It is based on a distant learning approach that uses several SVM classifiers that are combined to give a single result. The second tool is based on a symbolic procedure to obtain the English translation of each medical term and looks for normalization information in public accessible resources.

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

1 Introduction

We developed a semantic tagger for the medical domain [1] performing on English Wikipedia pages[4] (*WP*) previously selected as belonging to the domain using a distant learning approach. Our aim here is exploring whether the approach can be applied to other language (French), other genre (scientific documents) and other tagset, and to normalize the semantic tags to the Unified Medical Language System (*UMLS*). We performed these experiments within the framework of CLEF2016 eHealth contest[5] (see details in [2]). More specifically in Task 2, Multilingual Information Extraction as described in [3][6].

Semantic Tagging is 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, while the latter is closely related to Named Entity Classification.

The key elements of *Semantic Tagging* task are: (i) *the document*, or document genre, to be processed, (ii) *the linguistic units* to be tagged and (iii) *the*

[4] http://en.wikipedia.org
[5] https://sites.google.com/site/clefehealth2016/
[6] https://sites.google.com/site/clefehealth2016/task-2

tagset. All these elements play a crucial role for the success of the task. In this concrete task our constraints are the following: (i) documents of medical domain, mainly scientific articles indexed in MEDLINE and some drug monographs published by the European Medicines Agency (EMEA), (ii) the linguistic units to be tagged are the terminological strings found in the source documents, (iii) the tagset will be a subset of the top UMLS categories. Such resources will be used also for the normalization on the medical entities as defined in the phase II.

Our approach consists of learning a binary classifier for each of the categories, whose results are combined using a simple voting schema. The cases to be classified are the mentions in the document corresponding to TCs, to refer to any of the concepts in the tagset. No co-reference resolution is attempted and, so, co-referring mentions may be tagged differently. For the normalization of the entities found we used the resources available through BioPortal[7].

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 the current task. 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

English is, by far, the most supported language for biomedical resources and tools. The National Library of Medicine[8] (NLM®) maintains the Unified Medical Language System[9] (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 worth noting 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 are often incomplete and different variations may be found in the literature; therefore it is necessary to minimize this issue as much as possible.

2015 edition of CLEF eHealth contest contained two tasks focusing on information extraction and information retrieval. The topic of one of them was Clinical Named Entity Recognition in medical texts written in French (Task 1b) [4]. Seven teams participated in this task. Two types of biomedical documents were used: scientific articles indexed in the MEDLINE database, and full text drug monographs published by the European Medicines Agency (EMEA). The

[7] SPARQL Endpoint available at http://bioportal.bioontology.org/
[8] http://www.nlm.nih.gov/
[9] http://www.nlm.nih.gov/research/umls/

best system obtained F-measure of 0.756 for plain entity recognition, 0.711 for normalized entity recognition, and 0.872 for entity normalization [5].

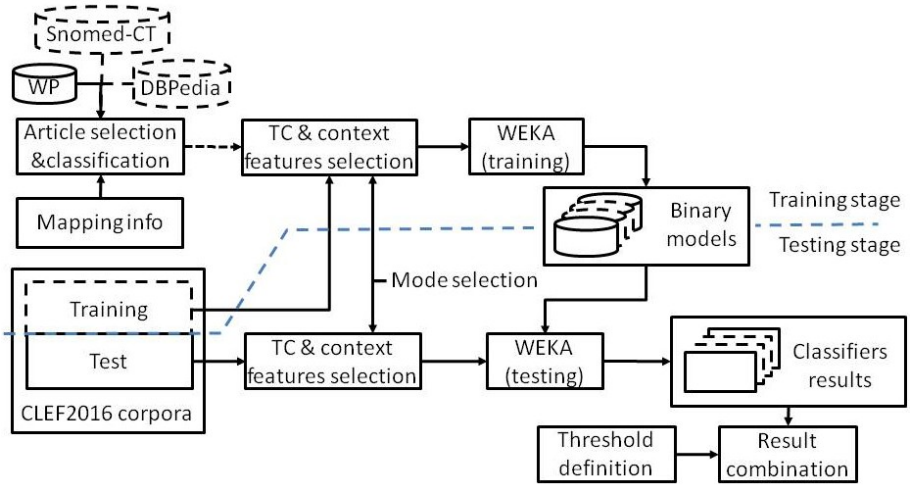

Fig. 1. Train and testing pipelines

3 Our approach

For participating in CLEF e-Health 2016 Task 2 we have submitted two runs to the Plain Entity Recognition Task and one run to the Normalization task. A description of our approaches followed in these three runs are presented below.

3.1 Run1: WP-based system using distant learning

In this run we follow basically the approach of our previous system, presented to CLEF e-Health 2015 [6], in turn based on a semantic tagging system aiming to detect and classify medical entities in English WP pages [1]. A multilingual extension (to Arabic, French, and Spanish) of this latter system, can be found in [7]. We sketch next the approach we follow, highlighting the differences between the current and the previous system. Details of the latter can be found in [6]. Figure 1 presents a overall vision of the system.

The core idea of the system is that for a term t known to belong to the semantic category c (one of the 10 *UMLS* categories we deal with) not only the occurrences of t in the training material can be considered positive examples for learning but also the occurrences of t in its associated *WP* page if existing. This hypothesis is important because for some semantic categories the training material contains not enough terms for accurate learning.

Following [8], 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.

We created a suite of binary contextual classifiers, one for each semantic class. The classifiers are learned using, as in [8], Support Vector Machine models utilizing *Weka* toolkit [9]. Each classifier makes a weighted decision as to whether a term belongs or not to its semantic class.

For every file of the training corpus, each tagged term is considered as a positive example for the tagged class and negative example for the rest of the classes. Features are the words occurring in the local context of mentions. The context size and POS of the context words are parametrizable.

Examples for learning correspond to the mentions of the seed terms in the corresponding *WP* pages. Let $t_1, t_2, \ldots, t_n$ the seed terms for the semantic class c, i.e. $t_i \in ST^c$. For each t_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 c 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.

For French we have used for processing documents, in learning and test phases, the *Freeling* toolbox[10] [10].

Term candidates, *TC*, are selected according to morpho-syntactic criteria. We have used for filtering the following regular expression: NA*(PNA*)+. Additionally, in order to take into account the peculiarities of the term selection of CLEF organizers we also decompose each complex term in its components (see section 4.1 for more details and examples).

The learning process has been performed using for each semantic category the most likely relevant documents including EMEA and MEDLINE training documents and *WP* pages obtained as described above. From the *WP* pages, besides those with purity less than 1, short pages and pages consisting mainly of itemized material or non-textual fragments were removed too.

For each example, the feature vector captures a context window of n words to its left and right[11] without surpassing sentence limits.

3.2 Run2: Knowledge-based approach

A careful analysis of our results on CLEF e-Health 2015 participation revealed that some apparently easy to detect terms were not detected or were classified incorrectly. For instance French terms occurring in French *DBPedia*[12] or translated English terms occurring in English (Princeton) *WN* or in English *DBPedia* were not detected. We decided, thus, combining, in our run2, the results of run1

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

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

[12] http://wiki.dbpedia.org/

with two other systems, one based on the performance of a state-of-the-art term extractor, YATE, tuned to work in the medical domain, and the other based on an external knowledge source, the *DBPedia*. Although domain independent, *DBPedia* has a nice coverage of medical classified terminology and offers good interlingual capabilities.

Extracting Term Candidates using YATE and wikiYATE *YATE* [11] basically performs using the taxonomic structure of the nominal part of *WN*. Given a domain *d*, the medical domain here, *YATE* obtains the called *Domain Borders*, synsets that are likely to belong, both them and their descendants to *d*. These *Domain Borders* are used later for extracting from a document the set of mentions corresponding to terms belonging to *d*, i.e. those *TC* whose synsets are placed below a *Domain Border*. Right part of Figure 2 shows this process.

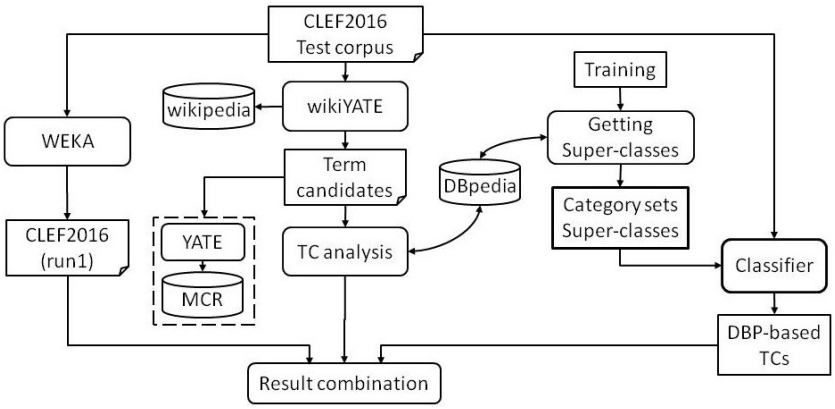

Fig. 2. Run 2 improvement

YATE uses English *WN*, therefore for applying this tool a translation of French *TC*s is needed. As the terms we are interested on are those represented in the French *WP*, we used the interwiki links between French and English *WP* (using *DBPedia* for getting these links). This process results on the extraction (tagging) of the medical terms occurring in the test documents. *YATE* is a term extractor, not a semantic tagger, so it is not able to classify the extracted terms, but, indirectly, these terms can be used as seed terms for learning the classifiers.

wikiYATE [12] is a similar term extractor but it is multilingual and uses *WP* as knowledge source for detecting terms. Both term extractors have been used in this task.

DBPedia-based approach Some useful information in *WP* is represented in infoboxes and, thus, has been automatically mapped into the corresponding *DBPedia* rdf triples. We take profit of several interesting properties of *DBPedia*:

– There exist *DBPedia* datasets for English (http://dbpedia.org/sparql) and French (http://fr.dbpedia.org/sparql).
– Entities (resources) in the two datasets are frequently linked through *sameAs* properties.
– Entities in the datasets are frequently mapped to one or more linguistic referents, words and phrases, through *label* properties, sometimes in several languages. So an entity in the English *DBPedia* can be labelled with French words or phrases.

As shown in Figure 3, iterating over French and English terms and resources in French and English *DBPedia* datasets through the *label* and *sameAs* properties, we are able to collect from an initial French *TC*, *t*, the set of English resources likely corresponding to translations of *t*.

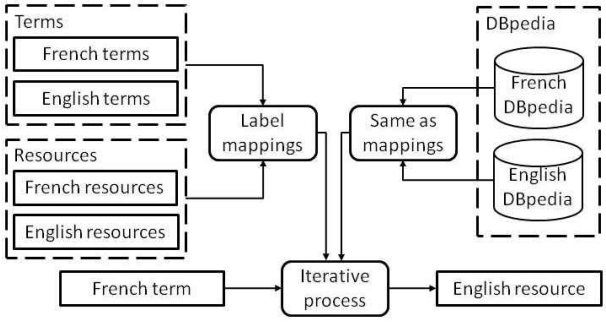

Fig. 3. French to English term translation

In order to be able to classify t into one of the 10 semantic categories we proceed on the following way:

During training, for each semantic category c we collect all the French terms t occurring in the training dataset and tagged with c. We filter out from these sets the terms not occurring in French *WP*. We then collect, as described above, the set of English *DBPedia* resources associated to them. For each of such resources r we obtain the set of classes to which r belongs (we reduce our search to *DBPedia* and *YAGO* classes) using the *type* property. From each class we recursively collect the set of super-classes using the *subClassOf* property. Some of the super-classes belong unambiguously to one semantic category. For instance, *http://dbpedia.org/class/yago/AliphaticCompound114601294* occurs as super-class of 6 terms all classified as *CHEM*. Others, as, *http://dbpedia.org/ontology/Eukaryote*, are ambiguous. This super-class occurs 4 times as *PHYS*, 8 times as *LIVB*, and 2 times as *DISO*.

For each semantic category we collect the set of unambiguous super-classes. For ambiguous cases we proceed as follows: If the total number of terms covered by the super-class is higher than a threshold *THR1* and the ratio between the

higher option and the second is higher than a threshold *THR2* we assign the super-class to the set corresponding to the first option. *THR1* and *THR2* have been manually set to 30 and 4. For instance in *http://dbpedia.org/class/yago/Location100027167*, occurring 10 times as *ANAT*, 13 times as *CHEM*, and 76 times as *GEOG*, the two conditions hold (76+13+10 >30 and 76/13 >4, so the super-class is included into the set of super-classes corresponding to the semantic category *GEOG*.

In this way a number of super-classes have been associated to each semantic category. See in Table 1 the size of each set and an example of their members.

Table 1. More frequent top super classes

Semantic class	Top super-classes	Example
ANAT	16	http://dbpedia.org/class/yago/BodyPart105220461
LIVB	27	http://dbpedia.org/class/yago/Animal100015388
CHEM	53	http://dbpedia.org/class/yago/Inhibitor114724436
PROC	21	http://dbpedia.org/class/yago/Treatment100658082
GEOG	31	http://dbpedia.org/ontology/PopulatedPlace
DISO	35	http://dbpedia.org/class/yago/Infection114174549

Once collected these sets (during the training phase) the process at test phase is quite straightforward. See left part of Figure 2. For each *TC* in the test dataset, following the approach described above, we obtain the set of English *DBPedia* resources. The *DBPedia* and *YAGO* classes of these resources are then obtained and from them the set of super-classes. This set is intersected with the sets of super-classes associated to each semantic category. The sizes of all the intersections are computed and the category associated to the higher size is returned as category of the term (ties are solved according to the most frequent semantic category in the training dataset). Finally the results of run1 (a semantic category), DBP-based (a semantic category), and YATE-based (a Boolean) are combined for getting the final result.

3.3 Normalization

For normalization we have used the *BioPortal SPARQL endpoint*. The a priori obvious way of accessing *UMLS* and obtaining the *CUI* of a term t consisted on getting a *UMLS* entity labelled with t and then getting the *CUI* of this entity. This simple procedure does not work because *UMLS* is not directly labelled with terms. We have used instead an indirect access to *UMLS* through other ontologies. We have employed for such process *Snomed-CT*, *Mesh*, and *RCD*. The process consists on translating French into English, using the approach described above and then accessing to any of the intermediate ontologies and

from them to *UMLS*. In Figure 4 one of the SPARQL templates used is presented. This template is instantiated into a real query just by replacing the placeholders '**ont**' by the name of one of our three ontologies and '**term**' by the name of the English *TC*. As can be seen in the Figure, this template uses the *prefLabel* (preferred label) link. We have also built templates using the *altLabel* (alternate label) link, and others using approximate string matching, for covering decreasingly confident matchings.

```
SELECT DISTINCT *
    FROM <http://bioportal.bioontology.org/ontologies/**ont**>
    FROM <http://bioportal.bioontology.org/ontologies/globals>
WHERE {
   ?x rdfs:label ?label .
   ?x <http://bioportal.bioontology.org/ontologies/umls/cui> ?y
   FILTER (regex( UCASE(str(?label)), "**term**"))}
```

Fig. 4. Example of SPARQL template for accessing to BioPortal

4 Experimental framework

Participants of CLEF2016 are requested to perform named entity recognition and normalization on a dataset of scientific article titles and full-text drug inserts. For performing such tasks we designed the working frameworks that are described in following subsections.

4.1 Entity recognition

Our basic working framework for entity recognition was the same than in CLEF2015. But taking into account the results obtained in such contest (see [6]) we perform some experimentation based on such material in order to decide to include or not *WP* pages in our learning framework. We test several configuration of features selection as well as different number of *WP* pages for using in training stage. Such framework foresees to automatically select *WP* pages for each class. We manually check a number of such pages in order to correct some issues with automatic selection phase. The results clearly shown that there was not any improvement in adding *WP* material in the training phase. Therefore we decided to train our model using only the training material provided by the CLEF2016 organization and provide such results as run 1.

As mentioned in section 3.1 the learning phase was made using the distant learning paradigm. For each mention of a TC the vector of features is built and the nine[13] learned binary classifiers are applied to it. For building such classifiers all the documents of the training corpus were linguistically processed

[13] In our run 1, the process of extracting GEOG entities was performed by a Geographic NER, and, so only nine classifiers were learnt.

using the *Freeling* suite (see [10] for details). The vector of features was built using the lemmas of the context words that within a window of 3 tokens of the *TC* (excluding determiners and punctuation signs). We used the lemmas to as features but we defined several ways to select such lemmas: (i) Mode 0: any context word within the window, (ii) Mode 1: only nouns and adjectives and (iii) Mode 2: only verbs, nouns and adjectives.

The Quaero corpus takes into account nested terms as different terms. Given this fact, when the TC is poly-lexical, all the possible combinations of components are taken into account. Table 2 shows some examples.

Table 2. Processing nested terms

TC found	Additional terms to be processed
cancers digestifs	cancers, digestif
dose de fentanyl	dose, fentanyl
clip péri-cave d' Adams	clip, péri-cave, clip péri-cave

We also decided to include a second run that, starting with run 1 results, improves them by doing some symbolic processes as shown in Figure 2 and described in the following paragraphs.

- Term extraction and analysis. For preparing run 2 we create a single document that includes all documents of the test corpus. We analyse such material with wikiYATE, a term extraction tool that uses *WP* for obtaining the TCs of a given text (see description in [12]). This tool ranks the *TC* according a termhood value; we create a set of string composed by: (i) those *TC* above a given threshold, (ii) those *TC* not found in WP and (iii) the list of all the adjectives that take part of the chosen TCs. We look in the DBpedia for the English translation of these units and if available processed them using YATE ([11]), a medical term extraction tool that uses the Multilingual Central Repository[14] (*MCR*) [13] for analysing the TCs. This tool in addition to give a termhood value for each TC, provides with some basic class information that we mapped to *UMLS* classes.
- DBpedia exploration as described in section 3.2.

Finally, the results of both analysis has been combined in a single result that was used to improve the result run 1.

4.2 Entity normalization

The process of Entity normalization was performed independently from the process of entity recognition. This is why we submitted runs for the task of plain

[14] http://adimen.si.ehu.es/web/MCR/

entity recognition and not to the task of normalized entity recognition. See 3.3 for details on our approach.

5 Results

Table 3 depicts the global results as reported by the organization of CLEF2016 (phase I task entity recognition). The material officially delivered included two runs. Unfortunately, for the run 2 we incorrectly submitted the same material as in run 1. After detecting such issue the organisation kindly accepted to evaluate our actual run as an unofficial result. For this reason we tagged with an "*" run2 results showed in Table 3. Additionally we performed an after challenge improvement of our system (noted as "/3*" in the table) introducing a simple voting mechanism for unifying the tags corresponding to multiple mentions of the same *TC* in the case enough evidence for one of the choices exists.

Table 3. Results as reported by the organization of the CLEF2016's (phase I)

–	entities exact match						entities inexact match					
docs/run	TP	FP	FN	Prec.	Recall	F1	TP	FP	FN	Prec.	Recall	F1
EMEA/1	512	3463	1835	0,1288	0,2182	0,1620	962	3013	1653	0,2420	0,3679	0,2920
EMEA/2*	420	4025	1816	0,0945	0,1878	0,1257	864	3581	1613	0,1944	0,3488	0,2496
EMEA/3*	654	1550	3538	0,2967	0.1560	0.2045	903	1301	2976	0.4097	0.2328	0.2969
MEDLINE/1	736	5053	2369	0,1271	0,2370	0,1655	1446	4343	1988	0,2498	0,4199	0,3132
MEDLINE/2*	969	5050	2138	0,1610	0,3119	0,2124	1759	4260	1684	0,2922	0,5109	0,3718
MEDLINE/3*	1078	2025	4933	0.3474	0.1793	0.2366	1575	1528	4113	0.5076	0.2769	0.3583

Table 4. Results as reported by the organization of the CLEF2016's (phase II)

–	entities exact match						entities inexact match					
docs	TP	FP	FN	Prec.	Recall	F1	TP	FP	FN	Prec.	Recall	F1
EMEA	517	558	558	0,4809	0,4809	0,4809	517	558	558	0,4809	0,4809	0,4809
MEDLINE	673	745	748	0,4746	0,4736	0,4741	673	745	748	0,4746	0,4736	0,4741

The results obtained for the Phase I (entity recognition) are poor (although better than those proposed in CLEF2015, see [6]) and far from the results obtained from our previous experiments on French Wikipedia pages[15].

[15] Using a very similar methodology to classify medical WP pages (over six classes) we obtained accuracies of 74.76% (exact match). See [7]

Table 5. Medical entities as tagged in file 49922.txt (MEDLINE)

Full sentence	Indications de la radiothérapie pour les tumeurs digestives
Entities	Indications, radiothérapie, tumeurs digestives, tumeurs, digestives

As already shown in [6], the terminological density of the QUAERO corpus is very high. As the same time, such density is obtained by a tagging methodology that nest several terms in a single polylexical term. An example of this situation is shown in Table 5. Undoubtedly, the tagging is correct but it is not clear that such concrete sentence actually contains 5 terms instead of just 3 (Indications, radiothérapie and tumeurs digestives) as most term extractors will do.

Another minor issue is that text seems to include some kind of extra 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 for TC delimitation ("l" and "d" will become a noun instead of a determiner and preposition respectively). Also, the generation of the final stand-off annotation becomes a bit more complicated.

Table 6 shows a detailed analysis of run 1 results. From one side, there are two classes (PHEN and GEOG) that does not produce any correct result and another class that only detects one valid term (OBJC). In these cases, the corresponding classifiers have a extremely low accuracy, probably due to the lack of training examples. So, acquiring additional examples for these cases should result on some improvement. From other side, there are some classes (DISO, PROC and ANAT) where the number of examples is much higher and therefore show a better result .

Table 7 shows the same analysis for run 2. There is an improvement in the performance for all the classes showing that: (i) the symbolic analysis partially solves the inaccuracies of the machine learning system and (ii) the combination of methods improves the global efficiency.

6 Conclusions and further work

The organizers of CLEF eHealth 2016 divided the task 2 in two phases: entity recognition and entity normalization on French medical text ofthe Quaero corpus. Our approach results in two different systems for solving each task.

For the first task, we have presented a system that automatically detects and tags medical terms in medical documents using a tagset derived from *UMLS* taxonomy. The results of the system for entity recognition, as discussed in previous section are poor, far from the obtained in our previous system (performing on medical English *wp* pages and confirmed for other languages, including French, [7]) but much better than those obtained in our participationin CLEF2015. The improvement was specially high in our run 2 that includes some symbolic pro-

Table 6. Error analysis run 1

Right class	Class proposed by the classifiers									
	DISO	PHEN	PROC	PHYS	ANAT	LIVB	CHEM	DEVI	OBJC	GEOG
DISO	451	26	180	56	185	92	184	19	24	16
PHEN	21	0	5	1	6	2	3	0	0	0
PROC	136	10	215	31	62	48	107	14	9	8
PHYS	13	1	17	11	11	1	18	2	0	0
ANAT	59	2	32	12	68	4	19	2	2	0
LIVB	51	3	35	16	26	131	16	2	3	3
CHEM	31	7	37	15	18	20	131	8	10	1
DEVI	15	0	7	4	4	3	1	4	0	0
OBJC	2	0	3	4	0	1	9	0	1	1
GEOG	7	0	1	1	6	4	3	1	0	0
Precision	57.38	0.00	40.41	7.28	17.62	42.81	26.68	7.69	2.04	0.00

Table 7. Error analysis run 2

Right class	Class improvement proposed by the symbolic process									
	DISO	PHEN	PROC	PHYS	ANAT	LIVB	CHEM	DEVI	OBJC	GEOG
DISO	499	24	160	49	118	93	162	18	21	7
PHEN	16	0	5	1	4	2	3	0	0	0
PROC	114	9	270	22	47	48	96	13	8	4
PHYS	12	5	17	37	8	1	17	2	0	0
ANAT	52	1	23	10	163	4	25	4	3	0
LIVB	45	3	27	14	21	131	15	2	3	2
CHEM	27	7	23	11	11	20	160	8	12	1
DEVI	13	0	4	3	3	2	1	4	0	0
OBJC	1	0	1	3	0	1	9	0	1	0
GEOG	7	0	2	1	11	4	3	1	1	15
Precision	63.49	0.00	50.75	24.50	42.23	42.81	32.59	7.69	2.04	51.72

cessing for improving the results. The working framework allowed us to experiment with several design parameters like the number of terms used for training, context width, features definition, etc. Undoubtedly, this is at the base of the improvement obtained for run 1. The use of some symbolic processing on the results of run 1 allow us to obtain some additional improvement.

It is interesting to observe that in all cases the improvement is higher in the inexact match than in the exact match. This fact may reveal some issues in the TC delimitation but also in the offset calculation. The latter issue is magnified

by the tokenization of the training corpus that difficulties the linguistic analysis and the offset calculation.

The second task was solved using a totally different system. It is based in obtaining the normalization information from public resources after obtaining the English translation of each medical term. The results were a bit below of the other participants . Again, tokenization is an issue that affects the performance of the system.

Several research lines will be followed in the next future:

- The integration of both entity recognition and normalization in a single task may bring mutual benefits.
- To enlarge the use of *BioPortal* for looking in the ontologies for the recognition and classification task seems to be a promising direction.
- A combination and/or the specialization of the resources for learning more accurate classifiers. The application of the *DBPedia* based approach, to all the semantic classes merits a deeper investigation.
- A careful combination of learning from the training dataset and from additional material, as *WP* should be experimented.
- The features currently used for learning the classifiers are rather crude and need some revision. We foresee to do some experimentation weighting the features, separating the features according its position in relation to the TC and adding new features as: start/end characters, typed features, etc.
- 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.

7 Acknowledgements

This work was partially supported by the TUNER project (Spanish Ministerio de Economía y Competitividad, TIN2015-65308-C5-5-R).

References

1. Vivaldi, J., Rodríguez, H.: Medical entities tagging using distant learning. In: CICLing 2015, Part II, LNCS. Volume 9042. (2015) 631–642
2. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2016. In: LNCS, Springer (September 2016)
3. Névéol, A., Goeuriot, L., Kelly, L., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Rey, G., Robert, A., Tannier, X., Zweigenbaum, P.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: CLEF Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (September 2016)
4. 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)

5. 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. LNCS, Springer (2015)
6. Cotik, V., Vivaldi, J., Rodríguez, H.: Semantic tagging of French medical entities using distant learning. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum. (2015)
7. Cotik, V., Vivaldi, J., Rodríguez, H.: Arabic medical entities tagging using distant learning in a multilingual framework. Submitted
8. 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
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. In: SIGKDD Explorations. (2009)
10. 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 (2012)
11. Vivaldi, J., Rodríguez, H.: Medical term extraction using EWN ontology. In: Proceedings of Terminology and Knowledge Engineering. (2002) 137–142
12. 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
13. Gonzalez-Agirre, A., Laparra, E., Rigau, G.: Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In: Proceedings of the Sixth International Global WordNet Conference (GWC'12)., Matsue, Japan (2012)