LIMSI @ CLEF eHealth 2015 - task 1b

Eva D'hondt, François Morlane-Hondère, Leonardo Campillos, Dhouha Bouamor, Swen Ribeiro, and Thomas Lavergne

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
(LIMSI-CNRS 3251),
Rue John von Neumann , 91400 Orsay, France
eva.dhondt@limsi.fr, francois.morlane-hondere@limsi.fr,
leonardo.campillos@limsi.fr, dhouha.bouamor@limsi.fr,
swen.ribeiro@limsi.fr, thomas.lavergne@limsi.fr

Abstract. This paper presents LIMSI's participation in the Clinical Named Entity Recognition task at the CLEF eHealth 2015 workshop. Our system is based on the combination of three classifiers: two CRFs to detect entities' boundaries and a SVM to identify their semantic class. These classifiers rely on a set of features used in state-of-the-art classification systems, including token/POS ngrams, morphologic features, and dictionary consultation in language-dependent external sources. Although our system was not fully operational when we submitted our run, we obtained above-average scores. In this paper we also present two additional runs which improve on the submitted system.

Keywords: named entity extraction, biomedical language, conditional random fields, classification, support vector machines

1 Introduction

With the increased availability of biomedical text and electronic patient records, the focus of the biomedical NLP community has shifted from data collection to text analysis over the last ten years. One of the most basic text analysis problems is Named Entity Recognition (NER): The identification and categorization of references to entities (*mentions*) in natural language text. While current NER systems for 'traditional' genres such as newswire text generally achieve high accuracy, those for the biomedical domain consistently lag behind. Biomedical NER (also called Clinical NER) is generally considered a more difficult task than traditional NER for several reasons. First, the biomedical domain is a fast-moving field and new entities and entity names are constantly added, which makes it hard for knowledge bases such as the Unified Medical Language System (UMLS) [1] to maintain an adequate coverage. Second, as in the patent domain, naming conventions in the biomedical domain are non-standardized. That is, a same entity may be referred to by different spelling variations: e.g. 'NF-Kappa B', 'NF Kappa B', 'NF KappaB', or 'NF-Kappa II'. Furthermore, authors in the biomedical domain often coin their own abbreviations at the beginning of an article or report. These are often highly ambiguous and do not necessarily appear in an existing dictionary or knowledge base. For example, 'AA' can stand for 'Alcoholic Anonymous', 'arachidonic acid', 'amino

acid' or 'amena'. Third, entity names in the biomedical domain are generally longer than those in newswire text, and contain many modifiers. This complicates the detection of entity boundaries considerably, and gives rise to two additional problems for linear labeling systems: (i) Conjunction and disjunction, and (ii) embedded entities. In the case of the former, two or more entities may share the same head noun, e.g. 'cancer du côlon et du rectum', in which 'cancer' is the shared head. An embedded (a.k.a. nested) entity is an entity that is completely encapsulated in a larger entity. For example, the entity 'virus de l' hépatite murine' also contains the following entities: 'virus', 'hépatite murine', 'hépatite' and 'murine'.

Over the last few years considerable research effort has been invested in biomedical NER. International shared tasks with associated, publicly available annotated corpora, such as the i2b2 challenge in 2010 [2] and the BioNLP/NLPBA 2004 shared task [3], have lead to major improvements in this particular task. Currently, the most well-known and widely used NER system for identification of biological entities is MetaMap [4], which is maintained by the National Library of Medicine (NLM). However, since Clinical NER depends heavily on external resources, the current systems are language-dependent, and—as is often the case—the vast majority of existing systems are geared towards the extraction from English texts only.

The organizers of the CLEF 2015 eHealth challenge [5] wanted to address this problem and proposed a separate task (task 1b) [6] that focused on constructing a Biomedical Named Entity recognition system for French biomedical text. The focus poses its own challenges: (1) The coverage of French in the existing biomedical resources such as the UMLS is much less extensive than that of English; (2) In general, fewer resources and dedicated tools for processing biomedical text are available.

This paper describes our participation in the Named Entity recognition task (task 1b). We present a hybrid system which consists of a two-level approach for entity boundary detection with Conditional Random Fields (CRFs) and a Support Vector Machine (SVM) classifier for Named Entity classification. In the next sections we outline the aim of the task and the test corpus (§2), describe our system (§3), report the runs and results (§4), discuss our outcomes (§5) and put forward some conclusions and future work (§6).

2 Task and Corpus Cescription of CLEF eHealth Task 1b

The CLEF 2015 eHealth task 1b addressed Clinical Named Entity Recognition in French medical texts. It is a follow-up of the task set forth in the 2013 CLEF-ER challenge[1] that dealt with multilingual Named Entity Recognition in parallel corpora. The organizers of task 1b proposed two subtasks: (i) *Entity recognition*, and (ii) *Entity normalization*. *Entity recognition* requires identifying a clinical entity in a given French text and annotating it with its corresponding UMLS Semantic Group [7] out of the following set: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures. *Entity normalization* involves linking the extracted entity to its corresponding UMLS concept(s). We opted to only participate in the first subtask ('Entity Recognition').

[1] https://sites.google.com/site/mantraeu/clef-er-challenge

The data set used in this task was taken from the QUAERO French Medical Corpus [8] and consisted of two separate subcorpora: MEDLINE titles, and texts from the European Medicines Agency (EMEA). The MEDLINE subcorpus contained 833 documents for training. Each document is a title of journal article in the PubMed database. The EMEA training subcorpus contained full-text documents for three public assessment reports. An assessment report contains a description of a particular medication and offers recommendations on the conditions of use. The text in both subcorpora are rather different. EMEA texts are full-text documents with full, sometimes grammatically complex sentences. Consequently, learning language models from them is feasible. In contrast, the MEDLINE subcorpus contains only PubMed titles, which are often complex noun phrases and not full sentences. Moreover, the EMEA corpus is more likely to contain (non-standardized) acronyms than the MEDLINE corpus. Table 1 shows the differences in distributions over both subcorpora in the training set (note that 'types' refers to unique entities).

Table 1. Statistics on training corpus

	# of sentences	average sent. length	# of entities	# of types	entity/type ratio
MEDLINE	833	12.67	2994	2296	1.30
EMEA	706	23.20	2695	923	2.92

Both corpora contain a various amount of discontinuous and embedded entities, which are known to be more difficult to extract than regular monoword ones. Discontinuous entities are multiword entities which are separated by tokens that are not part of the entity. In Figure 1, the sequence 'synthèse des' is part of the entity 'synthèse des ADN' but it is also linked to 'ARN' and 'protéines'. We found that the number of discontinuous entities in the training set is negligibly low (< 1%) and we therefore did not implement any preprocessing steps to deal with this problem. Embedded entities are a sub-part of a larger entity (like 'hépatiques' and 'cancers' in Figure 2). Their number is quite high: 16% of the entities in the training set are embedded in one or more entities. We decided to tackle this problem by introducing a two-level extraction method, as described in Section 3.

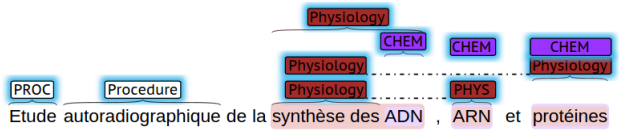

Fig. 1. Example of discontinuous entities

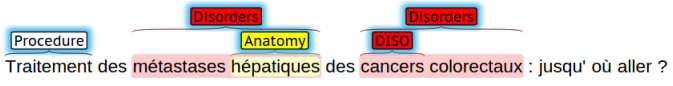

Fig. 2. Example of embedded entities

3 System Overview

Named Entity Recognition is generally considered as a combination of two subproblems: (i) Named entity boundary detection (a.k.a Entity Segmentation), and (ii) Named entity classification. Named entity boundary detection involves correctly recognizing the tokens that form an entity in running free text. Named entity classification deals with the classification of a recognized entity into one of the potential categories. While much research has focused on building systems that solve both problems at the same time [9], we opted to solve the problems in two separate systems. We used a sequential discriminative model to recognize named entities (Conditional Random Fields, hereafter, CRF), and Supervised Vector Machines (from here on, SVM) to label the recognized items with their corresponding Semantic Groups. This set-up allows us to add another CRF to the pipeline that dealt specifically with embedded entities. Figure 3 shows the pipeline used in both the one-level runs and two-level run.

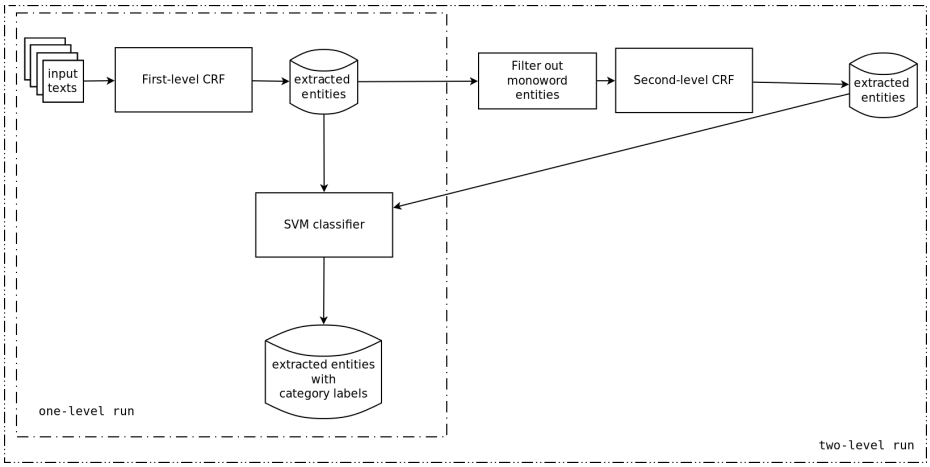

Fig. 3. System pipeline

3.1 Named Entity Boundary Detection

We adopt a two-fold strategy for identifying entity boundaries. A first CRF model is trained on the sentences of the EMEA and MEDLINE training corpora. Its aim is to identify, following the begin-in-out (BIO) tagging scheme [10], the boundaries of multiword entities and monoword ones that are not embedded in a larger entity. Thus, this model would extract entities like 'Traitement', 'métastases hépatiques' and 'cancers colorectaux' out of the sentence in Figure 2.

Then, a second CRF is trained to extract smaller entities that are embedded in multiword entities. This model aims to identify 'hépatiques' and 'cancers' in the multiword entities extracted by the first CRF. The main difference with the first-level CRF is the

context. The second CRF does not take the context of the whole sentence into account but is trained only on the multiword named entities in the training set. The intuition behind the inclusion of a second level CRF is that a model trained on phrase-level only will be better suited to capture which entities should be extracted from complex entities, i.e. mainly the head (noun) or preferably the modifying elements in the noun phrase?

Both models were trained using the Wapiti toolkit [11] on the same following features:

Token/POS bigrams and trigrams The value of the feature is the part of speech assigned to the token by TreeTagger [12]. A post-processing step to simplify the tagset—VER instead of VER:pres for verbs in the present tense—was not found to increase the efficiency of the system.

Token length This feature is the number of characters of the token.

Suffix We define 'suffix' as the four last characters of a token. This well-known pseudo-morphologic approach aims at identifying medical suffixes like '-aire' ('tissulaire', 'pulmonaire') or '-tion' ('maturation', 'évolution') and has been used with success by [13].

Stopwords The presence of a word in the list of stopwords used by Apache Solr is used as a binary feature. This list contains French pronouns, prepositions, conjunctions and the different forms of the auxiliaries 'être' and 'avoir'. The aim of this feature is to distinguish between potential entity-beginning words and others (we can safely assume that the words in the stopwords list belong to the latter category).

UMLS Obviously, the presence of the token in a medical thesaurus like the UMLS [1] is a strong indicator of its interest for our system. Given that a lot of short words (such as acronyms) tend to be ambiguous, we only applied this binary feature to words whose length is over four characters.

Head/modifier frequency We extracted head/modifier frequency information over the EMEA training corpus using the term extractor YaTeA [14]. Each of the two features, i.e. *head* and *modifier*, can take 4 values according to whether the frequency is 0, 1, 2 or superior or equal to 3. The aim of these features is to distinguish between the words that tend to be used as entity heads (e.g. 'effet', 'apparition', 'aggravation'), modifiers (e.g. 'clinique', 'cérébral', 'sévère') or both (e.g. 'traitement', 'solution', 'douleur').

Word shape This orthographical feature can take three values: 'AA' if the word is in uppercase, 'aa' if the word is in lowercase, and 'Aa' if only the first letter is in uppercase.

Number/punctuation checking This feature captures the presence of numbers and/or punctuation marks in the token.

3.2 Named Entity Classification

In a second step, the recognized entities were classified into one of ten possible Semantic Groups. The Wapiti output was processed to extract the recognized entities (both mono- and multiword). Since a vast majority of the entities in the training set (> 99.6%) only had one Semantic Group label, we considered this to be a monolabel classification task. For this task, we applied a Sequential Minimal Optimization (SMO) classifier as implemented in the Weka toolkit.[2] We used the SMO classifier with default parameters.

The set of machine learning features used by our SVM consists primarily of dictionary-lookups in various resources, rather than orthographic or lexical features. This choice of features was motivated by the abstract nature of the Semantic Group categories. These are very high-level categories, e.g. Chemicals & Drugs (CHEM) includes names of medications such as 'Refludan' but also generic words such as 'eau' (as an entity of organic chemistry) and chemical names, e.g. '1-Éthyl-3-(3-diméthylaminopropyl)-carbodiimide'. Features based on orthography such as presence of uppercase characters, presence of hyphens or word length did not appear to have much impact on the training set and were discarded for the official run. All classifiers were trained on the combination of the EMEA and MEDLINE training corpora.

Preprocessing Since the terminology in external sources often only contain base forms, without other surface variations, we added a preprocessing step to expand coverage of the used sources. We generated one or more normalized copies of each extracted entity that needed to be classified. To achieve this, we used information on spelling and surface form variation of French medical terms from the Unified Medical Lexicon for French (UMLF) [15]. In addition, acronyms were resolved using a list of known medical acronyms extracted from Wikipedia and the UMLS. An acronym recognition algorithm [16] was applied to obtain variants of UMLS entities with the semantic group Disorder (DISO). For example, for the term 'intoxication par lysergide (LSD)' (CUI C0274688), both variants with the acronym ('intoxication par LSD') and its expanded form ('intoxication par lysergide') were generated. When looking up the terms in the external resources, we used the different variants of the recognized entity in a back-off method. If the (longest) normalized variant was not found in the dictionary or list, a shorter variant was used. The non-normalized, original terms were used as the last option. The Named Entity classifiers used the following set of features:

Entity length Number of words in the recognized entity

Presence of word in training set The annotated entities in the training set were split up into tokens and their associated Semantic Group categories. We included a binary feature in the SVM that captured whether a token in a recognized entity had appeared in the training set as well. While some words were associated with multiple Semantic Groups, e.g. 'poumon' as Anatomy (ANAT) as well as Disorder (DISO) (from 'cancer du poumon'), we found that this feature captured the head nouns fairly well. Due to an encoding bug in the pipeline, this feature was not used in the classification of entities for the official submitted run, but was included in an additional run to gauge its impact (see Tables 2 and 3).

[2] http://www.cs.waikato.ac.nz/ml/weka/

Presence in ICD10 A binary feature that captures if the recognized entity features in the list of diseases included in the French version of the ICD10; the list was extracted from the UMLS.

Presence in Doctissimo disease list A binary feature that captures if the recognized entity features in the list of diseases extracted from Doctissimo. [3] We opted to include this additional list as it contains more common place names of frequently occurring diseases, e.g. 'La grippe' versus 'influenza'.

Presence in list of drugs A binary feature that captures if the recognized entity features in a list of known medications. This was a compilation of UMLS entities with the semantic type Pharmacological substance (T121), drug names in the VIDAL database[4] and MeSH entities with the descriptor Therapeutic Uses (D27.505.954).

Presence in a list of anatomical terms A binary feature that captures if the recognized term features in a list of anatomical terms extracted from the French part of the UMLS.

Semantic Type label in UMLF The UMLF contains a list of 24,480 CUIs (some with spelling variants) with links to their Semantic Types in the UMLS. This list was extracted from the UMLS during the creation of the UMLF and manually checked. The semantic types of the entities that featured in the list were mapped unto their corresponding groups.

Semantic Group label in CiSMeF portal CISMeF (Catalogue et Index des Sites Médicaux en langue Française) is a quality-controlled health gateway that combines—amongst other information sources—the existing terminologies of French medical texts. We queried its database online (which includes the automatic redirects incorporated in the CISMeF portal) to see if a given recognized entitiy was identified as a MeSH term in the database, which could be linked to a Semantic Group.

4 Run Descriptions and Results

In this section we present the results of our officially submitted run, and two additional runs that were performed after the competition deadline.

One-level official run Due to a run-time bug in our pipeline, we were not able to submit a two-level CRF run for official evaluation. Instead, the system of our official submitted run only contains the first CRF (see Figure 3), and consequently, in this run the embedded entities were ignored. Please note that in this run the *Presence of word in training set feature* is not present.

One-level run with lexical information on training corpus This run is identical to the officially submitted run with the addition of the *Presence of word in training set feature* feature as classification features for the SVM during Named Entity Classification.

[3] doctissimo.com
[4] www.vidal.fr

Two-level run This run utilizes the two two-level CRF approach and the same SVM classifier as was used for the *One-level run with lexical information on training corpus*.

Scores were calculated using the evaluation tool that was made available by the track organizers. The evaluation scores below capture both boundary detection and Named Entity classification at the same time. Two entities match when they to agree on the entity type and on the span of text, either by an *exact match* or by an overlap span match *inexact match*.

Table 2. Results on the EMEA test corpus.

	Exact match			Inexact match		
	Precision	Recall	F1	Precision	Recall	F1
One-level run (Submitted)	0.59	0.42	0.49	0.67	0.50	0.57
One-level additional run	**0.79**	0.56	0.65	0.87	0.65	0.74
Two-level additional run	0.78	**0.61**	**0.69**	0.87	0.70	0.77
Average scores	0.31	0.33	0.31	0.48	0.52	0.49
Median scores	0.21	0.18	0.22	0.58	0.55	0.55

Table 3. Results on the MEDLINE test corpus.

	Exact match			Inexact match		
	Precision	Recall	F1	Precision	Recall	F1
One-level run (Submitted)	0.57	0.38	0.45	0.72	0.54	0.62
One-level additional run	**0.61**	0.40	0.48	0.75	0.56	0.64
Two-level additional run	0.59	**0.49**	**0.54**	0.74	0.64	0.69
Average scores	0.35	0.50	0.40	0.52	0.72	0.58
Median scores	0.39	0.59	0.45	0.59	0.79	0.67

5 Discussion

While the evaluated systems generally scored quite high, we could see a marked difference in the performances for the EMEA corpus compared to the MEDLINE corpus. Overall, our systems dealt better with the long sentences and high frequencies in the training data of the EMEA corpus, and struggled more with the short phrases and low token/type ratio in the MEDLINE corpus. This was evidenced by the differences in accuracy scores between the submitted and additional one-level runs. For the EMEA corpus, adding lexical information from the training corpus led to a 0.16 points gain, compared to a 0.3 points gain for MEDLINE. An error analysis of the CRFs on the test

data (see Tables 4 and 5)[5] also showed the difficulties of the high-level CRF (i.e. CRF1) in recognizing entity boundaries in the short contexts of article titles. The addition of the second CRF (CRF2)[6] had more impact in the MEDLINE domain, as it generally had more embedded entities. In both domains we saw a marked increase in Recall while Precision dropped little.

Table 4. Results of the CRF entity recognition in the EMEA test corpus.

	Exact match			Inexact match		
	Precision	Recall	F1	Precision	Recall	F1
CRF1	0.82	0.58	0.68	0.91	0.73	0.82
CRF1+2	0.80	0.63	0.71	0.92	0.75	0.83

Table 5. Results of the CRF entity recognition in the MEDLINE test corpus.

	Exact match			Inexact match		
	Precision	Recall	F1	Precision	Recall	F1
CRF1	0.67	0.44	0.53	0.86	0.71	0.78
CRF1+2	0.65	0.54	0.59	0.86	0.78	0.82

An error analysis of the SVM classifier[7] on the test sets showed that the accuracy of the classifiers is 87% and 82% for the EMEA and MEDLINE test set, respectively. A visual inspection of the output did not yield consistent misclassifications. Overall, we found that performance of classifiers was quite high, especially considering that most features depended on external resources whose coverage is usually not perfect. Resources available for French, despite being scarce (when compared to those for English), seem to have an adequate coverage for classification purposes.

6 Conclusion and Future Perspectives

This paper presented our participation in the first subtask of task 1b of 2015 CLEF eHealth challenge, which focussed on Clinical Named Entity Recognition in French medical texts. We constructed a hybrid system in which a combination of CRFs performed entity boundary detection, and a SVM module classified the recognized entities. Special attention was given to the problem of embedded entities in a non-submitted run.

[5] These scores were obtained by running the evaluation software over copies of the run and reference files in which the category labels had been changed to just dummy category. Consequently, in these tables only entity boundaries are evaluated.

[6] This is the difference between the *One-level additional run* and the *Two-level additional run*.

[7] This was performed by running the classifier over *all* entities in the reference tests and can thus be seen as classification accuracy after a perfect entity boundary system.

We found that our high-level CRF had difficulties with the short contexts in the MEDLINE corpus. In future work we will examine how training data selection may improve performance for this domain. Furthermore, we found that the existing resources for medical French have a large enough coverage to be adequately for Named Entity Classification.

In a continuation of the task it would be interesting to tackle the problem of discontinuous entities. Another route for exploration is to use multilingual sources to expand the coverage of the knowledge sources even further.

Acknowledgments

This work was supported by the French National Agency for Research under grant Accordys[8] ANR-12-CORD-0007.

References

1. Bodenreider, O. (2004) The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic acids research 32, no. suppl 1: D267-D270.
2. Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011) 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. Journal of the American Medical Informatics Association 18, no. 5: 552-556.
3. Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004, August). Introduction to the Bio-entity Recognition Task at JNLPBA. In Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (pp. 70-75). Association for Computational Linguistics.
4. Aronson, A. R. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association, 2001.
5. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., and 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, September 2015.
6. 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. CLEF 2015 - 6th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September 2015.
7. Bodenreider, O., and McCray, A. T. (2003). Exploring Semantic Groups through Visual Approaches. Journal of biomedical informatics, 36(6), 414-432.
8. Néveol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014). The Quaero French medical corpus: A Resource for Medical Entity Recognition and Normalization. Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, 24-30.
9. Leaman, R., and Gonzalez, G. (2008, January). BANNER: An Executable Survey of Advances in Bio-medical Named Entity Recognition. In Pacific Symposium on Biocomputing (Vol. 13, pp. 652-663).

[8] Agrégation de Contenus et de COnnaissances pour Raisonner à partir de cas dans la DYSmorphologie foetale

10. Ramshaw L. A. and Marcus M. P. (1995), Text Chunking using Transformation-based Learning. Proceedings of the Third ACL Workshop on Very Large Corpora, pp. 82-94.
11. Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 504-513). Association for Computational Linguistics.
12. Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
13. Dimililer, N., and Varoğlu, E. (2006). Recognizing Biomedical Named Entities using SVMs: improving recognition performance with a minimal set of features. In Knowledge Discovery in Life Science Literature (pp. 53-67). Springer Berlin Heidelberg.
14. Aubin, S., Hamon, T. (2006) Improving Term Extraction with Terminological Resources. In Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T., eds.: Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006). LNAI 4139, Springer, 380-387
15. Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch P., Le Duff, F., Forget, J.-F., Douyère, M., and Darmoni, S. (2005) UMLF: A Unified Medical Lexicon for French. International Journal of Medical Informatics 74, 2, 119-124.
16. Schwartz, A. S. and Hearst, M. S. (2003). A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. Pacific Symposium on Biocomputing, 2003, 8, 451-62