Medical Entity Recognition: A Comparison of Semantic and Statistical Methods

Asma Ben Abacha
LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France
asma.benabacha@limsi.fr

Pierre Zweigenbaum
LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France
pz@limsi.fr

Abstract

Medical Entity Recognition is a crucial step towards efficient medical texts analysis. In this paper we present and compare three methods based on domain-knowledge and machine-learning techniques. We study two research directions through these approaches: (i) a first direction where noun phrases are extracted in a first step with a chunker before the final classification step and (ii) a second direction where machine learning techniques are used to identify simultaneously entities boundaries and categories. Each of the presented approaches is tested on a standard corpus of clinical texts. The obtained results show that the hybrid approach based on both machine learning and domain knowledge obtains the best performance.

1 Introduction

Medical Entity Recognition (MER) consists in two main steps: (i) detection and delimitation of phrasal information referring to medical entities in textual corpora (e.g. *pyogenic liver abscess*, *infection of biliary system*) and (ii) identification of the semantic category of located entities (e.g. Medical Problem, Test). Example 1 shows the result of MER on a sentence where the located entity and its category are marked with *treatment* and *problem* tags.

(1) *<treatment> Adrenal-sparing surgery </treatment> is safe and effective , and may become the treatment of choice in patients with <problem> hereditary phaeochromocytoma </problem>.*

This task is very important for many applications such as Question-Answering where MER is used in the question analysis step (to determine the expected answers' type, the question focus, etc.) and in the offline text tagging or annotation.

One of the most important obstacles to identifying medical entities is the high terminological variation in the medical domain (e.g. *Diabetes mellitus type 1*, *Type 1 diabetes*, *IDDM*, or *juvenile diabetes* all express the same concept). Other aspects also have incidence on MER processes such as the evolution of entity naming (e.g. new abbreviations, names for new drugs or diseases). These obstacles limit the scalability of methods relying on dictionaries and/or gazetteers. Thus, it is often the case that other types of approaches are developed by exploiting not only domain knowledge but also domain-independent techniques such as machine learning and natural language processing tools.

In this paper, we study MER with three different methods: (i) a semantic method relying on MetaMap (Aronson, 2001) (a state-of-the-art tool for MER) (ii) chunker-based noun phrase extraction and SVM classification and (iii) a last method using supervised learning with Conditional Random Fields (CRF), which is then combined with the semantic method. With these methods we particularly study two processing directions: (i) pre-extraction of noun phrases with specialized tools, followed by a medical classification step and (ii) exploitation of machine-learning techniques to detect simultaneously entity boundaries and their categories.

We also present a comparative study of the performance of different noun phrase chunkers on medical

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 56–64,
Portland, Oregon, USA, June 23-24, 2011.

texts: Treetagger-chunker, OpenNLP and MetaMap. The best chunker was then used to feed some of the proposed MER approaches. All three methods were experimented on the i2b2/VA 2010 challenge corpus of clinical texts (Uzuner, 2010). Our study shows that hybrid methods achieve the best performance w.r.t machine learning approaches or domain knowledge-based approaches if applied separately.

After a review of related work (Section 2), we describe the chunker comparison and the three MER methods (Section 3). We present experiments on clinical texts (Section 4), followed by a discussion and variant experiments on literature abstracts (Section 5), then conclude and draw some perspectives for further work (Section 6).

2 Related Work

Several teams have tackled named entity recognition in the medical domain. (Rindflesch et al., 2000) presented the EDGAR system which extracts information about drugs and genes related to a given cancer from biomedical texts. The system exploits the MEDLINE database and the UMLS. Protein name extraction has also been studied through several approaches (e.g. (Liang and Shih, 2005; Wang, 2007)). (Embarek and Ferret, 2008) proposed an approach relying on linguistic patterns and canonical entities for the extraction of medical entities belonging to five categories: Disease, Treatment, Drug, Test, and Symptom. Another kind of approach uses domain-specific tools such as MetaMap (Aronson, 2001). MetaMap recognizes and categorizes medical terms by associating them to concepts and semantic types of the UMLS Metathesaurus and Semantic Network. (Shadow and MacDonald, 2003) presented an approach based on MetaMap for the extraction of medical entities of 20 medical classes from pathologist reports. (Meystre and Haug, 2005) obtained 89.9% recall and 75.5% precision for the extraction of medical problems with an approach based on MetaMap Transfer (MMTx) and the NegEx negation detection algorithm.

In contrast with semantic approaches which require rich domain-knowledge for rule or pattern construction, statistical approaches are more scalable. Several approaches used classifiers such as decision trees or SVMs (Isozaki and Kazawa, 2002). Markov models-based methods are also frequently used (e.g. Hidden Markov Models, or CRFs (He and Kayaalp, 2008)). However, the performance achieved by such supervised algorithms depends on the availability of a well-annotated training corpus and on the selection of a relevant feature set.

Hybrid approaches aim to combine the advantages of semantic and statistical approaches and to bypass some of their weaknesses (e.g. scalability of rule-based approaches, performance of statistical methods with small training corpora). (Proux et al., 1998) proposed a hybrid approach for the extraction of gene symbols and names. The presented system processed unknown words with lexical rules in order to obtain candidate categories which were then disambiguated with Markov models. (Liang and Shih, 2005) developed a similar approach using empirical rules and a statistical method for protein-name recognition.

3 Medical Entity Recognition Approaches

Named entity recognition from medical texts involves two main tasks: (i) identification of entity boundaries in the sentences and (ii) entity categorization. We address these tasks through three main approaches which are listed in Table 1.

3.1 Noun Phrase Chunking

Although noun phrase segmentation is an important task for MER, few comparative studies on available tools have been published. A recent study (Kang et al., 2010), which claims to be the first to do such comparative experiments, tested six state-of-the-art chunkers on a biomedical corpus: GATE chunker, Genia Tagger, Lingpipe, MetaMap, OpenNLP, and Yamcha. This study encompassed sentence splitting, tokenization and part-of-speech tagging and showed that for both noun-phrase chunking and verb-phrase chunking, OpenNLP performed best (F-scores 89.7% and 95.7%, respectively), but differences with Genia Tagger and Yamcha were small.

With a similar objective, we compared the performance of three different noun-phrase chunkers in the medical domain: (i) Treetagger-chunker[1], a state-of-the-art open-domain tool, (ii) OpenNLP[2] and (iii)

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

[2] http://incubator.apache.org/opennlp

	Medical Entity Recognition	
	1. Boundary identification	**2. Type categorization (with *n* medical entity categories)**
Method 1 (MetaMap+)	Noun phrase segmentation	- Rule-based method, - *Noun phrase* classification, - Number of classes = $n + 1$
Method 2 (TT-SVM)	Noun phrase segmentation	- Statistical method with a SVM classifier, - *Noun phrase* classification, - Number of classes = $n + 1$
Method 3 (BIO-CRF)	- Statistical method with a CRF classifier, - and the BIO format, - *word-level* classification, - Number of classes = $2n + 1$	

Table 1: Proposed MER methods

	Corpus of clinical texts (i2b2)			Corpus of scientific abstracts (Berkeley)		
	MetaMap	**TreeTagger**	**OpenNLP**	**MetaMap**	**TreeTagger**	**OpenNLP**
Reference entities	58115	58115	58115	3371	3371	3371
Correct entities	6532	35314	26862	151	2106	1874
Found entities	212227	129912	122131	22334	19796	18850
Recall	11.14%	**60.06%**	46.62%	4.48%	**62.27%**	55.59%

Table 2: NP Segmentation Results

MetaMap. Regardless of the differences in corpora with (Kang et al., 2010) we chose these particular tools to compare medical-domain specific tools with open domain tools and to highlight the lower performance of MetaMap for noun-phrase chunking compared to other tools. This last point led us to introduce the MetaMap+ approach for MER (Ben Abacha and Zweigenbaum, 2011) in order to take advantage of MetaMap's domain-knowledge approach while increasing performance by relying on external tools for noun-phrase chunking.

We evaluate these tools on the subset of noun phrases referring to medical entities in our corpora (cf. Section 4.1 for a description of the i2b2 corpus and Section 5 for the Berkeley corpus). We consider that a noun phrase is correctly extracted if it corresponds exactly to an annotated medical entity from the reference corpora. Also, as our corpora are not fully annotated (only entities of the targeted types are annotated), we do not evaluate "extra noun-phrases" corresponding to non-annotated entities. The retrieved noun phrases are heterogeneous: many of them are not all relevant to the medical field and therefore not relevant to the MER task. Our goal is to obtain the maximal number of correct noun phrases and leave it to the next step to filter out those that are irrelevant. We therefore wish to maximize recall at this stage.

Table 2 shows that in this framework, Treetagger-chunker obtains the best recall. We thus used it for noun-phrase segmentation in the experimented MER approaches (cf. Sections 3.2 and 3.3).

3.2 Semantic and Rule-Based Method: MM+

MetaMap is a reference tool which uses the UMLS to map noun phrases in raw texts to the best matching UMLS concepts according to matching scores. MetaMap leads however to some residual problems, which we can arrange into three classes: (i) noun phrase chunking is not at the same level of performance as some specialized NLP tools, (ii) medical entity detection often retrieves general words and verbs which are not medical entities and (iii) some ambiguity is left in entity categorization since MetaMap can provide several concepts for the same term as well as several semantic types for the same concept. Several "term/concept/type" combinations

are then possible.

To improve MetaMap output, we therefore use an external noun phrase chunker (cf. Section 3.1) and stop-list based filtering to recover frequent/noticeable errors. MetaMap can propose different UMLS semantic types for the same noun phrase, thus leading to different categories for the same entity. In such cases we apply a voting procedure. For instance, if the process retrieves three UMLS semantic types for one noun phrase where two are associated to the target category "Problem" and one is associated to "Treatment", the "Problem" category is chosen as the entity's category. In case of a tie, we rely on the order output by MetaMap and take the first returned type.

More precisely, our rule-based method, which we call MetaMap+ (MM+), can be decomposed into the following steps:

1. Chunker-based noun phrase extraction. We use Treetagger-chunker according to the above-mentioned test (cf. Table 2).
2. Noun phrase filtering with a stop-word list.
3. Search for candidate terms in specialized lists of medical problems, treatments and tests gathered from the training corpus, Wikipedia, Health on the Net and Biomedical Entity Network.
4. Use MetaMap to annotate medical entities (which were not retrieved in the specialized lists) with UMLS concepts and semantic types.
5. Finally, filter MetaMap results with (i) a list of common/noticeable errors and (ii) the selection of only a subset of semantic types to look for (e.g. Quantitative Concept, Functional Concept, Qualitative Concept are too general semantic types and produce noise in the extraction process).

3.3 Statistical Method: TT-SVM

The second presented approach uses Treetagger-chunker to extract noun phrases followed by a machine learning step to categorize medical entities (e.g. Treatment, Problem, Test). The problem is then modeled as a supervised classification task with $n + 1$ categories (n is the number of entity categories). We chose an SVM classifier.

As noted by (Ekbal and Bandyopadhyay, 2010), SVMs (Support Vector Machines) have advantages over conventional statistical learning algorithms, such as Decision Trees or Hidden Markov Models, in the following two aspects: (1) SVMs have high generalization performance independent of the dimension of feature vectors, and (2) SVMs allow learning with all feature combinations without increasing computational complexity, by introducing kernel functions.

In our experiments we use the libSVM (Chang and Lin, 2001) implementation of the SVM classifier. We chose the following feature set to describe each noun phrase (NP):

1. Word features:
   - words of the NP
   - number of the NP words
   - lemmas of the NP words
   - 3 words and their lemmas before the NP
   - 3 words and their lemmas after the NP
2. Orthographic features (some examples):
   - first letter capitalized for the first word, one word or all words
   - all letters uppercase for the first word, one word or all words
   - all letters lowercase for the first word, one word or all words
   - NP is or contains an abbreviation
   - word of NP contains a single uppercase, digits, hyphen, plus sign, ampersand, slash, etc.
3. Part-of-speech tags: POS tags of the NP words, of the 3 previous and 3 next words.

3.4 Statistical Method: BIO-CRF

We conducted MER with a CRF in one single step by determining medical categories and entity boundaries at the same time. We used the BIO format: B (beginning), I (inside), O (outside) which represents entity tagging by individual word-level tagging. For instance, a problem-tagged entity is represented as a first word tagged B-P (begin problem) and other

(following) words tagged I-P (inside a problem). A problem entity comprising one single word will be tagged B-P. Words outside entities are tagged with the letter 'O'.

If we have n categories (e.g. Problem, Treatment, Test), we then have n classes of type B-, n classes of type I- (e.g. P-B and P-I classes associated to the *problem* category) and one class of type 'O'. Figure 1 shows an example sentence tagged with the BIO format. As a result, the classification task consists in a word classification task (instead of a noun-phrase classification task) into $2n + 1$ target classes, where n is the number of categories. As a consequence, relying on a chunker is no longer necessary.

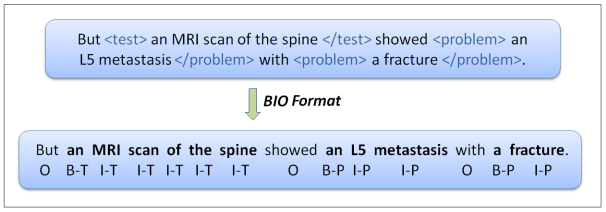

Figure 1: BIO Format (T = Test, P = Problem)

Words in a sentence form a sequence, and the decision on a word's category can be influenced by the decision on the category of the preceding word. This dependency is taken into account in sequential models such as Hidden Markov Models (HMMs) or Conditional Random Fields (CRF). In contrast with HMMs, CRF learning maximizes the conditional probability of classes w.r.t. observations rather than their joint probability. This makes it possible to use any number of features which may be related to all aspects of the input sequence of words. These properties are assets of CRFs for several natural language processing tasks, such as POS tagging, noun phrase chunking, or named entity recognition (see (Tellier and Tommasi, 2010) for a survey).

In our experiments we used the CRF++[3] implementation of CRFs. CRF++ eases feature description through *feature templates*. We list hereafter some of our main features. We instructed CRF++ to model the dependency of successive categories (instruction `B` in feature template).

For each word we use the following features:

1. Word features: The word itself, two words before and three words after, with their lemmas.
2. Morphosyntactic features: POS tag of the word itself, two words before and three words after.
3. Orthographic features (some examples):
   - The word contains hyphen, plus sign, ampersand, slash, etc.
   - The word is a number, a letter, a punctuation sign or a symbol.
   - The word is in uppercase, capitalized, in lowercase (AA, Aa, aa)
   - Prefixes of different lengths (from 1 to 4)
   - Suffixes of different lengths (from 1 to 4)
4. Semantic features: semantic category of the word (provided by MetaMap+)
5. Other features: next verb, next noun, word length over a threshold, etc.

Additionally, we tested semantic features constructed from MM+ results. More detail on these last features is given in Section 5.3.

4 Experiments on Clinical Texts

We performed MER experiments on English clinical texts.

4.1 Corpus

The i2b2 corpus was built for the i2b2/VA 2010 challenge[4] in Natural Language Processing for Clinical Data (Uzuner, 2010). The data for this challenge includes discharge summaries from Partners HealthCare and from Beth Israel Deaconess Medical Center (MIMIC II Database), as well as discharge summaries and progress notes from University of Pittsburgh Medical Center. All records have been fully de-identified and manually annotated for concept, assertion, and relation information. The corpus contains entities of three different categories: Problem, Treatment and Test, 76,665 sentences and 663,476 words with a mean of 8.7 words per sentence. Example 2 shows an annotated sentence from the i2b2 corpus.

[3] `http://crfpp.sourceforge.net/`

[4] `http://www.i2b2.org/NLP/Relations/`

(2) *<problem>CAD</problem> s/p <treatment>3v-CABG </treatment> 2003 and subsequent <treatment>stenting </treatment> of <treatment>SVG</treatment> and LIMA.*

Table 3 presents the number of training and test sentences.

i2b2 Corpus	Sentences	Words
Training Corpus	31 238	267 304
Test Corpus	44 927	396 172

Table 3: Number of training and test sentences

4.2 Experimental Settings

We tested the above-described five configurations (see Table 1):

1. MM: MetaMap is applied as a baseline method
2. MM+: MetaMap Plus (semantic and rule-based method)
3. TT-SVM: Statistical method, chunking with Treetagger and Categorization with a SVM classifier
4. BIO-CRF: Statistical method, BIO format with a CRF classifier
5. BIO-CRF-H: Hybrid method combining semantic and statistical methods (BIO-CRF with semantic features constructed from MM+ results)

We evaluate the usual metrics of Recall (proportion of correctly detected entities among the reference entities), Precision (proportion of correctly detected entities among those output by the system), and F-measure (harmonic means of Recall and Precision).

4.3 Results

Table 4 presents the results obtained by each configuration. BIO-CRF and BIO-CRF-H obtained the best precision, recall and F-measures. MM+ comes next, followed by TT-SVM and MetaMap alone.

Table 5 presents the obtained results per each medical category (i.e. Treatment, Problem and Test) for three configurations. Again, BIO-CRF-H obtains the best results for all metrics and all categories.

Setting	P	R	F
MM	15.52	16.10	15.80
MM+	48.68	56.46	52.28
TT-SVM	43.65	47.16	45.33
BIO-CRF	70.15	83.31	76.17
BIO-CRF-H	**72.18**	**83.78**	**77.55**

Table 4: Results per setting on the i2b2 corpus. R = recall, P = precision, F = F-measure

Setting	Category	P	R	F
MM+	Problem	60.84	53.04	56.67
	Treatment	51.99	61.93	56.53
	Test	56.67	28.48	37.91
TT-SVM	Problem	48.25	43.16	45.56
	Treatment	42.45	50.86	46.28
	Test	57.37	35.76	44.06
BIO-CRF-H	Problem	82.05	73.14	77.45
	Treatment	83.18	73.33	78.12
	Test	87.50	68.69	77.07

Table 5: Results per setting and per category on the i2b2 corpus

5 Discussion and Further Experiments

We presented three different methods for MER: MM+, TT-SVM, and BIO-CRF (with variant BIO-CRF-H). In this section we quickly present supplementary results obtained on a second corpus with the same methods, and discuss differences in results when corpora and methods vary.

5.1 Corpora

Different kinds of corpora exist in the biomedical domain (Zweigenbaum et al., 2001). Among the most recurring ones we may cite (i) clinical texts and (ii) scientific literature (Friedman et al., 2002). Clinical texts have motivated a long stream of research (e.g. (Sager et al., 1995), (Meystre et al., 2008)), and more recently international challenges such as i2b2 2010 (Uzuner, 2010). The scientific literature has also been the subject of much research (e.g. (Rindflesch et al., 2000)), especially in genomics for more than a decade, e.g. through the BioCreative challenge (Yeh et al., 2005).

Section 4 presented experiments in MER on English clinical texts. To have a complementary view on the performance of our methods, we performed additional experiments on the Berkeley corpus (Rosario and Hearst, 2004) of scientific literature abstracts and titles extracted from MEDLINE. The original aim of this corpus was to study the extraction of semantic relationships between problems and treatments (e.g. *cures*, *prevents*, and *side effect*). In our context, we only use its annotation of medical entities. The corpus contains two categories of medical entities: problems (1,660 entities) and treatments (1,179 entities) in 3,654 sentences (74,754 words) with a mean of 20.05 words per sentence. We divided the corpus into 1,462 sentences for training and 2,193 for testing.

We tested the MetaMap (MM), MetaMap+ (MM+) and BIO-CRF methods on the Berkeley corpus. Table 6 presents the results. BIO-CRF again obtain the best results, but it is not much better than MM+ in this case.

		P	R	F
MM	Problem	5.32	7.63	6.27
	Treatment	6.37	18.84	9.52
	Total	5.35	12.34	7.46
MM+	Problem	34.47	44.97	39.02
	Treatment	18.11	39.36	24.81
	Total	23.43	42.47	30.20
BIO-CRF	Problem	41.88	38.88	40.32
	Treatment	29.85	23.86	26.52
	Total	36.94	32.13	34.37

Table 6: Results on the Berkeley Corpus

We constructed three different models for the BIO-CRF method: a first model constructed from the Berkeley training corpus, a second model constructed from the i2b2 corpus and a third model constructed from a combination of the former two. We obtained the best results with the last model: F=34.37% (F=22.97% for the first model and F=30.08% for the second model). These results were obtained with a feature set with which we obtained 76.17% F-measure on the i2b2 corpus (i.e. words, lemmas, morphosyntactic categories, orthographic features, suffixes and prefixes, cf. set A4 in Table 7).

The results obtained on the two corpora are not on the same scale of performance. This is mainly due to the characteristics of each corpus. For instance, the i2b2 2010 corpus has an average words-per-sentence ratio of 8.7 while the Berkeley corpus has a ratio of 20.45 words per sentence. Besides, the i2b2 corpus uses a quite specific vocabulary such as conventional abbreviations of medical terms (e.g. *k/p* for *kidney pancreas* and *d&c* for *dilation and curettage*) and abbreviations of domain-independent words (e.g. *w/o* for *without* and *y/o* for *year old*).

However, according to our observations, the most important characteristic which may explain these results may be the quality of annotation. The i2b2 corpus was annotated according to well-specified criteria to be relevant for the challenge, while the Berkeley corpus was annotated with different rules and less control measures. We evaluated a random sample of 200 annotated medical entities in the Berkeley corpus, using the i2b2 annotation criteria, and found that 20% did not adhere to these criteria.

5.2 Semantic Methods

The semantic methods have the advantage of being reproducible on all types of corpora without a preprocessing or learning step. However, their dependency to knowledge reduces their performance w.r.t. machine learning approaches. Also the development of their knowledge bases is a relatively slow process if we compare it with the time which is necessary for machine learning approaches to build new extraction and categorization models.

On the other hand, a clear advantage of semantic approaches is that they facilitate semantic access to the extracted information through conventional semantics (e.g. the UMLS Semantic Network).

In our experiments we did not obtain good results when applying MetaMap alone. This is mainly due to the detection of entity boundaries (e.g. *"no pericardial effusion."* instead of *"pericardial effusion"* and *"(Warfarin"* instead of *"Warfarin"*).

We were able to enhance the overall performance of MetaMap for this task by applying several input and output filtering primitives, among which the use of an external chunker to obtain the noun phrases. Our observation is that the final results are limited by chunker performance. Nevertheless, the approach provided the correct categories for 52.28% correctly

extracted entities while the total ratio of the retrieved entities with correct boundaries is 60.76%.

5.3 Machine Learning Methods

We performed several tests with semantic features with the BIO-CRF method. For instance, applying MM+ on each word and using the obtained medical category as an input feature for CRF decreased performance from 76.17% F-measure to 76.01%. The same effect was observed by using the UMLS semantic type instead of the final category for each word, with an F-measure decrease from 76.17% to 73.55%. This can be explained by a reduction in feature value space size (22 UMLS types instead of 3 final categories) but also by the reduced performance of MetaMap if it is applied at the word level.

The best solution was obtained by transforming the output of the MM+ approach into BIO format tags and feeding them to the learning process as features for each word. Thus, each word in an entity tagged by MM+ has an input feature value corresponding to one of the following: B-problem, I-problem, B-treatment, I-treatment, B-test and I-test. Words outside entities tagged by MM+ received an 'O' feature value.

With these semantic features we were able to increase the F-measure from 76.19% to 77.55%. Table 7 presents the contribution of each feature category to the BIO-CRF method on the i2b2 corpus.

Features	**P**	**R**	**F**
A1: Words/Lemmas/POS	62.81	82.25	71.23
A2: A1 + orthographic features	63.72	82.19	71.78
A3: A2 + suffixes	67.91	82.89	74.65
A4: A3 + prefixes	70.15	83.31	76.17
A5: A4 + other features	70.22	83.28	76.19
A6: A5 + semantic features	**72.18**	**83.78**	**77.55**

Table 7: Contribution of each feature category (BIO-CRF method) on the i2b2 corpus

6 Conclusion

We presented and compared three different approaches to MER. Our experiments show that performing the identification of entity boundaries with a chunker in a first step limits the overall performance, even though categorization can be performed efficiently in a second step. Using machine learning methods for joint boundary and category identification allowed us to bypass such limits. We obtained the best results with a hybrid approach combining machine learning and domain knowledge. More precisely, the best performance was obtained with a CRF classifier using the BIO format with lexical and morphosyntactic features combined with semantic features obtained from a domain-knowledge based method using MetaMap.

Future work will tackle French corpora with both a semantic method and the BIO-CRF approach. We also plan to exploit these techniques to build a cross-language question answering system. Finally, it would be interesting to try ensemble methods to combine the set of MER methods tested in this paper.

Acknowledgments

This work has been partially supported by OSEO under the Quaero program.

References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *AMIA Annu Symp Proc*, pages 17–21.

Asma Ben Abacha and Pierre Zweigenbaum. 2011. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*. In Press.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Asif Ekbal and Sivaji Bandyopadhyay. 2010. Named entity recognition using support vector machine: A language independent approach. *International Journal of Electrical and Electronics Engineering*, 4(2):155–170.

Mehdi Embarek and Olivier Ferret. 2008. Learning patterns for building resources about semantic relations in the medical domain. In *LREC'08*, May.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35:222–235.

Ying He and Mehmet Kayaalp. 2008. Biological entity recognition with Conditional Random Fields. In *AMIA Annu Symp Proc*, pages 293–297.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of COLING-2002*, pages 390–396.

N Kang, EM van Mulligen, and JA Kors. 2010. Comparing and combining chunkers of biomedical text. *J Biomed Inform*, 44(2):354–360, nov.

Tyne Liang and Ping-Ke Shih. 2005. Empirical textual mining to protein entities recognition from PubMed corpus. In *NLDB'05*, pages 56–66.

Stéphane M. Meystre and Peter J. Haug. 2005. Comparing natural language processing tools to extract medical problems from narrative text. In *AMIA Annu Symp Proc*, pages 525–529.

S M Meystre, G K Savova, K C Kipper-Schuler, and J F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35:128–44.

Denys Proux, François Rechenmann, Laurent Julliard, Violaine Pillet, and Bernard Jacq. 1998. Detecting gene symbols and names in biological texts : A first step toward pertinent information extraction. In *Proceedings of Genome Informatics*, pages 72–80, Tokyo, Japan : Universal Academy Press.

Thomas C. Rindflesch, Lorraine Tanabe, John N. Weinstein, and Lawrence Hunter. 2000. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of Pacific Symposium on Biocomputing*, pages 517–528.

Barbara Rosario and Marti A. Hearst. 2004. Classifying semantic relations in bioscience text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, July.

N Sager, M Lyman, N T Nhàn, and L J Tick. 1995. Medical language processing: applications to patient data representation and automatic encoding. *Meth Inform Med*, 34(1–2):140–6.

G Shadow and C MacDonald. 2003. Extracting structured information from free text pathology reports. In *AMIA Annu Symp Proc*, Washington, DC.

Isabelle Tellier and Marc Tommasi. 2010. Champs Markoviens Conditionnels pour l'extraction d'information. In Éric Gaussier and François Yvon, editors, *Modèles probabilistes pour l'accès à l'information textuelle*. Hermès, Paris.

Özlem Uzuner, editor. 2010. *Working papers of i2b2 Medication Extraction Challenge Workshop*. i2b2.

Xinglong Wang. 2007. Rule-based protein term identification with help from automatic species tagging. In *Proceedings of CICLING 2007*, pages 288–298.

Alexander Yeh, Alexander Morgan, Marc Colosimo, and Lynette Hirschman. 2005. BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics*, 6 Suppl 1.

Pierre Zweigenbaum, Pierre Jacquemart, Natalia Grabar, and Benoît Habert. 2001. Building a text corpus for representing the variety of medical language. In V. L. Patel, R. Rogers, and R. Haux, editors, *Proceedings of Medinfo 2001*, pages 290–294, Londres.