Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task

Natalia Ponomareva, Paolo Rosso, Ferran Pla, Antonio Molina
Universidad Politecnica de Valencia
c/ Camino Vera s/n
Valencia, Spain
{*nponomareva, prosso, fpla, amolina*}@*dsic.upv.es*

Abstract

With a recent quick development of a molecular biology domain the Information Extraction (IE) methods become very useful. Named Entity Recognition (NER), that is considered to be the easiest task of IE, still remains very challenging in molecular biology domain because of the complex structure of biomedical entities and the lack of naming convention. In this paper we apply two popular sequence labeling approaches: Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) to solve this task. We exploit different stategies to construct our biomedical Named Entity (NE) recognizers which take into account special properties of each approach. Although the CRF-based model has obtained much better results in the F-score, the advantage of the CRF approach remains disputable, since the HMM-based model has achieved a greater recall for some biomedical classes. This fact makes us think about a possibility of an effective combination of these models.

Keywords

Biomedical Named Entity Recognition, Conditional Random Fields, Hidden Markov Models

1 Introduction

Recently the molecular biology domain has been getting a massive growth due to many discoveries that have been made during the last years and due to a great interest to know more about the origin, structure and functions of living systems. It causes to appear every year a great deal of articles where scientific groups describe their experiments and report about their achievements.

Nowadays the largest biomedical database resource is MEDLINE that contains more than 14 millions of articles of the world's biomedical journal literature and this amount is constantly increasing - about 1,500 new records per day [1]. To deal with such an enormous quantity of biomedical texts different biomedical resources as databases and ontologies have been created.

Actually NER is the first step to order and structure all the existing domain information. In molecular biology it is used to identify within the text which words or phrases refer to biomedical entities, and then to classify them into relevant biomedical concept classes.

Although NER in molecular biology domain has been receiving attention by many researchers for a decade, the task remains very challenging and the results achieved in this area are much poorer than in the newswire one.

The principal factors that have made the biomedical NER task difficult can be described as follows [11]:

(i) *Different spelling forms existing for one entity* (e.g. "N-acetylcysteine", "N-acetyl-cysteine", "NacetylCysteine").

(ii) *Very long descriptive names.* For example, in the Genia corpus (which will be described in Section 3.1) the significant part of entities has length from 1 to 7.

(iii) *Term share.* Sometimes two entities share the same words that usually are headnouns (e.g. "T and B cell lines").

(iv) *Cascaded entity problem.* There exist many cases when one entity appears inside another one (e.g. $< PROTEIN >< DNA > kappa3 < /DNA > bindingfactor < /PROTEIN >$) that lead to certain difficulties in a true entity identification.

(v) *Abbreviations*, that are widely used to shorten entity names, create problems of its correct classification because they carry less information and appear less times than the full forms.

This paper aims to investigate and compare a performance of two popular Natural Language Processing (NLP) approaches: HMMs and CRFs in terms of their application to the biomedical NER task. All the experiments have been realized using a JNLPBA version of Genia corpus [2].

HMMs [6] are generative models that proved to be very successful in a variety of sequence labeling tasks as Speech recognition, POS tagging, chunking, NER, etc.[5, 12]. Its purpose is to maximize the joint probability of paired observation and label sequences. If, besides a word, its context or another features are taken into account the problem might become intractable. Therefore, traditional HMMs assume an independence of each word from its context that is, evidently, a rather strict supposition and it is contrary to the fact. In spite of these shortcomings the HMM approach offers a number of advantages such as a simplicity, a quick learning and also a global maximization of the joint probability over the whole observation and label sequences. The last statement means that the decision of the best sequence of labels is made after the complete analysis of an input sequence.

CRFs [3] is a rather modern approach that has already become very popular for a great amount of NLP tasks due to its remarkable characteristics [9, 4, 8]. CRFs are indirected graphical models which belong to the discriminative class of models. The principal difference of this approach with respect to the HMM one is that it maximizes a conditional probability of labels given an observation sequence. This conditional assumption makes easy to represent any additional feature that a researcher could consider useful, but, at the same time, it automatically gets rid of the property of HMMs that any observation sequence may be generated.

This paper is organized as follows. In Section 2 a brief review of the theory of HMMs and CRFs is introduced. In Section 3 different strategies of building our HMM-based and CRF-based models are presented. Since corpus characteristics have a great influence on the performance of any supervised machine-learning model the first part of Section 3 is dedicated to a description of the corpus used in our work. In Section 4 the performances of the constructed models are compared. Finally, in Section 5 we draw our conclusions and discuss the future work.

2 HMMs and CRFs in sequence labeling tasks

Let $\mathbf{x} = (x_1 x_2 ... x_n)$ be an observation sequence of words of length n. Let $\mathbf{S}$ be a set of states of a finite state machine each of which corresponds to a biomedical entity tag $t \in T$. We denote as $\mathbf{s} = (s_1 s_2 ... s_n)$ a sequence of states that provides for our word sequence $\mathbf{x}$ some biomedical entity annotation $\mathbf{t} = (t_1 t_2 ... t_n)$.

HMM-based classifier belongs to naive Bayes classifiers which are founded on a joint probability maximization of observation and label sequences:

$$P(\mathbf{s}, \mathbf{x}) = P(\mathbf{x}|\mathbf{s})P(\mathbf{s})$$

In order to provide a tractability of the model traditional HMM makes two simplifications. First, it supposes that each state s_i only depends on a previous one s_{i-1}. This property of stochastic sequences is also called a Markov property. Second, it assumes that each observation word x_i only depends on the current state s_i. With these two assumptions the joint probability of a state sequence $\mathbf{s}$ with observation sequence $\mathbf{x}$ can be represented as follows:

$$P(\mathbf{s}, \mathbf{x}) = \prod_{i=1}^{n} P(x_i|s_i)P(s_i|s_{i-1}) \qquad (1)$$

Therefore, the training procedure is quite simple for HMM approach, there must be evaluated three probability distributions:

(1) initial probabilities $P_0(s_i) = P(s_i|s_0)$ to begin from a state i;

(2) transition probabilities $P(s_i|s_{i-1})$ to pass from a state s_{i-1} to a state s_i;

(3) observation probabilities $P(x_i|s_i)$ of an appearance of a word x_i in a position s_i.

All these probabilities may be easily calculated using a training corpus.

The equation (1) describes a traditional HMM classifier of the first order. If a dependence of each state on two proceding ones is assumed a HMM classifier of the second order will be obtained:

$$P(\mathbf{s}, \mathbf{x}) = \prod_{i=1}^{n} P(x_i|s_i)P(s_i|s_{i-1}, s_{i-2}) \qquad (2)$$

CRFs are undirected graphical models. Although they are very similar to HMMs they have a different nature. The principal distinction consists in the fact that CRFs are discriminative models which are trained to maximize the conditional probability of observation and state sequences $P(\mathbf{s}|\mathbf{x})$. This leads to a great diminution of a number of possible combinations between observation word features and their labels and, therefore, it makes possible to represent much additional knowledge in the model. In this approach the conditional probability distribution is represented as a multiplication of feature functions exponents:

$$P_\theta(\mathbf{s}|\mathbf{x}) = \frac{1}{Z_0} exp\left(\sum_{i=1}^{n}\sum_{k=1}^{m} \lambda_k f_k(s_{i-1}, s_i, \mathbf{x}) + \right. \\ \left. + \sum_{i=1}^{n}\sum_{k=1}^{m} \mu_k g_k(s_i, \mathbf{x})\right) \qquad (3)$$

where Z_0 is a normalization factor of all state sequences, $f_k(s_{i-1}, s_i, \mathbf{x})$, $g_k(s_i, \mathbf{x})$ are feature functions and λ_k,μ_k are learning weights of each feature function. Although, in general, feature functions can belong to any family of functions, we consider the simplest case of binary functions.

Comparing equations (1) and (3) there may be seen a strong relation between HMM and CRF approaches: feature functions f_k together with its weights λ_k are some analogs of transition probabilities in HMMs while functions $\mu_k f_k$ are observation probability analogs. But in contrast to the HMMs, the feature functions of CRFs may not only depend on the word itself but on any word feature, which is incorporated into the model. Moreover, transition feature functions may also take into account both a word and its features as, for instance, a word context.

A training procedure of the CRF approach consists in the weight evaluation in order to maximize a conditional log likelihood of annotated sequences for some training data set $D = (\mathbf{x}, \mathbf{t})^{(1)}, (\mathbf{x}, \mathbf{t})^{(2)}, ..., (\mathbf{x}, \mathbf{t})^{(|D|)}$

$$L(\theta) = \sum_{j=1}^{|D|} logP_\theta(\mathbf{t}^{(j)}|\mathbf{x}^{(j)})$$

We have used CRF++ open source [1] which implemented a quasi-Newton algorithm called LBFGS for the training procedure.

[1] http://www.chasen.org/ taku/software/CRF++/

3 Biomedical NE recognizers description

Biomedical NER task consists in the detecting in a raw text biomedical entities and assigning them to one of the existing entity classes. In this section the two biomedical NE recognizers, we constructed, based on the HMM and CRF approaches will be described.

3.1 JNLPBA corpus

Any supervised machine-based model depends on a corpus that has been used to train it. The greater and the more complete the training corpus is, the more precise the model will be and, therefore, the better results can be achieved. At the moment the largest and, therefore, the most popular biomedical annotated corpus is Genia corpus v. 3.02 which contains 2,000 abstracts from the MEDLINE collection annotated with 36 biomedical entity classes. To construct our model we have used its JNLPBA version that was applied in the JNLPBA workshop in 2004 [2]. In Table 1 the main characteristics of the JNLPBA training and test corpora are illustrated.

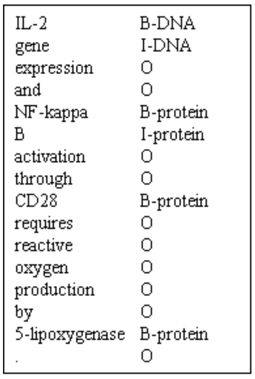
Table 1: *JNLPBA corpus characteristics*

Characteristics	Training corpus	Test corpus
Number of abstracts	2,000	404
Number of sentences	18,546	3,856
Number of words	492,551	101,039
Number of biomed. tags	109,588	19,392
Size of vocabulary	22,054	9,623
Years of publication	1990-1999	1978-2001

The JNLPBA corpus is annotated with 5 classes of biomedical entities: protein, RNA, DNA, cell type and cell line. Biomedical entities are tagged using the IOB2 notation that consists of 2 parts: the first part indicates whether the corresponding word appears at the beginning of an entity (tag B) or in the middle of it (tag I); the second part refers to the biomedical entity class the word belongs to. If the word does not belong to any entity class it is annotated as "O". In Fig. 1 an extract of the JNLPBA corpus is presented in order to illustrate the corpus annotation. In Table 2 a tag distribution within the corpus is shown. It can be seen that the majority of words (about 80%) does not belong to any biomedical category. Furthermore, the biomedical entities themselves also have an irregular distribution: the most frequent class (protein) contains more than 10% of words, whereas the most rare one (RNA) only 0.5% of words. The tag irregularity may cause a confusion among different types of entities with a tendency for any word to be referred to the most numerous class.

Table 2: *Entity tag distribution in the training corpus*

Tag name	Protein	DNA	RNA	cell type	cell line	no-entity
Tag distr.%	11.2	5.1	0.5	3.1	2.3	77.8



IL-2	B-DNA
gene	I-DNA
expression	O
and	O
NF-kappa	B-protein
B	I-protein
activation	O
through	O
CD28	B-protein
requires	O
reactive	O
oxygen	O
production	O
by	O
5-lipoxygenase	B-protein
.	O

Fig. 1: *Example of the JNLPBA corpus annotation*

3.2 Feature set

As it is rather difficult to represent in HMMs a rich set of features and in order to be able to compare HMM and CRF models under the same conditions we have not applied such commonly used features as orthografic or morphological ones. The only additional information we have exploited are parts-of-speech (POS) tags.

The set of POS tags was supplied by the Genia Tagger[2]. It is significant that this tagger was trained on the Genia corpus in order to provide better results in the biomedical texts annotation. As it has been shown by [12], the use of the POS tagger adapted to the biomedical task may greatly improve the performance of the NER system than the use of the tagger trained on any general corpus as, for instance, Penn TreeBank.

3.3 Two different strategies to build HMM-based and CRF-based models

As we have already mentioned, CRFs and HMMs have principal differences and, therefore, distint methodologies should be employed in order to construct the biomedical NE recognizers based on these models.

Due to their structure, HMMs cause certain inconviniences for feature set representation. The simplest way to add a new knowledge into the HMM model is to specialize its states. This strategy was previously applied to other NLP tasks, such as POS tagging, chunking or clause detection and proved to be very effective [5].

Thus, we have employed this methodology for the construction of our HMM-based biomedical NE recognizer. States specialization leads to the increasing of a number of states and to adjusting each of them to certain categories of observations. In other words, the idea of specialization may be formulated as a spliting of states by means of additional features which in our case are POS tags.

In our HMM-based system the specialization strategy using POS information serves both to provide an additional knowledge about entity boundaries and to diminish an entity class irregularity. As we have seen

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

in Section 3.1, the majority of words in the corpus does not belong to any entity class. Such data irregularity can provoke errors, which are known as false negatives, and, therefore, may diminish the recall of the model. It means that many biomedical entities will be classified as non-entity. Besides, there also exists a non-uniform distribution among biomedical entity classes: e.g. class "protein" is more than 100 times larger than class "RNA" (see Table 2).

We have constructed three following models based on HMMs of the second order (2):

(1) only the non-entity class has been splitted;
(2) the non-entity class and two most numerous entity categories (protein and DNA) have been splitted;
(3) all the entity classes have been splitted.

It may be observed that each following model includes the set of entity tags of the previous one. Thus, the last model has the greatest number of states.

Besides, we have carried out various experimens with a different number of boundary tags, and we have concluded that only adding two tags (E - end of an entity and S - a single word entity) to a standard set of boundary tags, supplied by the JNLPBA corpus annotation, can notably improve the performance of the HMM-based model.

Consequently, each entity tag of our models contains the following components:

(i) entity class (protein, DNA, RNA, etc.);
(ii) entity boundary (B - beginning of an entity, I - inside of an entity, E - end of an entity, S - a single word entity);
(iii) POS information.

With respect to the CRF approach, the specialization strategy seems to be rather absurd, because it was exactly developed to be able to represent a rich set of features. Therefore, instead of increasing of the states number the greater quantity of feature functions corresponding to each word should be used. Our CRF-based NE recognizer along with the POS tags information employes also context features in a window of 5 words.

4 Experiments

The standard evaluation metrics used for classification tasks are next three measures:

(1) Recall (R) which can described as a ratio between a number of correctly recognized terms and all the correct terms;
(2) Precision (P) that is a ratio between a number of correctly recognized terms and all the recognized terms;
(3) F-score (F), introduced by [10], is a weighted harmonic mean of recall and precision which is calculated as follows:

$$F_\beta = \frac{(1+\beta^2) * P * R}{\beta^2 * P + R} \qquad (4)$$

where β is a weight coefficient used to control a ratio between recall and precision. As a majority of researchers we will exploit an unbiased version of F-score - F_1 which establish an equal importance of recall and precision.

The first experiments we have carried out were devoted to compare our three HMM-based models in order to analyze what entity class splitting provides the best performance. In Table 3 our baseline (i.e., the model without class balancing procedure) is compared with our three models. Although all our models have improved the baseline, there is a significant difference between the first model and the other two models, which have shown rather similar results.

Table 3: *Comparison of the influence of different sets of POS to the HMM-based system performance*

Model	Tags number	Recall, %	Precision, %	F-score
Baseline	21	63.7	60.2	61.9
Model 1	40	68.4	61.4	64.7
Model 2	95	69.1	62.5	65.6
Model 3	135	69.4	62.4	65.7

In Table 4 the results we obtained with our CRF-based system are presented. Here, the baseline model takes into account only words and their context features. Model 1 is the final model which uses also POS-tag information.

Table 4: *The CRF-based system performance*

Model	Recall, %	Precision, %	F-score
Baseline	61.9	72.2	66.7
Model 1	66.4	71.1	68.7

At first glance, if only the F-score values are compared, the CRF-based model outperforms the HMM-based one with a significant difference (3 points). However, when the recall and precision are compared their opposite behaviour may be noticed : for the HMM-based model the recall almost always is higher than the precision whereas for the CRF-based model the contrary is true.

In Tables 5, 6 recall and precision values of the detection of two biomedical entities "protein" and "cell type" for the HMM and the CRF approaches are presented. The analysis of these tables shows the higher effectiveness of HMMs in finding as many biomedical entities as possible and their failure in the correctness of this detection. CRFs are more foolproof models but, as a result, they commit a greater error of the second order: the omission of the correct entities.

Table 5: *Recall values of a detection of "protein" and "cell type" for the HMM and the CRF medels*

Method	Protein	cell type
HMM	73.4	67.5
CRF	69.8	60.9

Table 6: *Precision values of a detection of "protein" and "cell type" for the HMM and the CRF models*

Method	Protein	cell type
HMM	65.2	65.9
CRF	70.2	79.2

The certain advantage of the CRF model with respect to the HMM one could also be disputed by the fact that the best biomedical NER system [12] is principally based on the HMMs. Nevertheless, the comparison does not seem rather fair, because this system, besides exploiting a rich set of features, employes some deep knowledge resources and techniques such as biomedical databases (SwissProt and LocusLink) and a number of post-processing operations consisting of different heuristic rules in order to correct entity boundaries.

Summarizing the obtained results we can conclude that the possibility of an effective combination of CRFs and HMMs would be very beneficial. Since generative and discriminative models have different nature, it is intuitive, that their integration might allow to capture more information about the object under investigation. The example of a successful combination of these methods can be a Semi-Markov CRF approach which was developed by [7] and is a conditionaly trained version of semi-Markov chains. This approach proved to obtain better results on some NER problems than CRFs.

5 Conclusions

In this paper we have presented two biomedical NE recognizers based on the HMM and CRF approaches. Both models have been constructed with the use of the same additional information in order to compare fairly their performance under the same conditions. Since CRFs and HMMs belong to different families of classifiers two distint strategies have been applied to incorporate an additional knowledge into these models. For the former model a methology of states specialization has been used whereas for the latter one all additional information has been presented in the feature functions of words.

The comparison of the results has shown a better performance of the CRF approach if only F-scores of both models are compared. If also the recall and the precision are taken into account the advantage of one method with respect to another one does not seem so evident. In order to improve the results, a combination of both approaches could be very useful. As future work we plan to apply a Semi-Markov CRF approach for the biomedical NER model construction and also investigate another possibility of the CRF-based and the HMM-based models integration.

Acknowledgments

This work has been partially supported by MCyT TIN2006-15265-C06-04 research project.

References

[1] K. B. Cohen and L. Hunter. *Natural Language Processing and Systems Biology.* Springer Verlag, 2004.

[2] J. D. Kim, T. Ohta, Y. Tsuruoka, and Y. Tateisi. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the Int. Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004)*, pages 70–75, 2004.

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning*, pages 282–289, 2001.

[4] A. McCallum. Efficiently inducing features of conditional random fields. In *In Proceedings of the 19th Conference in Uncertainty in Articifical Intelligence (UAI-2003)*, 2003.

[5] A. Molina and F. Pla. Shallow parsing using specialized hmms. *JMLR Special Issue on Machine Learning approaches to Shallow Pasing*, 2002.

[6] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77(2), pages 257–285, 1998.

[7] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In *In Advances in Neural Information Processing (NIPS17)*, 2004.

[8] B. Settles. Biomedical named entity recognition using conditional random fields and novel feature sets. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004)*, pages 104–107, 2004.

[9] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *In Proceedings of the 2003 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-03)*, 2003.

[10] J. van Rijsbergen. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow, 1979.

[11] J. Zhang, D. Shen, G. Zhou, S. Jian, and C. L. Tan. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics (special issue on Natural Language Processing in Biomedicine:Aims, Achievements and Challenge)*, 37(6), 2004.

[12] G. Zhou and J. Su. Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004)*, pages 96–99, 2004.