Biomedical and Chemical Named Entity Recognition with Conditional Random Fields: The Advantage of Dictionary Features

Christoph M. Friedrich, Thomas Revillion, Martin Hofmann and Juliane Fluck
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI); Department of Bioinformatics
Schloss Birlinghoven; 53754 Sankt Augustin; Germany
E-mail: friedrich@scai.fraunhofer.de

Abstract

We present our work on Chemical and Biomedical Named Entity Recognition (NER) using Machine Learning algorithms with different feature sets. It will be demonstrated, that the best results could be obtained using Conditional Random Fields. Furthermore we show the advantage of dictionary based features in this context. All results are obtained with the benchmark settings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004) on the GENIA corpus and show competetive results. Additionally, we provide first results on the recognition of chemical entities in IUPAC format from unstructured text.

1 Introduction

Most of the information in the life sciences is present as unstructured text. Amongst other uses, this wealth of information can be used to interpret the results of expression experiments or to derive pathways of biological or chemical interactions. Text Mining (e.g. (Jensen et al., 2006)) is a possible solution to obtain this information. The first step to efficiently extract information from text is to accurately assign meaningful tags from a well defined ontology to certain entities. Problems arise from the fact that there is no unified nomenclature for protein and gene names that is used by all scientists. Further problems lie in the use of ambiguous names, in the occurrence of multi-word terms and in the use of common-word names.

One possible way to resolve the use of ambiguities and to map synonyms to reference terms is the use of dictionaries with exact match or the use of approximate search techniques (Hanisch et al., 2005). The general disadvantage of this approach lies in the necessary dictionary update and curation when new protein or gene names are invented. Machine Learning algorithms offer an alternative way, learning the hidden properties of named entities from a large annotated text corpus. The main task here is the development of a feature subset, which is able to discriminate different named entities properly and the comparison of different Machine Learning algorithms on the same feature set. In the following sections we describe our work on this topic and demonstrate the feasability of this approach on a public benchmark set and for the recognition of chemical entities in text.

2 Description of task

A prerequisite for applying Machine Learning algorithms to entity recognition is the existence of an annotated representative text corpus. In the biomedical domain the GENIA Corpus (Kim et al., 2003) and associated Ontology is a frequently used benchmark set. The GENIA Corpus consists of 2000 Medline abstracts from the years 1990-1999 which are retrieved with the search terms: “human”, “blood cell” and “transcription factor”. It has about 400000 tokens and about 100000 manual annotations from a set of 36 classes. We used a modified version of this corpus, that has been provided as a benchmark setting for a predictive challenge at the JNLPBA-2004 (Kim et al., 2004). In this, entities from a subset of 5 classes from the GENIA ontology (protein, DNA, RNA, cell_line, cell_type) have to be automatically annotated.

The following sentence gives an impression of the task:

Kappa B-specific DNA binding proteins : role in the regulation of **human interleukin-2 gene** expression .

In this example protein names are underlined and genes (DNA) presented in **bold** font. As an additional obstacle, the occurrence of the entities in text is highly unbalanced (e.g. approx. 30000 proteins and 1000 RNA annotations).

The training-set is the complete GENIA corpus, whereas the test set of the task has been selected from 400 newly annotated abstracts from a different time period (1978-2001) and with search terms from a super-domain: "blood cell" and "transcription factor". It can be assumed, that the different time period and selection from a super-domain has to result in errors of the entity recognizer.

2.1 Testing of Machine Learning algorithms

For this task several different Machine Learning algorithms have been tested. Due to their successes in other Text Mining tasks e.g. (Joachims, 1998), Support Vector Machines (SVM) are used frequently in this domain. We used the multi-class implementation of SVMLight[1] with a linear kernel, which has been described in (Tsochantaridis et al., 2004). We tested an Instance Based Learning method TiMBL[2] (Daelemans et al., 2004) and the Maximum Entropy Markov Model (MEMM), Naive Bayes and Conditional Random Field (CRF) machine learners from the Mallet-toolkit (McCallum, 2002). During our initial experiments with an *ad hoc* feature set, it showed, that the CRF outperformed all other methods. We therefore concentrated our work on this Machine Learning method. The Conditional Random Field (see (McCallum and Sutton, 2006) for an introduction) is an undirected graphical machine learning model, specially suitable for sequence data. This method is not affected by the label-bias problem which is clearly present due to the unbalancedness of the classes in the training set. Although it is possible to build Conditional Random Fields from arbritrary graphical models, we stick to the linear-chain model constraining state transitions to transitions found in the training corpus. The L-BFGS (limited-memory breadth-first quasi-newton search) method (see (Malouf, 2002) for a comparison of optimization methods in this domain) has been used for training. After sensitivity tests, all regularization parameters and stopping criteria of the implementation have been left to their defaults. Additionally to the chosen feature set we automatically added the feature set of the left and right neighbour token.

[1] http://svmlight.joachims.org

[2] http://ilk.uvt.nl/timbl/

Feature	Example
InitCap	Interleukine
HasDash	IL-2
AllCap	IL
IsGreek	Gamma
IsRoman	VII
containsDigit	CD28
prefix3	acetyltransferase
suffix3	acetyltransferase
Begin Delimiter	the
tissue	artery
Porter Stemming	bind (binding)
Part-Of-Speech*	VB

Table 1: Partial list of used features. Marked* have not been used in our best model.

2.2 The used features

The main problem using Machine Learning algorithms for Named Entity Recognition tasks is the design of a proper set of features. The features should be able to generalize, which means to discriminate the entities correctly even on new, unseen samples. In general morphological, lexical, statistical and dictionary based features are suitable for this task. Morphological features discriminate for example between tokens which are non-numerical ("example"), consist only of capital letters ("EXAMPLE") or contain only abbreviated nucleid acids ("ATTTCG"). An example for an lexical feature is the Part-Of-Speech (POS), which assigns a token its position dependent function in a sentence. Statistical features assign frequeny depending properties, e.g. "the" is a frequent Begin Delimiter for proteins, that can be found in the training corpus. Dictionary based features allocate tokens to token classes, depending on their inclusion in certain dictionaries (e.g. token "Leu" in the dictionary of aminoacids). In table 1 a partial list of the used features is given.

Inspired from the work of (McDonald and Pereira, 2005) we used the AB_GENE Dictionaries mentioned in their work, that have been used in the BioCreaTive predictive challenge task 1A (Hirschmann et al., 2005). These dictionaries in-

Authors	F-Score	Method
(Zhou and Su, 2004)	72.6	SVM+CRF
Our method*	**71.5**	**CRF**
(Settles, 2005)*	70.5	CRF
(Finkel et al., 2004)	70.1	MEMM
(Settles, 2004)	69.8	CRF
(Song et al., 2004)	66.3	SVM+CRF
Our method*	**65.2**	**SVM**
(Zhao, 2004)	64.8	HMM
(Rössler, 2004)	64.0	SVM+HMM
(Park et al., 2004)	63.0	SVM
(Lee et al., 2004)	49.1	SVM

Table 2: Results for the JNLPBA 2004 shared task partly from (Kim et al., 2004). Marked* results have been obtained after the competition

clude amongst others aminoacids, tissues, genenames, units, minerals, stop-words and organisms.

3 Results for the JNPLBA 2004 shared task

In table 2 the results for our best Support Vector Machine and Conditional Random Field are given with comparisons to other published results on the tested benchmark set. The training for the multiclass SVM (with linear kernel) took approx. 60h on a standard PC with 3GHz and 2GB of memory. Training with gaussian kernels was judged unfeasible due to time limits and stopped after approx. 180h. Training time for the CRF was approx. 15h for 263 training iterations of 500 possible. This shows the actual limits for larger corpora, as the necessary training time is dependent on the number of features and size of the corpus. In the working phase 25 abstracts/s or 40kb/s could be processed.

The Conditional Random Field model showed a resubstitution F-Score of 91.0 on the full training set, hinting for possible annotation bias in the training corpus. The best model had an F-Score on the independent test-set of 71.5 (precision 70.0/recall 73.1). Without the dictionary based features, we only reached an F-Score of 70.5 on the test set, which is a comparable limit found in similar published approaches (Settles, 2005) and clearly shows the improvement adding dictionary based features.

To see if the different time period and superdomain of the test set, affects the performance, we compared 19-fold cross-validation results on the training-set with the performance on the independent test-set. The F-Score of $74.5 >> 71.5$ showed that the training-corpus is not completely representative for this super-domain. It can be concluded, that the method in general is dependent on the quality of the corpus.

4 Preliminary Results for Chemical Entity Recognition

Chemical names in text can be present in various forms, one standardized nomenclature comes from the International Union of Pure and Applied Chemistry (IUPAC)[3] and forms a systematic way of naming organic chemical compounds, that can be mapped to their structure. An example for IUPAC names is:

> 2-amino-3-(3-hydroxy-5-
> methylisoxazol-4-yl)propionic acid

To find these occurrences we used a trained Conditional Random Field. Due to the lack of an annotated corpus for the detection of chemical entities in text, we used the GENIA corpus described in the prior section and replaced the protein names with IUPAC names, obtained from public chemical databases. A Conditional Random Field trained with the feature set of the JNLPBA task gave an F-Score of 97.25 (precision 97.22/recall 97.27) on unseen IUPAC names. This is competetive, due to common misspellings that prevent rule-based approaches from being perfect. One of the problems that have been detected after inspection of the results is the tokenization of the input text. The occurrence of very long terms of short tokens (mostly special characters) call for the inclusion of a larger range of the surrounding tokens, which will affect the necessary computing time and memory demands. The existence of an annotated text-corpus from the chemical domain would facilitate comparisons between different methods and increase the significance of these results.

5 Conclusion

Conditional Random Fields have shown to be competetive for biomedical Entity Recognition tasks. The possible performance is highly dependent on the quality and representativeness of the training corpus. It could be demonstrated, that the

[3] http://www.iupac.org

chosen feature set affects the performance and that well chosen dictionary based features improve the results on unseen data. Finally, we demonstrated that CRFs might also be useful for the recognition of chemical entities such as IUPAC terms in text.

References

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2004. TiMBL: Tilburg Memory Based Learner; Version 5.1; Reference Guide. Technical Report ILK 04-02, Tilburg University.

J. Finkel, S. Dingare, H. Nguyen, M. Nissim, G. Sinclair, and C. Manning. 2004. Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pages 88–91, Geneva, Switzerland.

D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck. 2005. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 (Suppl 1)(S14).

L. Hirschmann, A. Yeh, C. Blaschke, and A. Valencia. 2005. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 (Suppl 1)(S1).

L. J. Jensen, J. Saric, and P. Bork. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews; Genetics*, 7:119–129, 2.

T. Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on machine learning (ECML 1998)*.

J. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 (Suppl 1):i180–i182.

J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pages 70–75, Geneva, Switzerland.

C. Lee, W.-J. Hou, and H.-H. Chen. 2004. Annotating Multiple Types of Biomedical Entities: A Single Word Classification Approach. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pages 80–83, Geneva, Switzerland.

R. Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, pages 49–55.

A. K. McCallum and C. Sutton. 2006. An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press.

A. K. McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

R. McDonald and F. Pereira. 2005. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6 (Suppl 1)(S6).

K.-M. Park, S.-H. Kim, D.-G. Lee, and H.-C. Rim. 2004. Boosting Lexical Knowledge for Biomedical Named Entity Recognition. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pages 75–79, Geneva, Switzerland.

M. Rössler. 2004. Adapting an NER-System for German to the Biomedical Domain. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pages 92–95, Geneva, Switzerland.

B. Settles. 2004. 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, Geneva, Switzerland.

B. Settles. 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Y. Song, E. Kim, G. G. Lee, and B.-K. Yi. 2004. POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pages 100–103, Geneva, Switzerland.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. 2004. Support Vector Learning for Interdependent and Structured Output Spaces. In *Proceedings of the International Conference on Machine Learning (ICML 2004)*.

S. Zhao. 2004. Named Entity Recognition in Biomedical Text using a HMM model. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pages 84–87, Geneva, Switzerland.

G. Zhou and J. Su. 2004. 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, Geneva, Switzerland.