Named Entity Recognition in Biomedical Texts using an HMM Model

Shaojun Zhao
Department of Computing Science
University of Alberta
Edmonton, Canada, T6G 2H8
shaojun@cs.ualberta.ca

Abstract

Although there exists a huge number of biomedical texts online, there is a lack of tools good enough to help people get information or knowledge from them. Named entity Recognition (NER) becomes very important for further processing like information retrieval, information extraction and knowledge discovery. We introduce a Hidden Markov Model (HMM) for NER, with a word similarity-based smoothing. Our experiment shows that the word similarity-based smoothing can improve the performance by using huge unlabeled data. While many systems have laboriously hand-coded rules for all kinds of word features, we show that word similarity is a potential method to automatically get word formation, prefix, suffix and abbreviation information automatically from biomedical texts, as well as useful word distribution information.

1 Introduction

In the Message Understanding Conference (MUC), Named entity Recognition aims to classify proper nouns, dates, time, measures and locations, etc. Many researchers adapt their systems from MUC to the biomedical domain, such as (Fukuda *et al* 1998), (Proux *et al* 1998), (Nobata *et al* 2000), (Collier *et al* 2000), (Gaizauskas *et al* 2000), (Kazama *et al* 2002), (Takeuchi *et al* 2002), (Lee *et al* 2003) and (Zhou *et al* 2004). As opposed to rule-based systems, machine learning-based systems could train their models on labeled data. But due to the irregular forms of biomedical texts, people still need to carefully choose word features for their systems. This work requires domain specific knowledge. How to get the domain knowledge automatically is a question that has not been fully investigated. Our system is built on an HMM model with the words themselves as the features. Huge unlabeled corpus is gathered from MEDLINE. Word similarity information is computed from the corpus and we use a word similarity-based smoothing to overcome the data sparseness.

2 Data Preparation

2.1 Labeled Data

Our labeled data is from GENIA 3.02 (Ohta *et al* 2002), which contains 2,000 abstracts (360K words). It has been annotated with semantic information such as DNA, protein annotations. These are useful for training models. It contains Part of Speech (POS) information as well. Although POS is not considered very useful for NER in newspaper articles, it can dramatically improve the performance of NER in biomedical texts (Zhou *et al* 2004). Our model is trained from this labeled data.

2.2 Unlabeled Data

We downloaded **17G** XML abstract data from MEDLINE, which contains **1,381,132** abstracts. Compared to the labeled data, we have far more unlabeled data, and the amount of available unlabeled data increases every day. We used this unlabeled data for computing word similarity. We extracted 66,303,526 proximity relationships from the unlabeled data.

3 Distributional Word Similarity

"Words that tend to appear in the same contexts tend to have similar meanings." (Harris 1968). For example, the words *corruption* and *abuse* are similar because both of them can be subjects of verbs like *arouse*, *become*, *betray*, *cause*, *continue*, *cost*, *exist*, *force*, *go on*, *grow*, *have*, *increase*, *lead to*, and *persist,* etc, and both of them can modify nouns like *accusation*, *act*, *allegation*, *appearance*, and *case*, etc.

Many methods have been proposed to compute distributional similarity between words, e.g., (Hindle, 1990), (Pereira *et al*. 1993), (Grefenstette 1994) and (Lin 1998). Almost all of the methods represent a word by a feature vector where each feature corresponds to a type of context in which the word appeared.

3.1 Proximity-based Similarity

It is natural to use *dependency relationship* (Mel'čuk, 1987) as features, but a parser has to be available. Since biomedical text is highly irregular, and is very different from text like newspaper, a parser developed for the newspaper domain may not perform very well on biomedical text. Since most dependency relationships involve words that are situated close to one another, the dependency relationships can often be approximated by co-occurrence relationships within a small window (Turney 2001); (Terra and Clarke 2003). We define the features of the word *w* to be the first non-stop word on either side of *w* and the intervening stop words (which can be defined as the top-k most frequent words in the corpus). For example, for a sentence "*He got a job from this company.*" (Considering *a*, *from* and *this* to be stop words.), the features of *job* provided by this sentence are shown in Table 1.

Features	Frequency
(*left, got*)	0.50
(*left, a*)	0.50
(*right ,from*)	0.33
(*right, this*)	0.33
(*right, company*)	0.33
…	…

Table 1: Features for word "job"

3.2 Computing Word Similarity

Once the contexts of a word are represented as a feature vector, the similarity between two words can be computed using their context vectors. We use $(u_1, u_2 \ldots u_n)$ and $(v_1, v_2 \ldots v_n)$ to denote the feature vectors for the words *u* and *v* respectively, where *n* is the number of feature types extracted from a corpus. We use f_i to denote the *i*th feature.

The point-wise mutual information (PMI) between a feature f_i and a word *u* measures the strength association between them. It is defined as:

$$pmi(f_i, u) = \log\left(\frac{P(f_i, u)}{P(f_i) \times P(u)}\right)$$

where $P(f_i,u)$ is the probability of f_i co-occurring with *u*; $P(f_i)$ is the probability of f_i co-occurring with any word; and $P(u)$ is the probability of any feature co-occurring with *u*.

The similarity between word *u* and *v* is defined as the Cosine of PMI:

$$sim_{word}(u, v) = \frac{\sum_{i=1}^{n} pmi(f_i, u) \times pmi(f_i, v)}{\sqrt{\sum_{i=1}^{n} pmi(f_i, u)^2} \times \sqrt{\sum_{i=1}^{n} pmi(f_i, v)^2}}$$

Different similarity measures of distributional similarity can affect the quality of the result to s statistically significant degree. (Zhao and Lin 2004) shows that the Cosine of PMI is a significantly better similarity measure than several other commonly used similarity measures.

Similar words are computed for each word in the unlabeled data. Only a subset of the similarity information is useful, because the similarity of words outside of the training data and test data vocabulary is not used. We only take into account the similar words that occur in the training data more than 10 times and only those word pairs which have point-wise mutual information greater than a threshold (0.04). Table 2 shows the computing result for "IL-0"[1]:

Similar Words	Similarity
interleukin-0	0.510891
IL-00	0.486665
IFN-gamma	0.44945
TNF-alpha	0.44702
GM-CSF	0.438226
TNF	0.37703
IL-0beta	0.365072
interferon-gamma	0.350704
IL0	0.336974
…	…

Table 2: Similar words for "IL-0"

Table 2 also shows that the similar words can capture word formation (IL-00, IL-0beta, and IL0 etc) and abbreviation (interleukin-0) information. A complete list of these word pairs and their similarity is available online[2]. The rule-based system may not able to capture words like IL-0ra, IL-0Ralpha, which are in the similar word list of IL-0, and it is very likely that they belong to the same semantic category. Many different kinds of expressions for numbers (like 0, 00-00, 00.00, -00, 0/0, five, six, 0-, iii, IV etc) are grouped together automatically.

4 HMM Model and Smoothing Schema

We follow the HMM model introduced in (Zhou *et al* 2004). The structure of an HMM model contains *States* and *observations.* In our model, each state is represented by a semantic tag, or a POS tag if the semantic tag is not available; each observation contains a word sequence. The main computing difficulty in (Zhou *et al* 2004) is the probability of a tag given a word sequence: formula (1). We use formula (2) to estimate formula (1). If the bigram is unseen in the training data, we use formula (3). If the unigram is also unseen, we use the *unknown* information which is

[1] We changed any single digit to 0.

[2] http://www.cs.ualberta.ca/~shaojun/biolist.txt

gathered from the low frequency words in the training data.

$$P(tag_t \mid wordsequence) \quad (1)$$

$$P(tag_t \mid word_t, word_{t+1}) \quad (2)$$

$$P(tag_t \mid word_t) \quad (3)$$

We find that about 26% of the bigrams *($word_t$, $word_{t+1}$)* in the testing data is unseen, so the smoothing is critical.

In order to compute formula (1), we can use the back-off (Katz 1987); (Bikel *et al* 1999) approach. *Baseline1* and *Baseline2* in our system use different back-off schema.

The following formula is introduced in (Lee 1999) for word similarity-based smoothing:

$$P(tag_t \mid w_t) = \frac{\sum_{w'_t \in S(w_t)} sim(w_t, w'_t) P(tag_t \mid w'_t)}{\sum_{w'_{t+1} \in S(w_t)} sim(w_t, w'_t)} \quad (4)$$

where S(w) is a set of candidate similar words and *sim(w,w')* is the similarity between word *w* and *w'*. Word similarity-based smoothing approach is used in our system to make advantage of the huge unlabeled corpus. In order to plug the word similarity-based smoothing into our HMM model, we made several extensions to formula (4).

For each word *w,* we define *p* as the distribution of *w*'s tags, which are annotated in the training data. We use the KL-Divergence to compute the distance between two distributions:

$$KLD(p_1 \,||\, p_2) = \sum_x p_1(x) \log\left(\frac{p_1(x)}{p_2(x)}\right)$$

We define the similarity between the tag distributions of word *w* and *w'* as:

$$sim_{tag}(w, w') = \frac{1}{1 + KLD(P(tag \mid w) \,||\, P(tag \mid w'))}$$

The harmonic average of word similarity and tag distribution similarity is defined as the similarity of word *w* and *w'* used in our system.

$$s(w, w') = \frac{2 sim_{word}(w, w') \times sim_{tag}(w, w')}{sim_{word}(w, w') + sim_{tag}(w, w')}$$

So, we get formula (5) and (6). Formula (5) is for bigram smoothing and formula (6) is for unigram smoothing.

$$P(tag_t \mid w_t, w_{t+1}) = \frac{\sum_{w'_{t+1} \in S(w_{t+1})} s(w_{t+1}, w'_{t+1}) P(tag_t \mid w_t, w'_{t+1})}{\sum_{w'_{t+1} \in S(w_{t+1})} s(w_{t+1}, w'_{t+1})} \quad (5)$$

$$P(tag_t \mid w_t) = \frac{\sum_{w'_{t+1} \in S(w_{t+1})} s(w_t, w'_t) P(tag_t \mid w'_t)}{\sum_{w'_{t+1} \in S(w_{t+1})} s(w_t, w'_t)} \quad (6)$$

Because it is natural to back-off from bigram to unigram, in our system, a back-off smoothing approach is combined with the word similarity-based smoothing. We follow these procedures to compute formula (1).

1. Check the frequency of the bigram (w_t, w_{t+1}). If the frequency is high (>10), use formula (2). Stop.
2. Check the frequency of the unigram (w_t). If the frequency of the unigram is high (>30), use formula (3). Stop.
3. Try formula (5) for bigram smoothing, and check the frequency summary of the similar words involved in the smoothing. If the summary is high (>10), use formula (5). Stop.
4. Try formula (6) for unigram smoothing, and check the frequency summary for this case. If the summary is high (>30), use formula (6). Stop.
5. If the bigram is not unseen, use formula (2). Stop.
6. If the unigram is not unseen, use formula (3). Stop.
7. Use low frequency (<5) word information in the training data and formula (3).

Our *Baseline1* uses step 5, 6 and 7; *Baseline2* uses step 1, 2, 5, 6 and 7.

5 Experiment Result

The experiment results are shown in Table 3:

Methods	R	P	F-score
Baseline1	64.77%	59.87%	62.22%
Baseline2	66.99%	61.25%	63.99%
Our system	69.41%	62.98%	66.04%

Table 3: Performance comparison

The *Baseline2* outperforms *Baseline1* because it prevents from using low frequency unigrams, and our system outperforms *Baseline1* and *Baseline2* because it prevents from using low frequency bigrams and unigrams. Our system benefits from huge unlabeled corpus.

6 Conclusion

We trained an HMM model on labelled data to recognize named entities in biomedical texts. Word similarity information was computed from huge unlabeled data. A word similarity-based smoothing method was integrated into the system, and improved the overall performance. We would like to see if it could also be plugged into other existing systems, and hopefully also improve their performance.

We also argue that the automatically acquired similar words are rich with word features, such as word formation, prefix, suffix, abbreviation, expression variation and clustering information. We will further investigate the usefulness of them in the future.

7 Acknowledgements

Thanks to Dekang Lin and other members in the Natural Language Processing Group at the University of Alberta for helpful discussion, the anonymous reviewers for their insightful comments. This material is based upon work supported by the Alberta Ingenuity Centre for Machine Learning (AICML).

References

Bikel, D., Schwartz, R., Weischedel, R. 1999. *An Algorithm that Learns What's in a Name*. In *Proc. of Machine Learning (Special Issue on NLP).* Collier, N., Nobata, C., Tsujii, J. 2000. *Extracting the names of genes and gene products with a hidden Markov model*. In *Proc. of COLING 2000*, pages 201-207.

Fukuda, K., Tsunoda, T., Tamura, A., Takagi, T. 1998. "*Toward Information extraction: Identifying protein names from biological papers*", in Proc. of the Pacific Symposium on Biocomputing 98 (PSB 98), Hawaii

Gaizauskas, R., Demetriou, G., Humphreys, K. 2000. *Term Recognition and Classification in Biological Science Journal Articles*. In *Proc. of the Computational Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on NLP*, pages 37-44.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers, Boston

Harris, Z.S. 1968. *Mathematical Structures of Language*. New York: Wiley.

Hindle, D. 1990. *Noun Classification from Predicate-Argument Structures*. In *Proceedings of ACL-90*. pp. 268-275. Pittsburgh, Pennsylvania

Katz, S.M. 1987. *Estimation of Component of a Speech Recognizer.* IEEE Transactions on Acoustics. Speech and Signal Processing. 35. 400-401.

Kazama, J., Makino, T., Ohta, Y., Tsujii, J. 2002. *Tuning Support Vector Machines for Biomedical Named Entity Recognition*. In *Proc. of the Workshop on Natural Language Processing in the Biomedical Domain (at ACL'2002)*, pages 1-8

Lee, K.J., Hwang, Y.S., Rim H.C. 2003. *Two phase biomedical NE Recognition based on SVMs*. In Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine. pp.33-40. Sapporo, Japan

Lee, L. 1999. *Measures of distributional similarity*. In Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, 1999, pp. 25-32.

Lin, D. 1998. *Automatic Retrieval and Clustering of Similar Words*. In *Proceedings of COLING-ACL98*. Montreal, Canada.

Mel'čuk, I. A., 1987. *Dependency Syntax: theory and practice*. State University of New York Press. Albany, NY.

Nobata, C., Collier, N., Tsujii, J. 2000. *Comparison between Tagged Corpora for the Named Entity Task*. In the Proceedings of ACL 2000 Workshop on Comparing Corpora. Hong Kong, China. pp. 20-27

Ohta, T., Tateisi, Y., Kim, J., Mima, H., Tsujii, J. 2002. *The GENIA corpus: An annotated research abstract corpus in molecular biology domain*. In Proc. of HLT 2002.

Pereira, F., Tishby, N., Lee, L. 1993. *Distributional Clustering of English Words*. In *Proceedings of ACL-93*. pp. 183-190. Columbus, Ohio.

Proux, D., Rechenmann, F., Julliard, L., Pillet, V., Jacq, B. 1998. *Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction*. In *Proc. of Genome Inform Ser Workshop Genome Inform*, pages 72-80.

Takeuchi, K., Collier, N. 2002. *Use of Support Vector Machines in Extended Named Entity Recognition.* In *Proc. of the Sixth Conference on Natural Language Learning (CONLL 2002)*, pages 119-125.

Terra, E. L., Clarke, C. 2003. *Frequency Estimates for Statistical Word Similarity Measures*. In *the Proceedings of the 2003 Human Language Technology Conference*, pp.244-251. Edmonton, Canada, May

Turney, P.D. 2001. *Mining the Web for synonyms: PMI-IR versus LSA on TOEFL*, *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, Freiburg, Germany, pp. 491-502.

Zhou, G., Zhang, J., Su, J., Shen, D., Tan, C. 2004. *Recognizing names in biomedical texts: a machine learning approach.* Bioinformatics Advance Access.

Zhao, S., Lin, D. 2004. *A Nearest-Neighbor Method for Resolving PP-Attachment Ambiguity.* In Proceedings of the First International Joint Conference on Natural Language Processing, 2004. Sanya, China.