Int. Computer Symposium, Dec. 15-17, 2004, Taipei, Taiwan.

Statistical Approaches to Biomedical Entities Recognition

Tyne Liang
Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan
tliang@cis.nctu.edu.tw

Ping-Ke Shih
Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan
gis91535@cis.nctu.edu.tw

Diang-Song Wu
Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan
gis92807@cis.nctu.edu.tw

Abstract

Named Entity Recognition (NER) is one of essential tasks for knowledge acquisition from scientific literature. In this paper, a full automatic named entities recognition from biomedical literature is presented by using Hidden Markov Model in which a rich set of features are concerned and back-off strategy is employed to overcome data sparseness problem. Experiments with GENIA corpora of different versions showed that the presented approach achieved promising results of 76% and 62% F-score for singular-type and multiple-type entities recognition respectively.

1. Introduction

With the rapid growth of biomedical research, huge amounts of biomedical resources are available. For example, the amount of biomedical citations available by PubMed increases 68.95% in recent ten years. Hence efficient named entity recognition (NER) becomes indispensable task for knowledge acquisition from research literature. Unlike the extraction in general domains in which efficient NER approaches may yield 94% and 97% F-scores in MUC-7 and MUC-6 respectively, the best result for multi-classes biomedical entities extraction in GENIA 3.0 (an annotated corpus in biomedical domain) is 66.5% F-score [7]. This is because the issues such as open vocabulary, synonyms, boundaries, semantic crossover become more complicated in biomedical domains. For example, the number of entries in SwissProt, a protein knowledge base, increases 277.36% in recent ten years [1]. Each protein entity contains 2.54 synonyms in average, and each synonym contains 2.74 tokens in average.

There are three NER methods, namely rule-based, statistical and hybrid methods, proposed in recent literature. Generally, rule-based approaches employ terms and rules (e.g. heuristic rules and decision tree rules) to produce candidates which then are verified by using lexical analysis. KeX [3] and Yapex [9] are two famous rule-based systems useful for protein entities extraction. Yet rule-based methods are essentially lack of portability and scalability.

Unlike rule-based approaches which demand more domain knowledge in rules construction, statistical approaches have been presented for their easy portability and scalability. Different statistical models have been applied to biomedical entities extraction, such as Hidden Markov Model (HMM), Support Vector Model (SVM), Maximum Entropy (ME), and Naïve Bayes. The recognition accuracy achieved by these models generally depends on a well-tagged training corpus and well set of input features [2,5,6,12,13,14,16]. Well-known training corpus like GENIA 3.01 [20] has been widely used for training models [5,6,7,11,14,16]. Different sets of features have different contribution in different phases of NER tasks [7].

On the other hand hybrid approaches were proposed by using coded rules, statistical model and outer resources like dictionaries. For example, Proux et al. [10] built a system to detect gene symbols and names in biological texts. The backbone of the system is a tagger for tokenization, lexical lookup, and disambiguation. To deal with unknown words, it used lexical rules to obtain candidates and used a HMM-based disambiguator for further verification. Similar approach can be found in [14,15] which extracted chemical names from biomedical texts by a rule-based segmentation and a Bayesian classification.

In this paper, the NER task was conducted purposely for biomedical entities. It is also an essential work at constructing the automation of biomedical interaction knowledge base. On the other hand the concise HMM-based extractor together with a back-off strategy was implemented. Experimental results on GENIA corpus showed that the presented approach could achieve promising results in terms of 76% and 62% F-score for singular-type and multiple-type entities recognition respectively.

The organization of this paper is as follows. Section 2 describes text preprocessing. Section 3 describes internal, external and global features. Section 4 presents the proposed extractors and corresponding experimental results. Section 5 gives the conclusion and future works.

2. Text Preprocessing

Two available processors *Sentence Splitter* [18] and *Penn Treebank Tokenizer* [19] were adopted for sentence segmentation and tokenization respectively. The POS tagging process was based on a traditional HMM which forward induction was applied. The goal of HMM is to optimize the probability of a POS sequence in which there are 35 kinds of POS tags. In order to train a better model to tag the articles other than GENIA 3.02p, we used the whole corpus as training set and it turned out that the presented POS tagger could achieve 94.84% accuracy.

3. Features Extraction

Extraction of those features useful for entity extraction was done on the basis of feature occurrences. In this paper rich set of features were concerned, including internal, external and global features. Internal features indicate those surface clues in tokens (e.g. initial character is upper case). There are 17 internal features, partly adopted from the set in [3,6]. For example, features INIT_UPPER, SUFFIX_NUM, LETTER_DIGITAL, and CONTAIN_HYPHEN will be assigned to ‘BK-2’. Besides we consider features not only current token but also preceding token in HMM. We also consider the prefix and suffix string, because they benefit the performance in our studies. We take the most frequent 1,000 three-character prefixes and suffixes strings.

External feature indicate the external information associated with tokens. In this paper we treated POS tag as our external feature set. This is because tokens of protein entities are normally tagged as nouns. Global features are the features extracted from whole training corpus by using statistical method such as Chi-square. The essence of the test is to compare the observed frequencies with the expected frequencies for independence. In this paper the global features are those significant nouns selected by chi-square test. Furthermore a complete-link clustering algorithm was applied to reduce the dimensions of features. The window size set to be three sentences long, we got 142 clusters in GENIA corpus 3.02p.

Table 1: Internal, external and global features.

Features Set	Features	Example
Internal	INIT_UPPER	BK-2
	INIT_LOWER	c-551
	INIT_NUM	5-HT1B
	INIT_SYMBOL	-p1
	SUFFIX_NUM	MDBP-2-H1
	CONTAIN_GREEK	3beta-hydroxysteroid
	LETTER_DIGITAL	A43
	TWO_CAPS	RasHua
	ALL_UPPER	ALP
	ALL_LOWER	bombesin
	NUM	35 kDa protein
	OTHER_SINGLE_SYMBOL	'
	CONTAIN_HYPHEN	5-HT1B
	SINGLE_UPPER	A protein
	CONTAIN_SLASH	C/EBP
	Prefix	acetyl-CoA
	Suffix	carboxylase
External	POS Tags	NNS
Global	Global Nouns	receptor

4. HMM-based Extraction

Given a token sequence $T_1^n = t_1 t_2 \cdots t_n$, the goal is to find an optimal state sequence $S_1^n = s_1 s_2 \cdots s_n$ that maximizes $\log Pr\left(S_1^n \mid T_1^n\right)$, the logarithm probability of state sequence T_1^n corresponding to the given token sequence S_1^n.

Traditional HMM

By applying Bayes’s rule to :

$$Pr\left(S_1^n \mid T_1^n\right) = \frac{Pr\left(S_1^n, T_1^n\right)}{Pr\left(T_1^n\right)} \tag{1}$$

we have

$$\arg\max_S \log Pr\left(S_1^n \mid T_1^n\right) = \arg\max_S \left(\log Pr\left(T_1^n \mid S_1^n\right) + \log Pr\left(S_1^n\right)\right) \tag{2}$$

where

$$Pr\left(T_1^n \mid S_1^n\right) = \prod_{i=1}^{n} Pr\left(t_i \mid s_i\right) \tag{3}$$

$$Pr\left(S_1^n\right) = \prod_{i=1}^{n} Pr\left(s_i \mid s_{i-1}\right) \tag{4}$$

under the assumption of conditional probability independence and considering preceding state. Therefore equation (2) can be rewritten as:

$$\arg\max_S \log Pr\left(S_1^n \mid T_1^n\right) = \arg\max_S \left(\sum_{i=1}^{n} \left(\log Pr\left(t_i \mid s_i\right) + \log Pr\left(s_i \mid s_{i-1}\right)\right)\right) \tag{5}$$

Mutual Information HMM

The mutual information HMM (*MI-HMM* for short) was presented in [17] and produced high F-scores in MUC-6 and MUC-7. Different from traditional HMM, *MI-HMM* is aimed to maximize the equation:

$$\arg\max_S \log Pr\left(S_1^n \mid T_1^n\right) = \arg\max_S \left(\log Pr\left(S_1^n\right) + \log \frac{Pr\left(S_1^n, T_1^n\right)}{Pr\left(S_1^n\right) \cdot Pr\left(T_1^n\right)}\right) \tag{6}$$

In order to simplify the computation, the mutual information independence is assumed to be:

$$MI\left(S_1^n, T_1^n\right) = \sum_{i=1}^{n} MI\left(s_i, T_1^n\right) \tag{7}$$

or

$$\log \frac{Pr\left(S_1^n, T_1^n\right)}{Pr\left(S_1^n\right) \cdot Pr\left(T_1^n\right)} = \sum_{i=1}^{n} \log \frac{Pr\left(s_i, T_1^n\right)}{Pr\left(s_i\right) \cdot Pr\left(T_1^n\right)} \tag{8}$$

Applying it to equation (6), we have:

$$\arg\max_S \log Pr\left(S_1^n \mid T_1^n\right) = \arg\max_s \left(\log Pr\left(S_1^n\right) - \sum_{i=1}^{n} \log Pr\left(s_i\right) + \sum_{i=1}^{n} \log Pr\left(s_i \mid T_1^n\right)\right) \tag{9}$$

Concise HMM

The presented concise HMM is based on the idea of maximizing the fundamental $\log Pr\left(S_1^n \mid T_1^n\right)$. In the equation (9), $\log Pr\left(S_1^n\right)$ and $\sum_{i=1}^{n} \log Pr\left(s_i\right)$ are found to carry less meaning because the weak probabilities of states and state transitions are merely 3-by-3 and 3-by-1 matrices respectively. Thus, concise HMM can be simplified as equation (10):

$$\arg\max_S \log Pr\left(S_1^n \mid T_1^n\right) = \arg\max_s \sum_{i=1}^{n} \log Pr\left(s_i \mid T_1^n\right) \tag{10}$$

The concise HMM does not take its state transition into account, therefore we put previous state in the model to ensure correct state induction. Because the presented HMM approach concerned many features mentioned above, it is possible to train a high-accuracy probability model. However, it is not enough to cover all data, so the data sparseness problem arises. To overcome this problem, we used a back-off model and it aims at the token sequence T_1^n in $Pr\left(S_1^n \mid T_1^n\right)$ or in $Pr\left(s_i \mid T_1^n\right)$ where T_1^n represents not only a token sequence but also the sequence's internal, external and global features. We then defined two back-off levels as follows:

(A) First level is based on different combinations of tokens and their features, and T_1^n will be assigned in the descending order:

1. $< s_{-1}, t_{-1}, t_0, f_0 >$
2. $< s_{-1}, t_0, f_0 >$
3. $< s_{-1}, t_{-1}, f_0 >$
4. $< s_{-1}, f_0 >$

Where ‘ f_i ’ represents the feature set including internal, external and global features. ‘ t_i ’ is a token, ‘ s_i ’ expresses a HMM state, and ‘$_i$’ is the *i*th one relative to current token.

(B) Second level is based on different combinations of features, and ‘ f_i ’ in first level is assigned in the descending order:

1. $< f_i^I, f_i^E, f_i^G >$
2. $< f_i^I, f_i^E >$
3. $< f_i^I >$

Where f_i^I, f_i^E and f_i^G represent internal, external and global features respectively.

4.1. Method Comparisons

In this paper, we presented two named entities recognition. One is singular-type entities recognition aimed to recognize protein entities, and the other is multiple-type entities recognition aimed to tag named entities with one of the four major concepts: Protein, DNA/RNA, Source, and Other. Table 2 is the mapping between the concepts we addressed and the ones in GENIA corpus 3.02p. Table 3 is the basic statistics in GENIA 3.02p.

Table 2: The target named entities in terms of GENIA ontology.

Class	Semantic
Protein	amino acid, protein, protein molecule, protein family or group, protein domain or region, protein structure, protein complex, protein N/A, peptide, amino acid monomer
DNA/RNA	DNA molecule, DNA family or group, DNA domain or region, DNA substructure, DNA N/A, RNA molecule, RNA family or group, RNA domain or region, RNA substructure, RNA N/A, polynucleotide, nuclotide
Source	multi cell, mono cell, virus, body part, tissue, cell type, cell line, other artificial source
Other	organic, lipid, carbohydrate, other organic compound, inorganic, atom

Table 3: The basic statistics in GENIA corpus 3.02p.

	Count	Average
Abstract	1,999	
Sentence	18,572	9.29(s/a)

Token	490,469	245.36(t/a)
		26.41(t/s)

Protein Entitiy	32,525	11.05(pn/a)
		1.14(pn/s)

Entity Token	58,220	1.79(tok/pn)

Method comparisons for the three HMM-based models were made on GENIA corpus for singular-type entities recognition in the same environment settings. We used the same back-off model for concise and mutual information HMM, but not for traditional HMM. Table 4 shows that concise HMM yielded the best result for singular-type entities recognition. Traditional HMM obtains good high precision, but low recall. This reason is that we chose a severe probability model to get the best F-score. It is also noticed that the performance of MI-HMM turned out to be the worst at the comparison. This is because the back-off model was used to optimize concise HMM. On the other hand, the impact of features was verified and the results as listed in Table 5 show that every feature turned out to be positive effect ($f^E > f^I > f^G$) for concise HMM.

However, the presented multiple-type recognizer turned out to yield 62.25% F-score in classification phase (67.07% F-score in identification phase) less than 13% for singular-type entities recognition. This is because there might be some entities whose semantic tags are decided by their last words. For instance, “hematopoietic gene” is tagged to be “DNA/RNA” while “hematopoietic gene cell” as “cells” and “hematopoietic cell specific molecules” as “protein”.

Moreover, biomedical named entity is often generated on the behavior of its source, which induces the problem of crossover between classes. For example “human NF-kappa B” should be chunked together as: “<cons sem="G#protein_molecule">human <cons sem="G#protein_molecule">NF-kappa B</cons></cons>”. But we chunk it as “<NE cl=Source>human</NE> <NE cl=Protein>NF-kappa B</NE>”. Even though such tagging result was acceptable, we still treated it as wrong answer.

Table 4: HMM-based models for singular-type recognizer.

HMM	tp + fn	tp + fp	tp	Recall	Precision	F-Score	Feature #
Concise	3,451	3,285	2,553	73.98%	77.72%	75.80%	19
MI	3,451	3,415	2,305	66.79%	67.50%	67.14%	19
Traditional	3,451	2,863	2,263	65.58%	79.04%	71.68%	18

Table 5: The effects of features in concise HMM.

	Features	tp + fn	tp + fp	tp	Recall	Precision	F-Score	Diff.
Protein Entities	All	3451	3285	2553	73.98%	77.72%	75.80%	
	All - f^G	3451	3267	2534	73.43%	77.56%	75.44%	-0.36%
	All - f^E	3451	3176	2442	70.76%	76.89%	73.70%	-2.10%
	All - f^I	3451	3213	2467	71.49%	76.78%	74.04%	-1.76%
Biomedical Entities	Features	tp + fn	tp + fp	tp	Recall	Precision	F-Score	Diff.
	All	8163	8175	5085	62.29%	62.20%	62.25%	
	All - f^G	8163	8205	5080	62.23%	61.91%	62.07%	-0.18%
	All - f^E	8163	8181	4990	61.13%	60.99%	61.06%	-1.19%
	All - f^I	8163	8152	5001	61.26%	61.35%	61.31%	-0.94%

The method comparisons were also made with other statistical models on different corpora. First comparison was made for singular-type entities recognition with the systems developed by Lee et al. [7] and Shen et al. [11] (for 22 and 23 classes NERs) respectively on GENIA version at least 3.0 (GENIA version 3.02p was based on version 3.0 but errors were fixed). Shen's HMM-based NER used semantic features including head noun and special verb. Moreover, abbreviation recognition and cascaded phenomena were conducted to recognize biomedical entities. Second comparison was made for multiple-type entities recognition on GENIA version 1.1 (671 abstracts). From tables 6 and 7 it is noticed that the presented concise model is very competitive.

Table 6: Comparisons with other systems in GENIA version 3.x.

System	Method	GENIA Ver.	Singular-type F-score	Multiple-type F-score	Class #
(Lee, 2003)	SVM	3.0p	69.20%	66.50%	22
(Shen, 2003)	HMM	3.0	70.81%	66.10%	23
KeX	Rule-based	3.02p	40.29%		1
Yapex	Rule-based	3.03p	47.48%		1
Ours	HMM	3.02p	75.80%	62.25%	4

Table 7: Comparisons with other systems in GENIA version 1.1.

System	Method	Singular-type F-score	Multiple-type F-score	Class #
(Kazama, 2002)	SVM	56.50%	54.40%	6
(Kazama, 2001)	ME	54.80%	53.20%	6
KeX	Rule-based	40.55%		1
Yapex	Rule-based	48.26%		1
Ours	HMM	60.82%	59.82%	4

5. Conclusions

In this paper a prototype system for biomedical entities extraction was presented. The kernel part of the presented extractor was built on a concise HMM-based model which was justified to outperform other HMM-based models by yielding 76% and 62% F-score for singular-type and multiple-type entities recognition respectively. The comparisons to other statistical models also show that the proposed model is competitive.

6. Acknowledgements

This research is partially supported by MediaTek Research Center, National Chiao Tung University, Taiwan and partly support by National Science Council, R. O. C. under the grant NSC 93-2213-E-009-074.

7. References

[1] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., Oonovan, C., Phan, I., Pilbout, S. and Schneider, M. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365-370(2003).

[2] Collier, N., Nobata, C., and Tsujii, J. (2000) Extracting the Names of Genes and Gene Products with a Hidden Markov Model. The 18th International Conference on Computational Linguistics (COLING 2000), pp 201-207.

[3] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. (1998) Towards Information Extraction: identifying Protein Names from Biological Papers. The 3rd Pacific Symposium on Biocomputing, pp 707-718.

[4] Jacquemin, C. and Tzoukermann, E. (1997) NLP for term variant extraction: Synergy between morphology, lexicon, and syntax. In: Strzalkowski T., ed, Natural Language Processing and Information Retrieval. Kluwer, Boston, Mass, 1997.

[5] Kazama, J., Makino, T., Ohta, Y., and Tsujii, J. (2001) A Maximum Entropy Tagger with Unsupervised Hidden Markov Models. In the Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium. pp. 333-340.

[6] Kazama, J., Makino, T., Ohta, Y., and Tsujii, J. (2002) Tuning Support Vector Machines for Biomedical Named Entity Recognition. Workshop on Natural Language Processing in the Biomedical Domain, Association for Computational Linguistics 2002, pp 1-8.

[7] Lee, K.J., Hwang, Y.S., and Rim, H.C. (2003) Two-Phase Biomedical NE Recognition based on SVMs. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp. 33-40.

[8] Nobata, C., Collier, N. and Tsujii, J. (1999) Automatic Term Identification and Classification in Biology Texts. The 5th Natural Language Processing Pacific Rim Symposium, pp 369-374.

[9] Olsson, F., Eriksson, G., Franzen, K., Asker, L., and Liden, P. (2002) Notions of Correctness when Evaluating Protein Name Taggers. Proceedings of the 19th International Conference on Computational Linguistics, pp. 765-771.

[10] Proux, D., Rechenmann, F., Julliard, L., Pillet, V., and Jacq, B. (1998). “Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction.” Genome Informatics, pp 72-80.

[11] Shen, D., Zhang, J., Zhou, G., Su, J., and Tan, C.L. (2003) Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp. 49-56.

[12] Takeuchi, K. and Collier, N. (2002) Use of Support Vector Machines in Extended Named Entity Recognition. The 19th International Conference on Computational Linguistics (COLING 2002).

[13] Takeuchi, K. and Collier, N. (2003) Bio-Medical Entity Extraction using Support Vector Machines. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp. 57-64.

[14] Tsuruoka, Y. and Tsujii, J. (2003) Boosting Precision and Recall of Dictionary-based Protein Name Recognition. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp. 41-48.

[15] Wilbur, W. J., Hazard, G. F., Divita, G., Mork, J. G., Aronson, A. R., and Browne, A. (1999). “Analysis of Biomedical Text for Chemical Names: A Comparison of Three Methods.” The 1999 American Medical Information Association Symposium, pp 176-180.

[16] Yamamoto, K., Kudo, T., Konagaya, A., and Matsumoto, Y. (2003) Protein Name Tagging for Biomedical Annotation in Text. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, pp. 65-72.

[17] Zhou, G.D., and Su, J. (2002) Named Entity Recognition using an HMM-based Chunk Tagger. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002).

[18] http://l2r.cs.uiuc.edu/~cogcomp/cc-software.html.

[19] http://www.cis.upenn.edu/~treebank/tokenization.html.

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