A Hybrid Approach to Biomedical Named Entity Recognition and Semantic Role Labeling

Richard Tzong-Han Tsai

Department of Computer Science and Information Engineering

National Taiwan University

Nankang, Taipei, Taiwan, 115

thtsai@iis.sinica.edu.tw

Abstract

In this paper, we describe our hybrid approach to two key NLP technologies: biomedical named entity recognition (Bio-NER) and (Bio-SRL). In Bio-NER, our system successfully integrates linguistic features into the CRF framework. In addition, we employ web lexicons and template-based post-processing to further boost its performance. Through these broad linguistic features and the nature of CRF, our system outperforms state-of-the-art machine-learning-based systems, especially in the recognition of protein names (F=78.5%). In Bio-SRL, first, we construct a proposition bank on top of the popular biomedical GENIA treebank following the PropBank annotation scheme. We only annotate the predicate-argument structures (PAS's) of thirty frequently used biomedical verbs (predicates) and their corresponding arguments. Second, we use our proposition bank to train a biomedical SRL system, which uses a maximum entropy (ME) machine-learning model. Thirdly, we automatically generate argument-type templates, which can be used to improve classification of biomedical argument roles. Our experimental results show that a newswire English SRL system that achieves an F-score of 86.29% in the newswire English domain can maintain an F-score of 64.64% when ported to the biomedical domain. By using our annotated biomedical corpus, we can increase that F-score by 22.9%. Adding automatically generated template features further increases overall F-score by 0.47% and adjunct (AM) F-score by 1.57%, respectively.

1 Introduction

The volume of biomedical literature available on the Web has experienced unprecedented growth in recent years, and demand for efficient methods to process this material has increased accordingly. Lately, there has been a surge of interest in mining biomedical literature. To this end, more and more information extraction (IE) systems using natural language processing (NLP) technologies have been developed for use in the biomedical field. Key biomedical IE tasks include named entity (NE) recognition (NER), such as the recognition of protein and gene names; and relation extraction, such as the extraction of protein-protein and gene-gene interactions.

NER identifies named entities from natural language texts and classifies them into specific classes according to a defined ontology or classification. In general, biomedical NEs do not follow any nomenclature and may comprise long compound words and short abbreviations. Some NEs contain various symbols and other spelling variations. On average, an NE has five synonyms (Tsai *et al*., 2006a), and it may belong to multiple categories intrinsically. Since biomedical language and vo-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 243–246,
New York, June 2006.

cabulary are highly complex and evolving rapidly, Bio-NER is a very challenging problem, which raises a number of difficulties.

The other main focus of Bio-IE is relation extraction. Most systems only extract the relation targets (e.g., proteins, genes) and the verbs representing those relations, overlooking the many adverbial and prepositional phrases and words that describe location, manner, timing, condition, and extent. However, the information in such phrases may be important for precise definition and clarification of complex biological relations.

This problem can be tackled by using semantic role labeling (SRL) because it not only recognizes main roles, such as agents and objects, but also extracts adjunct roles such as location, manner, timing, condition, and extent. (Morarescu *et al.*, 2005) has demonstrated that full-parsing and SRL can improve the performance of relation extraction, resulting in an F-score increase of 15% (from 67% to 82%). This significant result leads us to surmise that SRL may also have potential for relation extraction in the biomedical domain. Unfortunately, no SRL system for the biomedical domain exists.

In this paper, we tackle the problems of both biomedical SRL and NER. Our contributions are (1) employing web lexicons and template-based post-processing to boost the performance of Bio-NER; (2) constructing a proposition bank on top of the popular biomedical GENIA treebank following the PropBank annotation scheme and developing a Biomedical SRL system. We adapt an SRL system trained the World Street Journal (WSJ) corpus to the biomedical domain. On adjunct arguments, especially those relevant to the biomedical domain, the performance is unsatisfactory. We, therefore, develop automatically generated templates for identifying these arguments.

2 Biomedical Named Entity Recognition

Our Bio-NER system uses the CRF model (Lafferty *et al.*, 2001), which has proven its effectiveness in several sequence tagging tasks.

2.1 Features and Post-Processing

Orthographical Features
In our experience, ALLCAPS, CAPSMIX, and INITCAP are more useful than others. The details are listed in (Tsai et al., 2006a).

Context Features
Words preceding or following the target word may be useful for determining its category. In our experience, a suitable window size is five.

Part-of-speech Features
Part-of-speech information is quite useful for identifying NEs. Verbs and prepositions usually indicate an NE's boundaries, whereas nouns not found in the dictionary are usually good candidates for named entities. Our experience indicates that five is also a suitable window size. The MBT POS tagger is used to provide POS information. We trained it on GENIA 3.02p and achieved 97.85% accuracy.

Word Shape Features
As NEs in the same category may look similar (e.g., IL-2 and IL-4), we have to find a simple way to normalize all similar words. According to our method, capitalized characters are all replaced by 'A', digits are all replaced by '0', non-English characters are replaced by '_' (underscore), and non-capitalized characters are replaced by 'a'. To further normalize these words, we reduce consecutive strings of identical characters to one character.

Affix Features
Some affixes can provide good clues for classifying named entities (e.g., "ase"). In our experience, an acceptable affix length is 3-5 characters.

Lexicon Features
Depending on the quality of a given dictionary, our system uses one of two different lexicon features to estimate the possibility of a token in a biomedical named entity. The first feature determines whether a token is part of a multi-word NE in the dictionary, while the second feature calculates the minimum distance between the given token and a dictionary. In our experience, the first feature is effective for a dictionary containing high-quality items, for example, human-curated protein dictionaries. The second feature is effective for a dictionary that has a large number of items that are not very accurate, for example, web or database lexicons. Details can be found in (Tsai et al., 2006a).

Post-Processing
We count the number of occurrences of a word x appearing in the rightmost position of all NEs in each category. Let the maximum occurrence be n,

and the corresponding category be *c*. The total number of occurrences of *x* in the rightmost position of an NE is *T*; c/T is the consistency rate of *x*. According to our analysis of the training set of the JNLPBA 2004 data, 75% of words have a consistency rate of over 95%. We record this 75% of words and their associated categories in a table. After testing, we crosscheck all the rightmost words of NEs found by our system against this table. If they match, we overwrite the NE categories with those from the table.

2.2 Experiments and Summary

We perform 10-fold cross validation on the GENIA V3.02 corpus (Kim *et al.*, 2003) to compare our CRF-based system with other biomedical NER systems. The experimental results are reported in Table 1. Our system outperforms other systems in protein names by an F-score of at least 2.6%. For DNA names, our performance is very close to that of the best system.

BioNER System	Protein	DNA
Our System (Tsai et al., 2006a)	78.4	66.3
HMM (Zhou et al., 2004)	75.8	63.3
Two Phase SVM (Lee et al., 2003)	70.6	66.4

Table 1. Performance of protein and DNA name recognition on the GENIA V3.02 corpus

We have made every effort to implement a variety of linguistic features in our system's CRF framework. Thanks to these features and the nature of CRF, our system outperforms state-of-the-art machine-learning-based systems, especially in the recognition of protein names.

Our system still has difficulty recognizing long, complicated NEs and coordinated NEs and distinguishing between overlapping NE classes, e.g., cell-line and cell-type. This is because biomedical texts have complicated sentence structures and involve more expert knowledge than texts from the general newswire domain. Since pure machine learning approaches cannot model long contextual phenomena well due to context window size limitations and data sparseness, we believe that template-based methods, which exploit long templates containing different levels of linguistic information, may be of help. Certain errors, such as incorrect boundary identification, are more tolerable if the main purpose is to discover relations between NEs (Tsai *et al.*, 2006c). We shall exploit more linguistic features, such as composite features and external features, in the future. However, machine leaning approaches suffer from a serious problem of annotation inconsistency, which confuses machine learning models and makes evaluation difficult. In order to reduce human annotation effort and alleviate the scarcity of available annotated corpora, we shall learn from web corpora to develop machine learning techniques in different biomedical domains.

3 Biomedical Semantic Role Labeling

In this section, we describe the main steps in building a biomedical SRL system: (1) create semantic roles for each biomedical verb; (2) construct a biomedical corpus, annotated with verbs and their corresponding semantic roles; (3) build an automatic semantic interpretation model, using the annotated text as a training corpus for machine learning. However, on adjunct arguments, especially on those highly relevant to the biomedical domain, such as AM-LOC (location), the performance is not satisfactory. We therefore develop a template generation method to create templates that are used as features for identifying these argument types.

3.1 Biomedical Proposition Bank -- BioProp

Our biomedical proposition bank, BioProp, is based on the GENIA Treebank (Yuka *et al.*, 2005), which is a 491-abstract corpus annotated with syntactic structures. The semantic annotation in BioProp is added to the proper constituents in a syntactic tree.

Basically, we adopt the definitions in PropBank (Palmer *et al.*, 2005). For the verbs not in PropBank, such as "phosphorylate", we define their framesets. Since the annotation is time-consuming, we adopt a semi-automatic approach. We adapt an SRL system trained on PropBank (Wall Street Journal corpus) to the biomedical domain. We first use this SRL system to automatically annotate our corpus, and then human annotators to double check the system's results. Therefore, human effort is greatly reduced.

3.2 Biomedical SRL System -- SEROW

Following (Punyakanok *et al*., 2004), we formulate SRL as a constituent-by-constituent (C-by-C) tagging problem. We use BioProp to train our biomedical SRL system, SEROW (Tsai *et al.*, 2006b), which uses a maximum entropy (ME) machine-learning model. We use the basic features described in (Xue & Palmer, 2004). In addition, we automatically generate templates which can be used to improve classification of biomedical argument types. The details of SEROW system are described in (Tsai *et al.*, 2005) and (Tsai *et al.*, 2006b).

3.3 Experiment and Summary

Our experimental results show that a newswire English SRL system that achieves an F-score of 86.29% can maintain an F-score of 64.64% when ported to the biomedical domain. By using SEROW, we can increase that F-score by 22.9%. Adding automatically generated template features further increases overall F-score by 0.47% and adjunct (AM) F-score by 1.57%, respectively.

4 Conclusion

NER and SRL are two key topics in biomedical NLP. For NER, we find broad linguistic features and integrate them into our CRF framework. Our system outperforms most machine learning-based systems, especially in the recognition of protein names (78.4% of F-score). In the future, templates that can match long contextual relations and coordinated NEs may be applied to NER post-processing. Web corpora may also be used to enhance unknown NE detection. In Bio-SRL, our contribution is threefold. First, we construct a biomedical proposition bank, BioProp, on top of the popular biomedical GENIA treebank following the PropBank annotation scheme. We employ semi-automatic annotation using an SRL system trained on PropBank thereby significantly reducing annotation effort. Second, we construct SEROW, which uses BioProp as its training corpus. Thirdly, we develop a method to automatically generate templates that can boost overall performance, especially on location, manner, adverb, and temporal arguments. In the future, we will expand BioProp to include more biomedical verbs and will also integrate a parser into SEROW.

References

Kim, J.-D., Ohta, T., Teteisi, Y., & Tsujii, J. i. (2003). Genia corpus - a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl. 1).

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Paper presented at the ICML-01.

Lee, K.-J., Hwang, Y.-S., & Rim, H.-C. (2003). Two phase biomedical ne recognition based on svms. Paper presented at the ACL-03 Workshop on Natural Language Processing in Biomedicine.

Morarescu, P., Bejan, C., & Harabagiu, S. (2005). Shallow semantics for relation extraction. Paper presented at the IJCAI-05.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1).

Punyakanok, V., Roth, D., Yih, W., & Zimak, D. (2004). Semantic role labeling via integer linear programming inference. Paper presented at the 20th International Conference on Computational Linguistics (COLING-04).

Tsai, R. T.-H., Chou, W.-C., Wu, S.-H., Sung, T.-Y., Hsiang, J., & Hsu, W.-L. (2006a). Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. Expert Systems with Applications, 30(1), 117-128.

Tsai, R. T.-H., Lin, W.-C. C. Y.-C., Ku, W., Su, Y.-S., Sung, T.-Y., & Hsu, W.-L. (2006b). Serow: Adapting semantic role labeling for biomedical verbs: An exponential model coupled with adapting semantic role labeling for biomedical verbs: An exponential model coupled with automatically generated template features. To appear in BioNLP-2006.

Tsai, R. T.-H., Wu, C.-W., Lin, Y.-C., & Hsu, W.-L. (2005). Exploiting full parsing information to label semantic roles using an ensemble of me and svm via integer linear programming. Paper presented at the CoNLL-2005.

Tsai, R. T.-H., Wu, S.-H., Chou, W.-C., Lin, Y.-C., He, D., Hsiang, J., et al. (2006c). Various criteria in the evaluation of biomedical named entity recognition. BMC Bioinformatics, 7(92).

Xue, N., & Palmer, M. (2004). Calibrating features for semantic role labeling. Paper presented at the EMNLP 2004.

Yuka, T., Yakushiji, A., Ohta, T., & Tsujii, J. (2005). Syntax annotation for the genia corpus.

Zhou, G., Zhang, J., Su, J., Shen, D., & Tan, C. (2004). Recognizing names in biomedical texts: A machine learning approach. Bioinformatics, 20, 1178-1190.