Biomedical Term Recognition With the Perceptron HMM Algorithm

Sittichai Jiampojamarn and **Grzegorz Kondrak** and **Colin Cherry**
Department of Computing Science,
University of Alberta,
Edmonton, AB, T6G 2E8, Canada
{sj,kondrak,colinc}@cs.ualberta.ca

Abstract

We propose a novel approach to the identification of biomedical terms in research publications using the Perceptron HMM algorithm. Each important term is identified and classified into a biomedical concept class. Our proposed system achieves a 68.6% F-measure based on 2,000 training Medline abstracts and 404 unseen testing Medline abstracts. The system achieves performance that is close to the state-of-the-art using only a small feature set. The Perceptron HMM algorithm provides an easy way to incorporate many potentially interdependent features.

1 Introduction

Every day, new scientific articles in the biomedical field are published and made available on-line. The articles contain many new terms and names involving proteins, DNA, RNA, and a wide variety of other substances. Given the large volume of the new research articles, it is important to develop systems capable of extracting meaningful relationships between substances from these articles. Such systems need to recognize and identify biomedical terms in unstructured texts. Biomedical term recognition is thus a step towards information extraction from biomedical texts.

The term recognition task aims at locating biomedical terminology in unstructured texts. The texts are unannotated biomedical research publications written in English. Meaningful terms, which may comprise several words, are identified in order to facilitate further text mining tasks. The recognition task we consider here also involves term classification, that is, classifying the identified terms into biomedical concepts: proteins, DNA, RNA, cell types, and cell lines.

Our biomedical term recognition task is defined as follows: given a set of documents, in each document, find and mark each occurrence of a biomedical term. A term is considered to be annotated correctly only if all its composite words are annotated correctly. Precision, recall and F-measure are determined by comparing the identified terms against the terms annotated in the gold standard.

We believe that the biomedical term recognition task can only be adequately addressed with machine-learning methods. A straightforward dictionary look-up method is bound to fail because of the term variations in the text, especially when the task focuses on locating exact term boundaries. Rule-based systems can achieve good performance on small data sets, but the rules must be defined manually by domain experts, and are difficult to adapt to other data sets. Systems based on machine-learning employ statistical techniques, and can be easily re-trained on different data. The machine-learning techniques used for this task can be divided into two main approaches: the word-based methods, which annotate each word without taking previous assigned tags into account, and the sequence based methods, which take other annotation decisions into account in order to decide on the tag for the current word.

We propose a biomedical term identification

system based on the Perceptron HMM algorithm (Collins, 2004), a novel algorithm for HMM training. It uses the Viterbi and perceptron algorithms to replace a traditional HMM's conditional probabilities with discriminatively trained parameters. The method has been successfully applied to various tasks, including noun phrase chunking and part-of-speech tagging. The perceptron makes it possible to incorporate discriminative training into the traditional HMM approach, and to augment it with additional features, which are helpful in recognizing biomedical terms, as was demonstrated in the ABTA system (Jiampojamarn et al., 2005). A discriminative method allows us to incorporate these features without concern for feature interdependencies. The Perceptron HMM provides an easy and effective learning algorithm for this purpose.

The features used in our system include the part-of-speech tag information, orthographic patterns, word prefix and suffix character strings. The additional features are the word, IOB and class features. The orthographic features encode the spelling characteristics of a word, such as uppercase letters, lowercase letters, digits, and symbols. The IOB and class features encode the IOB tags associated with biomedical class concept markers.

2 Results and discussion

We evaluated our system on the JNLPBA Bio-Entity recognition task. The training data set contains 2,000 Medline abstracts labeled with biomedical classes in the IOB style. The IOB annotation method utilizes three types of tags: <B> for the beginning word of a term, <I> for the remaining words of a term, and <O> for non-term words. For the purpose of term classification, the IOB tags are augmented with the names of the biomedical classes; for example, <B-protein> indicates the first word of a protein term. The held-out set was constructed by randomly selecting 10% of the sentences from the available training set. The number of iterations for training was determined by observing the point where the performance on the held-out set starts to level off. The test set is composed of new 404 Medline abstracts.

Table 1 shows the results of our system on all five classes. In terms of F-measure, our system achieves the average of 68.6%, which a substantial improvement over the baseline system (based on longest string matching against a lists of terms from training data) with the average of 47.7%, and over the basic HMM system, with the average of 53.9%. In comparison with the results of eight participants at the JNLPBA shared tasks (Kim et al., 2004), our system ranks fourth. The performance gap between our system and the best systems at JNLPBA, which achieved the average up to 72.6%, can be attributed to the use of richer and more complete features such as dictionaries and Gene ontology.

Class	Recall	Precision	F-measure
protein	76.73 %	65.56 %	70.71 %
DNA	63.07 %	64.47 %	63.76 %
RNA	64.41 %	59.84 %	62.04 %
cell_type	64.71 %	76.35 %	70.05 %
cell_line	54.20 %	52.02 %	53.09 %
ALL	70.93 %	66.50 %	68.64 %

Table 1: The performance of our system on the test set with respect to each biomedical concept class.

3 Conclusion

We have proposed a new approach to the biomedical term recognition task using the Perceptron HMM algorithm. Our proposed system achieves a 68.6% F-measure with a relatively small number of features as compared to the systems of the JNLPBA participants. The Perceptron HMM algorithm is much easier to implement than the SVM-HMMs, CRF, and the Maximum Entropy Markov Models, while the performance is comparable to those approaches. In the future, we plan to experiment with incorporating external resources, such as dictionaries and gene ontologies, into our feature set.

References

M. Collins. 2004. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

S. Jiampojamarn, N. Cercone, and V. Keselj. 2005. Biological named entity recognition using n-grams and classification methods. In *Proceedings of PACLING*.

J. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of JNLPBA*.