Biological Named Entity Recognition Using n-grams and Classification Methods

Sittichai Jiampojamarn, Nick Cercone and Vlado Kešelj

Faculty of Computer Science, Dalhousie University,
Halifax, NS, B3H 1W5, Canada
{jiampoja, nick, vlado}@cs.dal.ca

Abstract

We propose a biological named entity recognition system which uses classification methods and a n-gram model to annotate terms in text. A novel method is presented to express lexical features in a pattern notation. Prefix and suffix characters are used instead of lists of potential terms or other external resources. Creating classification exemplars is conducted from text by using a word n-gram model. We evaluate our system based on the GENIA version 3.02 corpus which contains 2,000 paper abstracts. The system obtains an 0.705 F-score on exact match term performance. Biological concept markers are also assigned to each located term indicating its meaning. Our system retains simplicity and generalizability.

Key words: Biological named entity recognition, information extraction, word n-gram, and classification algorithms

1 Introduction

Knowledge discovery in biological text focuses on identifying interesting patterns, relations, or meaningful messages from text. Providing users information which they seek for specific questions is desirable. This desire translates into an opportunity for computational linguistics to focus on extracting information from a large collection of articles. Annotating important terms in text such as protein names, DNA, and RNA instance names can facilitate a system that discovers relations between one substance and others. The task is different from traditional information retrieval. While an information retrieval system is to find a list of documents which are relevant to user queries, our task is to extract information from input text and provide users with useful messages by annotating important terms in text. For example, if we wanted to know "what are three missense mutations that have been identified in Alzheimer disease" [1], we would not like to read all articles which are related to missense mutations or Alzheimer disease. Instead, we would like to obtain a list of relationships among those terminology terms leading us to the answers. Even better, we would like a system that provides us the answers: Presenilin-1 (PS1) gene mutations, Amyloid precursor protein (APP), and Presenilin-2 (PS2).

In this paper, we explore machine learning (ML) and natural language processing (NLP) techniques to extract biological terms in unstructured text. We apply classification methods and the n-gram model to annotate biological terms. Given unstructured text in biological research, the annotation system locates biological terms and assigns each locating term with biological concept markers indicating whether the term is "protein", "DNA" or "RNA" instance name.

2 Related work

Biological named entity recognition task is to recognize biological name instances such as protein names, DNA, RNA, and so on. In the biological domain, we are looking for such a system which provides information on, for example, cellular localization, protein-protein interactions, gene regulation and the context of these interactions [2]. The current methodology in biological named entity recognition can be divided into two main ap-

proaches, rule-based methods and learning-based methods.

The PROPER system, introduced by K. Fukada et al. [3], is one of the earliest systems which extract protein names in biological research publications. Based on hand-coded rules, PROPER obtains an 0.967 F-score evaluated on 30 abstracts which are retrieved from MEDLINE database on SH3 domain. It is noticeable that PROPER obtains high performance, 0.967 F-score on 30 abstracts because all hand-coded rules are defined based on observation of the data set. It turns out that the system obtains an 0.47 F-score on a different data set which is used to evaluate the annotation system in Yapex [4].

There were eight participants in the Bio-Entity Recognition Task at JNLPBA. All of these systems were learning-based systems; Support Vector Machines, Hidden Markov Models, Maximum Entropy Markov models, and Conditional Random Fields [5]. The task was evaluated by using the GENIA corpus 3.02 [6] as the training data, and the newly 404 annotated MEDLINE abstracts from the GENIA project as the testing data. Lexical features, affix information (character n-grams), part-of-speech information and previously predicted entity tags were widely used in most of the systems. The best system [7] obtained an 0.726 F-score in overall performance. The system also used dictionary resources, constructed from SWISS-PROT and LocusLink as well as the existing terms in training data.

Our proposed system is based on identifying word position tags which indicate whether a word in text is a beginning, middle or ending word of a multi-word biological term or a single word term. The identifying methodology is based on a classification method. An n-gram model is used to create train and test instances for a classification model.

3 Methodology

Our objective is to annotate biological terms in unstructured text. In our work, the unstructured text is defined as biological research literature written in the English language. The biological named entity recognition task is to identify and classify biological terms in unstructured text. Considering term structure, there are two biological term types: single word terms, and multi-word terms. Locating term boundaries is our main focus because wrong boundaries will give less meaning or misunderstanding. Each identified term is assigned a biological concept marker which we focus on protein, DNA, RNA and miscellaneous concept names.

Word n-gram model Word n-grams are collected using a sliding window of fixed size which starts from the beginning until the end of each sentence. We use classification methods to classify each word n-gram into word position tag classes which indicate whether a word at the mid-point of each word n-gram is the beginning, in between or ending of the term. The class label is considered at the mid-point of each word n-gram; therefore, the classifier will have both pre-word and post-word information to decide which class label should be assigned at the mid-word. Thus, we add "dummy" word, which has "none" feature attribute, to make a word at the beginning and ending of each sentence considered as a mid-word.

Feature attributes From word n grams, we extract feature attributes including part-of-speech tag information, prefix and suffix characters, and word feature pattern. We use a bi-word Hidden Markov Model Part-of-speech tagger [8] to tag each word providing Part-of-speech tag information. The prefix and suffix character features are l characters at the beginning and ending characters of each word. The word feature pattern is a sequence of "uppercases", "lowercases", "digits", and "symbols" occurring in a word. For example, "CD-28-mediated" has the word feature pattern as "uppercase, symbol, digit, symbol, lowercase".

The choice of class labels Biological terms can be divided into two types, a single word term and multi-word term. A single word term indicates that one term consists of only one word; for example, "Adenovirus", "E1A", "tublin", and so on. Another type, multi-word term consists of many words combined as a term which is considered as one substance. It is required that we annotate these multi-words as one term; otherwise, the annotation will mean a different thing. For

example, “clonogenetic fetal calf serum”, “Human adult blood”, “mouse interleukin-1 receptor alpha gene” are multi-word terms. Obviously, we cannot separate these terms at any word point. This will lead to a different meaning. The annotation should cover all words from the beginning until the end of each multi-word biological term. To avoid any ambiguity in annotating single word and multi-word biological terms, we trained the classifier to identify word position tags which are “Beginning”, “Middle”, “Ending”, and “Single”.

These word position tags aid our system to annotate biological terms by rule matching. Only a consistent sequence is annotated as a biological term. We describe this rule using the regular expression as “Beginning (Middle)* Ending”. A single word biological term is indicated by the “Single” class label. Considering biological concept classes, we embed concept classes with the word position classes. Each word position tag is specific to biological class concept it belongs to; for example, “Beginning-Protein”, “Ending-DNA”, etc. To classify a semantic concept marker for each term, we relax the constrained rule by using a majority vote. Clearly, a biological term is annotated by a consistent word tag position sequence of labels “Beginning”, “Middle”, “Ending” and “Single” but it is classified into a biological conceptual class by the most prevalent biological concept marker which is assigned to each word in the term.

4 Evaluation, results and discussion

We evaluate our system using the GENIA version 3.02 corpus which contains 2,000 paper abstracts. We randomly select 1,800 abstracts from the corpus as the training data set which is used to train a classification model and 200 abstracts are used as the testing data set. To annotate biological terms along with semantic classes, we associate each term with the biological concepts: “Protein”, “DNA”, “RNA”, and “Others”.

Several classification method results In the first experiment, we applied several classification algorithms provided by WEKA, machine learning tool [9], to classify each word into word position tags which indicate whether a word is the beginning, middle, or ending of the term. The biological term is located by considering only consistent tag sequences. The results reported in table 1 are obtained by using the ten-fold cross-validation method. We fix $n = 3$ words for the word n-grams model, $m = 4$ character feature pattern, and $l = 4$ for characters both suffix and prefix features. The C4.5 classifier is the best classifier to classify each word into word position tags when comparing with others. In the decision tree, a part-of-speech tag feature is placed at the root of the tree indicating the largest information gain.

The words which belong to the class label “Single” have direct impact to the exact biological matching annotation because the class label “Single” indicates a single word biological term. However, an error which occurs in any one of “Beginning”, “Middle” and “Ending” classes leads the system to annotate multi-word terms incorrectly. Consequently, the accumulating errors influence the exact matching annotation performance as

Classifiers	Word position tags performance on training data (F-score)			
	Beginning	Middle	Ending	Single
Naive Bayes	0.685	0.661	0.704	0.635
C4.5	0.762	0.745	0.810	0.748
SVM	0.567	0.559	0.531	0.556
IBk	0.593	0.625	0.736	0.585
Holte’s OneR	0.196	0.001	0.001	0.424

Table 1. Word position tag performance (F-score) on training data which are evaluated on various classification methods

Classifier	Exact matching annotation performance on testing data		
	Precision	Recall	F-score
Naive Bayes	0.576	0.669	0.619
C4.5	0.748	0.666	0.705
SVM	0.713	0.414	0.524
IBk	0.556	0.558	0.557
Holte’s OneR	0.436	0.146	0.219

Table 2. Exact matching annotation performance on testing data which is evaluated on various classification methods

shown in table 2. In the same way, the C4.5 method performs the best compared with others in exact matching annotation performance.

Recognize terms into biological concepts The results of annotating each biological conceptual class are reported in table 3 and 4. Although the system provides comparable performance of each biological concept in classification word position tag task, the exact matching annotation performance of each biological concept is different. In table 4, the results illustrate the task consistently on training data in term of precision, recall and F-score regarding to overall classes performance. However, the performance on testing set suggests the overfitting problem, whereas the training sets especially on DNA, RNA classes are too small to produce a representative sample of the true classes on testing data.

Biological concept	Classification performance (F-score)			
	Beginning	Middle	Ending	Single
Protein	0.725	0.676	0.814	0.804
DNA	0.731	0.668	0.829	0.620
RNA	0.678	0.542	0.859	0.643
Others	0.767	0.760	0.849	0.735
All classes	0.817	0.791	0.849	0.807

Table 3. The word position tag performance using the C4.5 classification

We evaluate our system on the BioNLP/NLPBA data sets. The results are shown in table 5. Zhou [7], Finkel [10], ABNER [11], and Song [12] system are reported from the recent JNLPBA 2004 workshop. These systems represent the top four systems from the competition. The baseline performance of this workshop is evaluated by collecting a list of entities from the training set, and performing the longest match search for entities through the testing set [5]. Without using any dictionary or other specific domain resource, our system, named "ABTA", is comparable to those systems in term of precision. However, we achieved less performance in recall which led to the lower F-score in overall performance.

Biological concept	Exact matching annotation performance on testing data		
	Precision	Recall	F-score
Protein	0.579	0.596	0.588
DNA	0.585	0.400	0.475
RNA	0.848	0.280	0.421
Others	0.710	0.521	0.601
All classes	0.748	0.666	0.705

Biological concept	Exact matching annotation performance on training data		
	Precision	Recall	F-score
Protein	0.813	0.680	0.740
DNA	0.879	0.50	0.635
RNA	0.908	0.544	0.680
Other	0.839	0.666	0.743
All classes	0.807	0.740	0.772

Table 4. The system performance using the C4.5 classifier to annotate term into biological concepts

System	Precision	Recall	F-score
Zhou [7]	69.4	76.0	72.6
Finkel [10]	68.6	71.6	70.1
ABNER [11]	69.3	70.3	69.8
Song [12]	64.8	67.8	66.3
Baseline [5]	47.6	50.8	49.1
ABTA	64.9	52.6	58.1

Table 5. System comparision within JNLPBA workshop

5 Conclusion

We presented an automatic biological named entity recognition system which uses classification techniques to annotate terms in unstructured text. We explored our system's capabilities using several classification algorithms. Our experiments demonstrated that the C4.5 algorithm is a suitable classification method for annotating biological terms among other algorithms. The classifier classifies each word into a word position tag that indicates whether a word is at a beginning position, in-between position, or ending position, or a single biological term. Based on the word tag positions, only consistent sequences are identified as a biological term. We assign each biological term a biologi-

cal concept marker which is based on a user's interest. In this work, we define the biological concept markers as "protein", "DNA", "RNA" and "other" names. We found that the task of detecting term boundary is hard due to ambiguity of terms, and complexity of the language used in biological research.

The word n-gram model was proposed instead of noun phrase model, which is normally used, because some biological terms are partial noun phrases. Each word n-gram is treated as an instance in the classification model and each instance's class label is the label of the middle word in the n-gram. In this case, the classifier will have both pre-word and post-word information to classify the middle (current) word into a word tag class. Without using any dictionary or other specific domain resource, our system obtains an 0.705 F-score in exact matching on the GENIA 3.02 corpus which contains 2,000 abstracts.

The goal of this work is a system that can annotate biological related terms in research publications. While we explore many techniques to solve the problem of annotating terms in biological publications, there are some issues that can be improved in the future work which include applying biological resources, improving learning methods and defining feature attributes.

Acknowledgement

This work was supported through an NSERC research grant.

References

[1] Flash Med. Answers to medical questions. `http://www.flash-med.com/Topics3.asp?RX_Drug=Alzheimer`, accessed Jan 2005.

[2] Lynette Hirschman, Jong C. Park, Junichi Tsujii, and Limsoon Wong. Accomplishments and challenges in literature data mining for biology. In *BioInfomatics Review, 18(12)*, pages 1553–1561.

[3] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 707–718, 1998.

[4] Kristofer Franzn, Gunnar Eriksson, Fredrik Olsson, Lars Asker Per Lidn, and Joakim Cster. Protein names and how to find them. In *International Journal of Medical Informatics special issue on Natural Language Processing in Biomedical Applications*, pages 49–61, 2002.

[5] Jin-Dong KIM, Tomoko OHTA, Yoshimasa TSURUOKA, Yuka TATEISI, and Nigel COLLIER. 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)*, 2004.

[6] Tsujii laboratory of the University of Tokyo. Genia corpus. `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/`, accessed June 2004.

[7] GuoDong Zhou and Jian Su. 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)*, 2004.

[8] Maciej Ceglowski and Eric Nichols. Lingua::EN::Tagger Part-of-speech tagger for English natural language processing. `http://cpan.uwinnipeg.ca/htdocs/Lingua-EN-Tagger/Lingua/EN/Tagger.html`, accessed Nov 2003.

[9] Ian H. Witten and Eibe Frank. *Data Mining Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, 1999.

[10] Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Gail Sinclair, and Christopher Manning. 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), 2004.

[11] Burr Settles. 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)*, 2004.

[12] Yu Song, Eunju Kim, Gary Geunbae Lee, and Byoung kee Yi. 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)*, 2004.