Computers in Biology and Medicine 37 (2007) 1327–1333

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Rich features based Conditional Random Fields for biological named entities recognition

Chengjie Sun*, Yi Guan, Xiaolong Wang, Lei Lin

School of Computer Science, Harbin Institute of Technology, Mailbox 319, West Da-zhi Street 92, Harbin, Heilongjiang 150001, China

Received 3 June 2006; received in revised form 17 November 2006; accepted 4 December 2006

Abstract

Biological named entity recognition is a critical task for automatically mining knowledge from biological literature. In this paper, this task is cast as a sequential labeling problem and Conditional Random Fields model is introduced to solve it. Under the framework of Conditional Random Fields model, rich features including literal, context and semantics are involved. Among these features, shallow syntactic features are first introduced, which effectively improve the model's performance. Experiments show that our method can achieve an F-measure of 71.2% in an open evaluation data, which is better than most of state-of-the-art systems.

Keywords: Conditional Random Fields; Named entities recognition; Chunking; Sequential labeling problem; Text mining

1. Introduction

With the development of computational and biological technology, the amount of biological literature is increasing fleetingly. MEDLINE database has collected 11 million biological related records since 1965 and is increasing at the rate of 1500 abstracts a day [1]. The research literature is a major repository of knowledge. From them, researchers can find knowledge, such as connections between diseases and genes, the relationship between genes and specific biological functions and the interactions between different proteins and so on.

The explosion of literature in the biological field has provided an opportunity for natural language processing techniques to aid researchers and curators of databases in the biological field by providing text mining services. Yet typical natural language processing tasks such as named entity recognition (NER), information extraction, and word sense disambiguation are particularly challenging in the biological domain with its highly complex and idiosyncratic language.

Biological NER is a critical task for automatically mining knowledge from biological literature. Two special workshops for biological NER BioCreAtIvE [2] (Critical Assessment for Information Extraction in Biology) and JNLPBA [3] (Joint Workshop on Natural Language Processing in Biomedicine and its Applications) were held in 2004 and each of them contained an open evaluation of biological NER technology. The data and guidelines afforded by the two workshops greatly promote the biological NER technology. According to the evaluation results of JNLPBA2004, the best system can achieve an F-measure of 72.6%. This is somewhat lower than figures for similar tasks from the news wire domain. For example, extraction of organization names has been done at over 90% F-measure [2]. Therefore, biological NRE technology needs further study in order to make it applicable.

Current research methods for NER can be classified into three categories: dictionary-based methods [4], rule-based methods [5] and machine learning based methods. In biological domain, dictionary-based methods suffer from low recall due to new entities appearing continually with the advancing biology research. Biological named entities do not follow any nomenclature, which makes rule-based method hard to be perfect. Besides, rule-based method itself is hard to port to new applications. More and more machine learning methods are introduced to solve the biological NER problem, such as Hidden Markov

* Corresponding author. Tel.: +86 451 86413322 89; fax: +86 451 86413322 93.

E-mail address: cjsun@insun.hit.edu.cn (C. Sun).

0010-4825/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2006.12.002

Model [6] (HMM), Support Vector Machine [7] (SVM), Maximum Entropy Markov Model [8] (MEMM) and Conditional Random Fields [1,9] (CRFs).

Biological NER problem can be cast as a sequential labeling problem. CRFs for sequences labeling offer advantages over both generative models like HMM and classifiers applied at each sequence position [10]. In this research, we utilize CRFs model involving rich features to extract biological named entities from biological literature. The feature set includes orthographical features, context features, word shape features, prefix and suffix features, Part of Speech (POS) features and shallow syntactic features. Among these features, shallow syntactic features are first introduced to CRFs model and do boundary detection and semantic labeling at the same time, which effectively improve the model's performance. Although some features have been used by some researchers, we show the effect of each kind of feature in detail, which can afford valuable reference to other researchers. Our method does not need any dictionary resources and post-processing, so it has strong adaptability. Experiments show that our method can achieve an F-measure of 71.2% in JNLPBA test data which is better than most of state-of-the-art systems.

The remainder of this paper is structured as follows. In Section 2, we define the problem of biological NER and introduce its unique characteristics compared to news wire domain. In Section 3, a brief introduction of linear-chain CRFs model is given. In Section 4 we explain the features involved in our method. Experiment results are shown in Section 5. Section 6 is a brief conclusion.

2. Biological NER

Biological NER can be addressed as a sequential labeling problem. It is defined as recognizing objects of a particular class in plain text. Depending on required application, NER can recognize objects ranging from protein/gene names to disease/virus names. In practice, we regard each word in a sentence as a token and each token is associated with a label. Each label with a form of B–C, I–C or O indicates not only the category of a named entity (NE) but also the location of the token within the NE. In this label denotation, C is the category label; B and I are location labels, standing for the beginning of an entity and inside of an entity, respectively. O indicates that a token is not part of an NE. Fig. 1 is an example of biological NER.

Table 1
Biological named entities label list

Meaning	Label		
Beginning of protein	B-protein	Inside protein	I-protein
Beginning of DNA	B-DNA	Inside DNA	I-DNA
Beginning of RNA	B-RNA	Inside RNA	I-RNA
Beginning of cell_type	B-cell_type	Inside cell_type	I-cell_type
Beginning of cell_line	B-cell_line	Inside cell_line	I-cell_line
Others	O		

Biological NER is a challenging problem. There are many different aspects to deal with compared to news wire domain. In general, biological NEs do not follow any nomenclature [11] and can comprise long compound words and short abbreviations. Biological NEs are often English common nouns (as opposed to proper nouns, which, are the nouns normally associated with names) and are often descriptions [12]. For example, some Drosophila (fruit fly) gene names are *blistery*, *inflated*, *period*, *punt and midget*. Some NEs contain various symbols and other spelling variations. On average, any NE of interest has five synonyms. An NE may also belong to multiple categories intrinsically; an NE of one category may contain an NE of another category inside it [13].

In natural language processing domain, Generative Models and Discriminative Models are often used to solve the sequential labeling problem, such as NER. Recently, Discriminative Models are preferred due to their unique characteristics and good performance [14]. Generative Models define a joint probability distribution $p(X, Y)$ where X and Y are random variables, respectively, ranging over observation sequences and their corresponding label sequences. In order to define a joint distribution of this nature, generative models must enumerate all possible observation sequences—a task which, for most domains, is intractable unless observation elements are represented as isolated units, independent from the other elements in an observation sequence. Discriminative Models directly solve the conditional probability $p(Y|X)$. The conditional nature of such models means that no effort is wasted on modeling the observations and one is free from having to make unwarranted independent assumptions about these sequences; arbitrary attributes of the observation data may be captured by the model, without the modeler having to worry about how these attributes are related.

This paper utilizes a Discriminative Model—CRFs to solve biological NER problem. Using the definition in [2], we recognize five categories of entities. There are 11 labels in all using BIO notation mentioned above. All labels are shown in Table 1. Each token in the biological text will be assigned with one of the 11 labels in the recognition results.

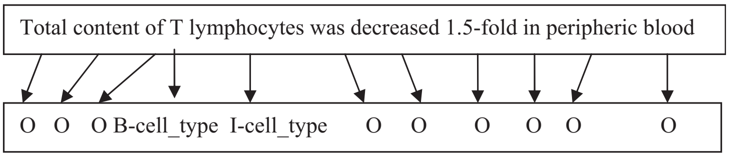
Total content of T lymphocytes was decreased 1.5-fold in peripheric blood
O O O B-cell_type I-cell_type O O O O O O

Fig. 1. An example of biological NER.