CBR-Tagger: a case-based reasoning approach to the gene/protein mention problem

Mariana Neves
Biocomputing Unit
Centro Nacional de Biotecnología - CSIC
Madrid, 28049, Spain
mlara@cnb.csic.es

Monica Chagoyen
Biocomputing Unit
Centro Nacional de Biotecnología - CSIC
Madrid, 28049, Spain
monica.chagoyen@cnb.csic.es

José M. Carazo
Biocomputing Unit
Centro Nacional de Biotecnología - CSIC
Madrid, 28049, Spain
carazo@cnb.csic.es

Alberto Pascual-Montano
Departamento de Arquitectura de Computadores
Facultad de Ciencias Físicas, UCM
Madrid, 28040, Spain
pascual@fis.ucm.es

Abstract

This work proposes a case-based classifier to tackle the gene/protein mention problem in biomedical literature. The so called gene mention problem consists of the recognition of gene and protein entities in scientific texts. A classification process aiming at deciding if a term is a gene mention or not is carried out for each word in the text. It is based on the selection of the best or most similar case in a base of known and unknown cases. The approach was evaluated on several datasets for different organisms and results show the suitability of this approach for the gene mention problem.

1 Introduction

This paper proposes a new method to the gene mention problem by using a case-based reasoning approach that performs a binary classification (gene mention or not) for each word in a text. In a first step cases are stored in two bases (known and unknown cases), followed by a search in these bases for the case most similar to the problem. The classification decision is given by the class of the case selected. The system was developed using Java and MySQL technologies and is available for download as part of the Moara project[1].

2 Proposed method

The method here proposed identifies gene mentions in a text by means of classifying each token into two possible classes: gene mention or not. The system consists of two main steps: the construction of the case bases, and the testing phase, when the test dataset is presented to the system to identify the possible mentions. The words extracted from the training documents were the tokens used to construct the two case bases, one for known cases and the other for unknown cases, as proposed for the part-of-speech tagging problem in (Daelemans, Zavrel, Berck, & Gillis, 1996).

The known cases are the ones used by the system to classify those words that are not new, i.e. those that have were present in the training dataset. The attributes used to represent a known case are the word itself, the class of the word (if it is a gene mention or not), and the class of the preceding word (if it is a gene mention or not).

The system uses a second case base to decide about words that are unknown to the system, i.e. those that are not present in the training set. The attributes of the unknown cases were the shape of the word, the class of the word (if it is a gene mention or not), and the class of the preceding word (if it is a gene mention or not). Note that instead of saving the word itself, a shape of the word is kept in order to allow the system to be able to classify unknown words by means of looking for cases with similar shape. The shape of the word is given by its transformation in a set of symbols according to the type of character found.

In the construction of cases, each word represents a single case, and in order to account for repetitions, the frequency of the case is incremented to indicate the number of times that it appears in the training dataset. The training

[1] http://biocomp.cnb.csic.es/~mlara/moara/index.html

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 108–109,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

documents are read twice, one in the forward (from left to right), and one in the backward (from right to left) directions, in order to allow a more variety of cases. This is important as the classification of a token may be influenced by its preceding and following words.

CBR-Tagger has also been trained with additional corpora in order to better extract mentions from different organisms. These extra corpora are the datasets for gene normalization of the BioCreative task 1B (Hirschman, Colosimo, Morgan, & Yeh, 2005) for to yeast, mouse and fly and the BioCreative 2 Gene Normalization task (Morgan & Hirschman, 2007) for human.

In the classification procedure, the text is tokenized and a sliding window is applied first in the forward and then in the backward direction. In each case, the system keeps track of the class of the preceding token (false at the beginning), gets the shape of the token and tries to find in the bases a case most similar to it. The search procedure is divided in two parts, for the known and unknown cases. Priority is always given to the known cases since it saves the word exactly as they appeared in the training documents and the classification may be more precise than using the unknown cases.

A token already classified as positive by the forward reading may be used for the backward reading as preceding class and might help recognizing mentions composed by many tokens that would not have been totally recognized by one of the reading procedures only. After the identification of the best case for each token, some post-processing procedures are executed to check boundaries (for mentions composed of more than one token) as well as abbreviations and corresponding full names.

3 Results

The results obtained with the BioCreative 2 gene mention task for the CBR-Tagger are shown in Table 1 along with the best result of the competition. Results are showed according to the datasets used for the training of the CBR-tagger: BioCreative 2 Gene Mention task (Wilbur, Smith, & Tanabe, 2007) corpus only (CbrBC2), and the combination of it with the BioCreative task 1B gene normalization corpus (Hirschman et al., 2005) for the yeast (CbrBC2y), mouse (CbrBC2m), fly (CbrBC2f) and the three of them (CbrBC2ymf).

Taggers	*P*	*R*	*FM*
CbrBC2	77.8	75.9	76.9
CbrBC2y	82.7	52.6	64.7
CbrBC2m	83.1	47.1	60.1
CbrBC2f	82.0	65.9	73.0
CbrBC2ymf	82.5	39.7	53.6
Best BC2 result	88.5	86.0	87.2

Table 1: Results for the BC2 gene mention task.

CBR-Tagger has also been applied to the gene normalization problem in conjunction with two other available taggers: Abner[2] and Banner[3]. Table 2 summarizes the best mix of taggers configuration for each organism. Detailed results may be found at the author's research page[4].

Organism	*Best configuration*
Yeast	Abner+CbrBC2
Mouse	Abner+CbrBC2m
Fly	CbrBC2f
Human	Banner+CbrBC2ymf

Table 2: Best taggers for each organism.

Acknowledgments

This work has been partially funded by the Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006, TIN2005-5619. APM acknowledges the support of the Spanish Ramón y Cajal program. The authors acknowledge support from Integromics, S.L.

References

Daelemans, W., Zavrel, J., Berck, P., & Gillis, S. (1996). *MBT: A Memory-Based Part of Speech Tagger-Generator*. Paper presented at the Fourth Workshop on Very Large Corpora, Copenhagen, Denmark.

Hirschman, L., Colosimo, M., Morgan, A., & Yeh, A. (2005). Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics, 6 Suppl 1*, S11.

Morgan, A., & Hirschman, L. (2007). *Overview of BioCreative II Gene Normalization*. Paper presented at the Second BioCreative Challenge Evaluation Workshop, Madrid-Spain.

Wilbur, J., Smith, L., & Tanabe, L. (2007). *BioCreative 2. Gene Mention Task*. Paper presented at the Second BioCreative Challenge Evaluation Workshop, Madrid, Spain.

[2] http://pages.cs.wisc.edu/~bsettles/abner/

[3] http://banner.sourceforge.net/

[4] http://biocomp.cnb.csic.es/~mlara/mention.html