Constructing a Dictionary of Biological Terms for Information Extraction

Yasunori Yamamoto
yas@ims.u-tokyo.ac.jp
Yoshihiro Ohta
yoh@ims.u-tokyo.ac.jp
Ikuo Uchiyama
uchiyama@ims.u-tokyo.ac.jp
Toshihisa Takagi
takagi@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1, Shirokanedai, Minato-ku, Tokyo 108 Japan

Abstract

In information extraction (IE) systems a keyword dictionary which is a kind of knowledge base on domain specific information is very important. We are developing technologies that construct a keyword dictionary for IE.

1 Introduction

In recent years, an amount of information in genome area has been growing rapidly. Although many data about biomolecular structures have been comprehensively compiled into public databases such as GenBank, SWISS-PROT, and GDB, most of data as biological functions such as molecular interactions which are crucial for the next stage of genome science, are still only in the literatures. Therefore, information extraction (IE) from biological papers is an important theme for computational biology.

A dictionary of domain specific terms is one of the central knowledge sources of IE systems[1][2]. In this paper, we present a strategy for constructing a dictionary of biological terms such as protein names. With our strategy at first, terms categorized in a specific class are collected from the public databases and transformed into appropriate forms. We call this initial collection a base dictionary. Next, context of each term in the actual literatures are examined and hierarchical term clustering is performed. The resultant clusters can be used as a thesaurus for IE.

2 Base dictionary construction

In biological papers, gene names, protein names or disease names often work as keywords. These terms can be collected easily from the public databases. However, their description forms in the databases are not always consistent with those in the literatures. In case of protein names, this problem is more complicated because they often consist of multiple compound words. To solve this problem, bigram analysis of the protein names in the database is carried out. Information which is incorporated into the base dictionary is two types: 1)Extracted collocations and 2)other words which compose protein names selected by frequency in a corpus.

3 Term clustering based on context comparison

Once the base dictionary is constructed, context around each term can be analyzed by locating them into the corpus sentences. Here, frequencies of nouns cooccuring with each term in the same sentence are considered. Based on these frequencies, Hierarchical Bayesian Clustering (HBC)[3] is performed to cluster terms. Briefly, HBC merges a pair of clusters in such a way that the maximum posterior probability $P(C|D)$ is obtained, where $P(C|D)$ is a probability that a collection of data D is classified into a set of clusters C. This procedure is originally developed for text classification, but here we apply this method to the term categorization.

4 Experiments and results

We obtained 49,340 protein names from the 'DE' fields in the SWISS-PROT database to construct a base dictionary. A corpus containing 169,532 sentences were collected from MEDLINE abstracts. Tagger by Brill [4] was used for tagging sentences after protein names in the base dictionary were marked.

By collocations analysis, our system found several useful collocations such as "zinc finger protein". These collocations were used for constructing the base dictionary. Each term in this base dictionary was marked in the corpus sentences and 181 most frequent terms were clustered by HBC. The result was interesting such that proteins relating to cancer were formed as a cluster.

Acknowledgement

This work was supported in part by a Grant-in-Aid (08283103) for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] William B.Frakes,Ricardo Baeza-Yates, "Information Retrieval," *PRENTICE HALL*, 1992.

[2] Stephen Soderland, David Fisher, Jonathan Aseltine, Wendy Lehnert, "CRYSTAL: Inducing a Conceptual Dictionary," *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence(IJCAI'95)*, 1995.

[3] Iwayama Makoto, Tokunaga Takenobu, "Hierarchical Bayesian Clustering for Automatic Text Classification," *Proceedings of the International Joint Conference on Artificial Intelligence(IJCAI'95)*, 1995.

[4] Eric Brill, "Some Advances in Transformation-Based Part of Speech Tagging," *Proceedings of the Twelfth National Conference on Artificial Intelligence(AAAI-94)*, 1994.