Information Finding from Biological Papers

Yoshihiro Ohta [1]
yoh@ims.u-tokyo.ac.jp
Yasunori Yamamoto [2]
yas@cs.titech.ac.jp
Ikuo Uchiyama [1]
uchiyama@ims.u-tokyo.ac.jp
Toshihisa Takagi [1]
takagi@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science,
The University of Tokyo
Shiroganedai, Minato-ku, Tokyo 108, Japan

[2] Graduate School of Information Science and Engineering,
Tokyo Institute of Technology
Ookayama, Meguro-ku, Tokyo 152, Japan

Abstract

We have developed computer technologies for a system that extracts domain specific knowledge from human written biological papers. This system consists of two components, Information Retrieval (IR) and Information Extraction (IE). We propose a query modification method using automatically constructed thesaurus for IR and a statistical keyword prediction method for IE. Although by a purely statistical model with no heuristics, the experimental result has shown the good performance.

1 Introduction

As the volume of biological information increases, the demand for the domain specific knowledge base grows. When a researcher constructs the knowledge base, however, the task of finding necessary information from the amount of papers is labor-intensive and time-consuming.

In this paper, we present the method to support biologists who are willing to construct a knowledge base. The task here is divided into two parts, IR and IE. IR returns papers that is relevant to a user's needs, and IE returns a structured representation of information within the retrieved paper. With this method, we could collect the interesting papers from text database in IR, and automatically extract keywords from retrieved papers in IE (See Figure 1).

2 Information Retrieval

When researchers search for interesting papers, they prepare query and throw it against text database, for example MEDLINE. Retrieval query is often defined using some documents or keywords in user's mind that express their interest [1].

It is almost impossible, however, to collect enough papers from text database to construct knowledge base by such initial query. We propose a method of query modification, where the initial query is expanded and term reweighted by the term distance matrix obtained in the way of constructing thesaurus [2].

3 Information Extraction

IE starts with a collection of papers retrieved in IR, then transforms them into information that is more rapidly digested and analyzed [3]. The extraction process is divided into three phases: (1) Predict keywords in a collocation level (2) Classify keywords using the linguistic pattern mined from corpus (3) Extract knowledge from keywords and neighbors of them. In this paper, we focus on keyword prediction.

Although the knowledge-based approach, using heuristics, has been proved effective for information extraction on limited domains, there are difficulties for constructing a large number of domain-specific patterns. With this reason, we bring the purely statistical model using large corpus. Because the domain specific term often appears as collocation, we define the score of collocation to resolve the ambiguity of compound words recognition.

4 Experimental Results

As subject of our analysis, we have used *the Signaling Pathway Database*[4]. Using our modified query, good retrieval performance, in terms of *recall* and *precision*, was shown in IR phase experiments. In IE, keywords are extracted automatically in a collocation level. Additionally, this method doesn't depend on domain, it is possible to port the system to a new domain.

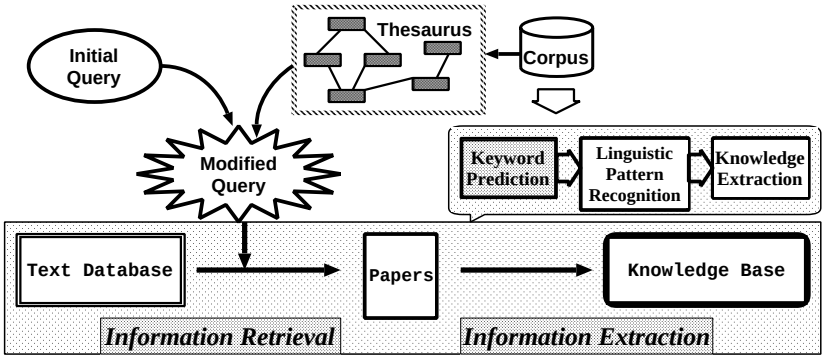

Figure 1: A model of information finding

Acknowledgment

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] G. Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," *Addison-Wesley*, 1988.

[2] T.Tokunaga, M. Iwayama, and H. Tanaka, "Automatic thesaurus construction based on grammatical relations," *In Proceedings of the International Joint Conference on Artificial Intelligence*, 1995.

[3] J. Cowie and W. Lehnert, "Information Extraction," *Communications of the ACM*, Vol. 35, No. 12, pp. 80-91, 1996.

[4] N, Tateishi, H. Shiotari, S, Kuhara, T, Takagi and M, Kanehisa, "An integrated database SPAD(Signaling PAthway Database) for signal transduction and genetic information," *Proceedings of Genome Informatics Workshop*, 1995.