The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers

Nigel Collier, Hyun Seok Park, Norihiro Ogata
Yuka Tateishi, Chikashi Nobata, Tomoko Ohta
Tateshi Sekimizu, Hisao Imai, Katsutoshi Ibushi, Jun-ichi Tsujii
{nigel,hsp20,ogata,yucca,nova,okap,sekimizu,hisao,k-ibushi,tsujii}@is.s.u-tokyo.ac.jp
Department of Information Science, Graduate School of Science
University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan

Abstract

We present an outline of the genome information acquisition (GENIA) project for automatically extracting biochemical information from journal papers and abstracts. GENIA will be available over the Internet and is designed to aid in information extraction, retrieval and visualisation and to help reduce information overload on researchers. The vast repository of papers available online in databases such as MEDLINE is a natural environment in which to develop language engineering methods and tools and is an opportunity to show how language engineering can play a key role on the Internet.

1 Introduction

In the context of the global research effort to map the human genome, the Genome Informatics Extraction project, GENIA (GENIA, 1999), aims to support such research by automatically extracting information from biochemical papers and their abstracts such as those available from MEDLINE (MEDLINE, 1999) written by domain specialists. The vast repository of research papers which are the results of genome research are a natural environment in which to develop language engineering tools and methods.

This project aims to help reduce the problems caused by information overload on the researchers who want to access the information held inside collections such as MEDLINE. The key elements of the project are centered around the tasks of information extraction and retrieval. These are outlined below and then the interface which integrates them is described.

1.1 Terminology identification and classification

Through discussions with domain experts, we have identified several classes of useful entities such as the names of proteins and genes. The reliable identification and acquisition of such class members is one of our key goals so that terminology databases can be automatically extended. We should not however underestimate the difficulty of this task as the naming conventions in this field are very loose.

In our initial experiments we used the ENGCG shallow parser (Voutilainen, 1996) to identify noun phrases and classify them as proteins (Sekimizu et al., 1998) according to their cooccurrence with a set of verbs. Due to the difficulties caused by inconsistent naming of terms, we have decided to use multiple sources of evidence for classifying terminology.

Currently we have extended our approach and are exploring two models for named entity recognition. The first is based on a statistical model of word clustering (Baker and McCallum, 1998) which is trained on pre-classified word lists from Swissprot and other databases. We supplemented this with short word lists to identify the class from a term's final noun if it existed in a head final position. In our first experiments on a judgement set of 80 expert tagged MEDLINE abstracts the model yielded F-scores for pre-identified phrases as follows: 69.35 for 1372 source entities, 53.00 for 3280 proteins, 66.67 for 56 RNA and 45.20 for 566 DNA. We expect this to improve with the addition of better training word lists. The second approach is based on decision trees (Quinlan, 1993), supplemented with word lists for classes derived from Swissprot and other databases. In these tests the phrases for terms were not pre-identified. The model was trained on a corpus of 60 expert tagged MEDLINE abstracts and tested on a corpus of 20 articles yielding F-scores of: 55.38 for 356 source, 66.58 for 808 protein entities. The number of RNA

and DNA entities was too small to train with.

As part of the overall project we are creating an expert-tagged corpus of MEDLINE abstracts and full papers for training and testing our tools. The markup scheme for this corpus is being developed in cooperation with groups of biologists and is based on a conceptual domain model implemented in SGML. The corpus itself will be cross-validated with an independent group of biologists.

1.2 Information extraction

We are using information extraction methods to automatically extract named entity properties, events and other domain-specific concepts from MEDLINE abstracts and full texts. One part of this work is the construction and maintenance of an ontology for the domain which is executed by a system which we are now developing called *Ontology Extraction-Maintenace System (OEMS)*. *OEMS* extracts three types of information about the domain-ontology, (Ogata, 1997), called *typing information*, from the abstracts: *taxonomy* (a subtype structure), *mereology* (a part-whole structure), *synonymy* (an identity structure). Eventually we hope to be able to identify and extract domain specific facts such as protein-protein binding information from full texts and to aid biochemists in the formation of cell signalling diagrams which are necessary for their work.

1.3 Thesaurus building

A further goal of our work is to construct a thesaurus automatically from MEDLINE abstracts and domain dictionaries consisting of medical domain terms for the purpose of query expansion in information retrieval of databases such as MEDLINE, e.g. see (Jing and Croft, 1994). We are currently working with the Med test set (30 queries and 1033 documents) on SMART (e.g. see (Salton, 1989),(Buckley et al., 1993)). Eventually we plan on building a specialised thesaurus for the genome domain but this currently depends on the creation of a suitable test set.

1.4 Interface

A key aspect of this project is providing easy interaction between domain experts and the information extraction programs. Our interface provides a link to the information extraction programs as well as clickable links to aid in querying for related information from publically available databases on the WWW within a single environment. For example, a user can highlight proteins in the texts using the named entity extraction program and then search for the molecule structure diagram.

2 Conclusion

This paper has provided a synopsis of the GENIA project. The project will run for a further two years and aims to provide an online demonstration of how language engineering can be useful in the genome domain.

References

L.D. Baker and A.K. McCallum. 1998. Distributional clustering of words for text classification. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia.*

C. Buckley, J. Allan, and G. Salton. 1993. Automatic routing and ad-hoc retrieval using SMART: TREC-2. In D. K. Harman, editor, *The second Text REtrieval Conference (TREC-2)*, pages 45–55. NIST.

GENIA. 1999. Information on the GENIA project can be found at:. http://www.is.s.u-tokyo.ac.jp/~nigel/GENIA.html.

Y. Jing and W. Croft. 1994. An association thesaurus for information retrieval. In *Proceedings of RIAO'94*, pages 146–160.

MEDLINE. 1999. The PubMed database can be found at:. http://www.ncbi.nlm.nih.gov/PubMed/.

Norihiro Ogata. 1997. Dynamic constructive thesaurus. In *Language Study and Thesaurus: Proceedings of the National Language Research Institute Fifth International Symposium: Session 1*, pages 182–189. The National Language Research Institute, Tokyo.

J.R. Quinlan. 1993. *c4.5 Programs for Machine Learning.* Morgan Kaufmann Publishers, Inc., San Mateo, California.

G. Salton. 1989. *Automatic Text Processing - The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Publishing Company, Inc., Reading, Massachusetts.

T. Sekimizu, H. Park, and J. Tsujii. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. In *Genome Informatics.* Unviersal Academy Press, Inc.

A. Voutilainen. 1996. Designing a (finite-state) parsing grammar. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing.* A Bradford Book, The MIT Press.