

Joint proposal for a tutorial: Mining the Biomedical Literature

Dietrich Schuhmann, Lynette Hirschman,
Junichii Tsuji, Sophia Ananiadou,
Alfonso Valencia

LION Bioscience AG, Waldhofer Strasse 98, 69123 Heidelberg, Germany
Tel.: +49-6221-4038-179, email: dietrich.schuhmann@lionbioscience.com

February 13, 2001

Contents

1 Organisational issues:	1
1.1 The Plan:	2
1.2 Modifications to the plan:	2
2 Abstract	3
3 Introduction (Lynette Hirschman)	3
4 Integrating information extraction and text mining: state of the art (Lynette Hirschman)	4
5 Ontology building and text annotation for information extraction in the field of molecular biology (Junichi TSUJII)	4
6 Automatic Term Recognition for Molecular Biology (Sophia Ananiadou)	5
7 Information extraction in molecular biology: tools, evaluation of retrieval and linguistical particularities (Alfonso Valencia, Dietrich Schuhmann)	5

1 Organisational issues:

ISMB has become one of the main reference points for the work in mining in the biomedical literature. Key papers and posters in this area have been presented in ISMB since 1996. The main teams working in this area used to be active in ISMB and the Pacific Symposium Conferences. All this scientific activity will benefit from a complementary activity in teaching and cross-fertilization between more theoretical and biological approaches. Something particularly needed in this area where the fusion of very different approaches has still not completely emerged.

1.1 The Plan:

This is a proposal for teaching text mining and information extraction to biologists, bioinformaticians and informaticians. The presentations of Dr. Lynette Hirschman and Dr. Sophia Ananiadou are closer to the pure technology, and the presentations of Prof. Tsujii, Prof. Alfonso Valencia and Dr. Dietrich Schuhmann are closer to the biology.

- Integrating information extraction and text mining: state of the art
Dr. Lynette Hirschman, The Information Technology Center, MITRE Corporation, USA
- Ontology building and text annotation for information extraction in the field of molecular biology
Prof. Junichi Tsujii, Department of Computer Science, University of Tokyo, Tokyo, Japan
- Automatic term recognition for molecular biology
Dr. Sophia Ananiadou, Salford University, Manchester, Uk
- Information extraction in molecular biology: tools, evaluation of retrieval and linguistic particularities
Dr. Alfonso Valencia, CNB, Campus de la Universidad Autonoma Cantoblanco, Madrid, Spain
Dr. Dietrich Schuhmann, LION Bioscience AG, Heidelberg, Germany

The five partners work in different projects together.

Lynette Hirschman, Junichi Tsujii, Sophia Ananiadou and Alfonso Valencia are well-known experts in the field of text mining and information extraction. They have numerous publications and have extensive teaching experience on the University level.

Dietrich Schuhmann is running a large research project in the field of text mining in molecular biology, which is funded by the German government. Within one year he was able to implement different software components, which establish text mining at LION Bioscience AG. He has several years of teaching experience on a school level.

1.2 Modifications to the plan:

The four parts can be combined in different ways and the authors will take care, that links between the different talks will be quite visible during the tutorial. For example the more general information of Lynette Hirschman will be introductory to the other presentations, and the results of Sophia Ananiadou address a core problem, which is readdressed in the work of Junichii Tsujii and the one of Schuhmann / Valencia.

The more general and theoretical aspects come from issues in text mining and in assessment of the state of the art, and will explore the cross fertilization between natural language processing tools and corpus-based methods using annotated data, with specific attention to extraction of biological terms. The more practical and concrete aspects come from the usage of ontologies in the tools, from the problems with the curation process, and last but not least from the presentation and discussion of existing tools.

The proposal could serve as a single tutorial, or as two half-day tutorials, the first focused on the status and applicability of the technology, and the second focused on the biological implications. Altogether we expect students to gain from details such as the materials that will be presented to the students, some references to papers, books or monographies that could be used for the students, and a clear description of what part will be practical and what part theoretical, including the hardware support that will be required for the practical part.

2 Abstract

This tutorial will survey the state of the art of tools for accessing the information in the biomedical literature. There will be four parts to the tutorial. The first part (Hirschman) will survey the state of the art of the tools used for natural language processing. We will describe the availability of these tools, their performance, and the issues in "retargeting" the tools to new genres (journal abstracts, instead of news report), and new domains (biology instead of crisis management or business news). The second part (Tsujii) will focus specifically on ontology building and text annotation in biological texts. The third part (Ananiadou) will focus specifically on term extraction from biology text. The fourth part (Valencia, Schuhmann) will look specifically at the application of information retrieval techniques and how these can be applied to identify relevant information in the scientific literature in molecular biology.

Our goal is to provide the participants with a survey of what tools exist, how these tools perform, what resources are available for the biomedical domain, and how these tools can be applied to help the working biologist and bioinformatics researcher.

3 Introduction (Lynette Hirschman)

Researchers in bioinformatics are often faced with the task of searching the literature for specific kinds of information, such as homologous genes, genes with similar functions, or members of a particular metabolic pathway. In doing so, a researcher may create collections of annotated documents, or even an annotated database to support these information needs. This can be a tedious and frustrating task; to date, there has been little automated support for such searches. There is now growing interest on the part of bioinformatics researchers in finding ways to improve the ability to find relevant documents in the literature, and to extract specific kinds of information from these documents. This tutorial will describe the state of the art in applying natural language tools and methods to the biomedical literature.

The bioinformatics community is developing the necessary underpinnings for rich annotations, including ontologies, nomenclature and XML-based data exchange standards. For syntax, XML is becoming the standard of choice, both for natural language processing and for biology. For semantics, the issues are more complex. For consistent annotation, it is necessary to define exactly what things should go into a particular class to be identified as gene "names" or protein "names" when they occur in free text. Fortunately, the biology community is addressing these issues with work on nomenclature and ontologies. Annotation work in the data management community is addressing techniques for efficient storage; the tutorial will provide pointers to this work. By integrating syntax, semantics, and the appropriate ontologies and nomenclatures, expressed via emerging XML standards, the natural language processing tools can locate and mine information in free text. These techniques can be coupled with Information retrieval techniques to provide improved topic clustering and topic detection, and also to provide question answering - the ability to find specific facts (rather than pointers to documents) within a large collection of articles.

4 Integrating information extraction and text mining: state of the art

(Lynette Hirschman)

The central focus of this section of the tutorial will be to review the status of work in natural language processing. The natural language community has developed a range of tools that can be integrated to locate and extract specific kinds of information from free text, such as journal articles or abstracts, or comment fields in databases. There now exist commercial systems as well as open source software modules that can be used to extract proper names and numerical expressions. There are also techniques drawn from machine learning and statistical processing that allow these systems to be retrained to identify biological entities such as gene names or proteins. However, these modules must be trained from a corpus of annotated data - that is, examples of occurrences of such entities in the relevant literature. The requirement for training such a module is that there be sufficient numbers of examples consistently and reproducibly annotated.

In addition to identifying classes of entities, there are research tools that can identify simple relations among entities of interest, e.g., a gene codes for a protein, and complex relations, e.g., an enzyme regulates the rate of protein production.

The tutorial will review what tools exist, and how to integrate them for different applications. We will then discuss issues of retraining the tools for a new domain (biology) in terms of amount of training data, consistency of annotation standards (in terms of human interannotator agreement), and expected performance of modules, given sufficient training data. Finally, we will review how these tools are combined with information retrieval techniques to provide question answering: the ability to provide an answer to a question by coupling question analysis, information retrieval and information extraction techniques.

5 Ontology building and text annotation for information extraction in the field of molecular biology

(Junichi TSUJII)

There have been substantial attempts of text annotation in general fields (newspaper articles, etc.) for information extraction and other NLP applications. However, scientific papers in general and their abstracts in particular have many characteristics that make text annotation more troublesome. In order to provide an appropriate set of semantic tags, we also have to develop domain specific ontology that is consistent with their textual realization.

This part of the tutorial consists of :

1. Survey of work on ontology: Several key projects of ontology building for natural language processing will be surveyed.
2. Ontology for biology: The group of Tokyo University, the NLP group in Computer Science and the Genome Informatics Group in Medical Science Institute, is now building ontology for Cell-Signal Pathway and mapping terms (around 1 million terms) into this ontology. Using this work as an example, we will illustrate what ontology looks like, what difficulties one encounter in building ontology for biology, and how our tool (TMIS) helps domain specialists and linguists to build ontology.

3. Text Annotation for biology: Different from proper nouns such as persons, company names, etc., most of terms in biology refer to classes, instead of individuals. While abbreviations are frequent, there are many systematic metonymies. These imply that terms in the fields are highly ambiguous, i.e. the annotators have to consider contexts when they assign semantic tags to occurrences of terms. Following the first part of the tutorial by Lynette Hirschman, we illustrate these specific difficulties by examples and show how they can be resolved.
4. Information extraction (IE) from abstracts: A sample IE system for Cell-Signal Pathway will be demonstrated, which uses the domain ontology.

6 Automatic Term Recognition for Molecular Biology (Sophia Ananiadou)

With the overwhelming increase of textual information in molecular biology, there is a need to assist domain experts in gathering and classifying information. Existing scientific databases are a source of knowledge but they suffer from terminological confusion. In order to obtain terminology in a consistent manner, to update and maintain specialized lexical resources, we need tools, which automatically acquire and classify terminology.

In this tutorial we will present the main techniques from NLP and machine learning which have been used for automatic term recognition and classification. These are information retrieval based, statistical, hybrid (combination of linguistic and statistical information), decision tree based and HMM. We will discuss how to extract terminology using these techniques and evaluate the performance of some systems.

We will focus only on term extraction systems based on molecular biology. We will also discuss how the output of a good term recognizer can provide more than just a list of words, i.e. how we can exploit contextual information around terms for term disambiguation and clustering. Term clustering can be used for annotation purposes and ontology building.

Finally we will demonstrate our system ATRACT developed within the BioPath project. ATRACT is an interface combining automatic term extraction (based on C/NC value) and term clustering which can be parameterized according to the biologist's application needs. Finally, we will discuss issues of evaluation and what to expect from term recognition and clustering system in molecular biology.

7 Information extraction in molecular biology: tools, evaluation of retrieval and linguistical particularities (Alfonso Valencia, Dietrich Schuhmann)

This part will focus on the different steps which are currently undertaken to go from the literature and from the database information towards curated information. More precisely to show the necessities, that have to be put in place to come up with reasonable results from the biological point of view. It gives an overview about different strategies, shows software components that have been put in place and shows the data at different levels of the extraction and curation process.

The Name Problem:

In biology we deal with objects which have evolved over a period of time. This leads to the consequence, that e.g. genes and proteins are not well-defined according to their names. Sequences

are similar, but may encode functionally different proteins, or functional related proteins are encoded by unrelated sequences. This has led to inconsistencies in the naming of these objects. If we compare the same 'object' across different species, we again come up with different results. Furthermore, names for genes and proteins have to be found, which on the one hand offer good categorization capabilities and on the other hand take into consideration the mentioned variability. The conclusion is, that text mining in biology has to cope with the problem of the definition of gene and protein entities by name and with name disambiguation.

The working process:

Different methods are used to do text mining in biology. Among these methods we can distinguish matching techniques to do precise information extraction on the one hand and "ignorance based" techniques to do classification and clustering on the other hand. Both techniques are in a way complementary, but do a good job, if they are put together (see talk of Sophia Ananiadou). We can demonstrate the different steps of information extraction. This starts out with tagging, NP extraction and leads into rule based information extraction. For the last step the user or team defines templates for the expected information structure. Any template matches to patterns, that are used in the text to find the information. The definition and identification of a pattern requires a trained person. Once a pattern has matched, the information can be put to a database according to the template. Special tools with statistical components and components from natural language processing support the pattern definition and the information extraction. Afterwards the curator can go over the retrieved data and can distinguish the correct hits from the hits with a low quality. This again needs special tools to catch the curators opinion and to have a supervisor in case of inconsistencies.

A short discussion on linking information together and ontologies:

The textmining and curation process offers information, which is related to biological models, that the community of writers and biologists had/has in mind. These models are also part of other software components: ontologies and databases, and have to be consistent with the data that is provided from laboratory methods. From the practical as well as from the philosophical point of view, these different systems now seem to grow together.