

Mining the Biomedical Literature

What is the expected gain, if we process scientific text automatically?

Scientific publications form the major tool for the exchange of new scientific facts. Therefore, scientists have to process large amounts of publications, without being able to read everything. This explains, why automatic methods for the analysis of scientific text become very popular. If an automatic method could extract the details needed with good quality, the scientist would save a lot of time reading. If an automatic method could retrieve and pre-select at 100% certainty the relevant documents from the complete set of documents available, then the scientist would save a lot of time searching for the best publications, but he still would have to read them to find the facts.

Text mining methods have been developed and applied in all fields, where large amounts of data were available: analysis of large archives of news documents, internet documents and in recent years scientific publication in the biomedical field. The strong increase in publications in the biomedical field is the consequence of the strong increase of information gathered throughout numerous sequencing projects.

What is text mining (TM)?

The term 'text mining' refers to a variety of techniques which fulfill different tasks mainly:

- Information retrieval (IR): Such automatic methods extract the features of a document (words, parts of sentences), classify them and then compare them to a profile, which consists in general of a list of terms. If the document features match to the profile, then the document is selected.
- Information extraction (IE): Such automatic methods seek written evidence of facts in the text, which are in general represented by combinations of words (patterns). Such patterns select a sequence of words and take their morphological changes into consideration. Other techniques include grammar rules to 'understand' the sentence (Natural Language Processing (NLP)) or complex classification techniques, e.g. Support Vector Machines (SVM) or Bayesian Networks, to identify patterns of high complexity.

Other approaches use similar techniques to archive documents for better retrieval or to answer questions on the basis of existing documents.

What data is used for TM?

Scientific abstracts from Medline (PubMed) form the standard data resource. In addition scientific groups analyze complete scientific publications. Patent text from the European Patent Office is an additional source.

The importance of correct terminology

The use of correct terminology is necessary for the identification of the object behind the described fact and for the correct link between the literature and scientific databases. Unfortunately natural language is often ambiguous in the use of terminology. The biological community is addressing these issues with work on nomenclature and ontologies. The following ambiguities are most prominent.

- Syntactic ambiguity occurs when a word has a noun or a verb interpretation.
- Semantic ambiguity occurs when two identical terms have different meanings. Disambiguation has to take the context of the term into consideration.

Which advantages result from natural language processing (NLP) ?

A variety of different systems is available to do natural language processing, but no system can cope with the complexity of natural language in a spoken language. In addition, the time for the analysis of a sentence increases with its complexity. As a conclusion, NLP based systems provide solutions on chunks of sentences. The main advantage in NLP based approaches can be found in the fact that sentence analysis can be shaped following a hand-crafted linguistic model and thus they do not rely on the correct annotation of a training data set.

Text Mining for 2Can

Which TM tasks are important in bioinformatics ?

The most prominent information extraction task is the extraction of protein-protein interactions from Medline (numerous publications). Other tasks are concerned with the association of a gene to a phenotype, e.g. a disease or a symptom, or the identification of a gene in conjunction with a mutation. In addition, research groups are working on the correct identification of protein and gene names from Medline, which is again a non-trivial task due to the complexity of the applied names.

Evaluation of extracted results

The evaluation quantifies how many correct facts were selected from all facts available (= recall) and how many correct facts are contained in all selected facts (= precision). The sweat part of this task is to generate a random sample from the complete set of documents and to count how many facts are contained in the sample. Other evaluation methods compare the extracted information by the systems with independent information stored in databases and large experimental repositories.