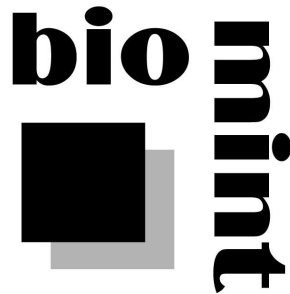


Deliverables Report
QLRI-2002-02770 BioMinT
February 2004



Deliverable Report 10.5 (Year 1)
Yearly technology watch report

Authors: **Compiled on behalf of the consortium by**
 Kristof Van Belleghem
 PharmaDM

Status: **Final**

Distribution: **Consortium**

Version: **1.0**

Checkers:

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

PROJECT MANAGER

Name: Professor Terri Attwood

Address: Bioinformatics Group, School of Biological Sciences,

Stopford Building University of Manchester, Oxford Road, Manchester, M13 9PT

Phone Number: +44 161 275 5082

Fax Number: +44 161 275 5082

E-mail: attwood@bioinf.man.ac.uk

TABLE OF CONTENTS

<u>1. Executive Overview</u>	3
<u>2. State of the Art in Text Mining</u>	3
<u>2.1 Natural Language Processing</u>	3
<u>2.2 Information Retrieval</u>	4
<u>2.3 Information Extraction</u>	5
<u>2.4 Text Mining</u>	5
<u>3. Recent Text Mining Systems</u>	5
<u>3.1 PreBind/Textomy</u>	5
<u>3.2 Textpresso</u>	6
<u>3.3 Pasta</u>	6
<u>4. Resources</u>	7
<u>4.1 Annotated Corpora</u>	7
<u>4.1.1 GENIA</u>	7
<u>4.1.2 Integrated Annotation of Biomedical Text at Pennsylvania Univeristy</u>	7
<u>4.2 Ontologies and Vocabularies</u>	8
<u>4.3 Assessments through competitions</u>	9
<u>5. Conclusion</u>	9

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

1. Executive Overview

During the BioMinT project, we evidently want to stay informed about the state of the art and any new developments in text mining, information extraction, information retrieval and natural language processing. The yearly reports give a brief overview of the state of the art and of recent developments in the field.

Especially for the 2003 report, we try to give an overview of the state of the art, with somewhat more stress on recent work. The following reports of 2004 and 2005 should then merely add the newest developments of those years. We try and discuss not only algorithms, techniques and new complete systems, but also useful resources that can be used to boost any text mining system.

2. State of the Art in Text Mining

The research on text mining can be seen as a merge of various semi-independent research areas, like Natural Language Processing, Information Retrieval, and Information Extraction. We discuss the current state of the art in each of these fields here.

2.1 Natural Language Processing

Natural language processing is concerned with finding syntactic and semantic information in free text through various analysis steps. Typically these steps include

- tokenization : breaking the text up into units, in particular sentences and words. Many algorithms exist for this task.
- part of speech tagging : labeling words with their semantic content type, e.g. *Noun*, *Verb* etc.. Two main approaches exist to this task: on the one hand probabilistic approaches based on Hidden Markov Models, and on the other hand rule-based approaches which take context and morphological information into account. The BioMinT tagger relies on yet another technique called Memory-based Learning, where tags are determined based on the previously seen examples nearest to the case to be decided. All types of approach rely on appropriate pre-tagged training corpora.
- chunking : grouping syntactically related words, such as a noun and a preceding adjective, together into higher-level chunks, such as noun phrases and prepositional phrases. In BioMinT, a memory-based learner is used for this step.

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

- (shallow) parsing : finding the partial or complete syntactic structure of a sentence. Shallow parsing typically involves tokenization, tagging and chunking. Full parsing goes beyond chunking, finding ever higher levels of structure until it has determined the structure of the entire sentence. Full parsing is often accomplished using a grammar of rewrite rules, whereas shallow parsing can more easily be tackled by various learning methods.
- named entity recognition : finding those words or groups of words in a sentence that describe a member of a specific category (e.g. proteins). Several approaches exist, some just relying on extensive dictionaries, others using e.g. morphological or context information. It depends on the specific entity type which approach is most appropriate.
- anaphora resolution : finding out what term in the preceding text words like “this”, “it” etc. refer to. Typically knowledge-based approaches are useful here, especially for resolving anaphora of the type “the *X*” where *X* is a term describing a class of objects (e.g. *gene*).

2.2 Information Retrieval

Information Retrieval is the task of identifying those documents that are most relevant to a particular need, within a large document set. The need can be expressed in a few different forms, tied to the approaches for identifying the right documents.

The simplest approach is the boolean keyword query, where index files of the text collection are searched for the specified keywords, and results combined using the logical operators AND, OR, NOT as specified in the query.

A more advanced approach uses the Vector-Space model, wherein documents and queries are represented as vectors over terms. A distance measure over these vectors is then used to find the most similar documents to a query. Both the mapping of documents to term vectors and the distance measure come in many flavours and levels of granularity.

Other approaches try to alleviate the dependency on the specific search terms, e.g. using probabilistic models of documents and queries, or, based on the vector model, building matrices from document vectors and finding classes of terms with a common “hidden semantics” using singular value decomposition.

One specific subfield of information retrieval deals with text categorization, i.e. labeling documents with thematic categories. Two main approaches are the manual rule coding approach, where a set of rules is defined encoding expert knowledge, and a machine learning approach, where documents are classified based on what has been learned from a training set of labeled documents.

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

2.3 Information Extraction

Information extraction is the fine-grained task of finding the actual facts and events described in a document. Typically, IE heavily uses NLP techniques, up to full or (more often) shallow parsing. The results of this analysis can be combined with domain knowledge (in the form of dictionaries or ontologies) to find the facts of interest. Ideally, domain knowledge can also be used to help in anaphora resolution, to find out exactly what a reference like “the protein” refers to earlier in the text.

2.4 Text Mining

Text mining is the entire automated process of discovering information in unstructured natural language text. Many different systems exist for different purposes. Typically, a text mining system needs an information retrieval component to find relevant documents for processing, and an information extraction component for extracting the relevant facts as structured data. Once the data is available in structured form, data mining techniques can be applied to find new knowledge.

3. Recent Text Mining Systems

In this section we describe some recent text mining systems, in particular PreBind/Textomy and Textpresso.

3.1 PreBind/Textomy

In 2003, the PreBind/Textomy system was introduced ([2]). This is a text mining system with a very similar setup and goal as the BioMinT tool. The system is used as a database curation aid for the BIND database ([1]) of protein-protein interactions. It contains an information retrieval component, an information extraction part, and a domain knowledge module.

The information retrieval component uses a Support Vector Machine classifier to distinguish PubMed abstracts that discuss protein-protein interactions from those that do not. As a first step in information extraction, this same classifier is used to find the interaction-related sentences in the selected documents. No deep linguistic analysis is performed on the sentences, i.e. the information extraction component is strictly based on co-occurrence of terms of the right types. Lists of protein names and synonyms are derived from public databases, and are used as domain knowledge, in combination with morphological and contextual rules, to find candidate interacting proteins. Interaction phrases are found in a similar fashion. The selected abstracts, with the

Deliverables Report

QLRI-2002-02770 BioMinT

February 2004

highest-scoring sentences highlighted, can be inspected by a human curator, who validates every interaction.

The authors of the system expect improved results from a use of deeper linguistic analysis. However, a human validation step will always remain necessary. Also, they observe that to get a reasonable recall of “real” interactions (tested against MIPS, an independent interaction database), it is necessary to mine full papers rather than abstracts.

3.2 Textpresso

Another text mining system is Textpresso, available on the web at <http://www.textpresso.org/>. This is a quite specialised web application for *C. elegans* literature, which is continually being extended with new features and new documents. It now contains over 15000 abstracts and over 3000 full papers from various sources, which should be about 35 relevant literature.

The system supports information retrieval for simple boolean keyword queries, but also has a more advanced query form where categories (e.g. localization) can be specified i.o. keywords. Between 30 and 40 categories are defined by a fixed set of terms, both *entity* categories and *relationship* categories. Some categories have additional attributes (e.g. pathway type - cellular) or specifications attached to their terms, further separating them into a kind of sub-categories, which can also be used in the query. In all, this defines an ontology of limited depth.

Matching is based on simple co-occurrence (one can specify if the match must be within one sentence or within the full text). To facilitate this, all texts are marked up with the relevant categories and attributes beforehand.

Information extraction is apparently planned as the next extension, but is not yet available as of the time of this writing. Another apparently planned extension is a deeper ontology.

According to the available information, the entire system is not using any learning module, i.e. all information markup is added to the texts manually.

In this system as well, the authors stress the importance of access to full papers rather than abstracts.

3.3 Pasta

A slightly older system was developed in the *PASTA* (Protein Active Site Template Acquisition) project (1998-2001), and has recently been described in the *Journal of Bioinformatics* ([3]). *PASTA* is an information extraction system that uses sophisticated NLP techniques to extract information on the roles of amino acid residues in protein active sites. The system architecture consists of several components or modules that combine with each other to perform to-

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

kenization, part-of-speech tagging, named entity recognition, parsing, discourse interpretation (i.e. successive sentences are integrated into a model of the text, a process which is guided by a pre-existing model of the domain), and template extraction. The templates are used to fill a relational database. More data on the project can be found on the project's web site, <http://www.dcs.shef.ac.uk/research/groups/nlp/pasta/>.

4. Resources

The advancement of the text mining field is not only dependent on new systems and techniques, but also on underlying resources, in particular ontologies/dictionaries, training material in the form of annotated corpora, and good assessment methods. Here we list the most important of those resources.

4.1 Annotated Corpora

4.1.1 GENIA

Since many systems depend on the availability of large volumes of good training data, it is essential to generate reliable corpora. Several efforts are underway to generate such corpora for the biomedical domain, the most important one being made in the context of the GENIA project (<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>).

The key task addressed by GENIA is the extraction of event information about protein interactions. To this end - and to provide much-needed annotated language resources for biomedical informatics - the project is developing a corpus of MEDLINE abstracts, which is being marked-up for terms from a domain-specific ontology, as well as for other types of linguistic knowledge, e.g. part-of-speech tags. Updates of the corpus are regularly made available on the website. This corpus and the markup ontology are used as a starting point for annotating texts in the BioMinT project.

4.1.2 Integrated Annotation of Biomedical Text at Pennsylvania Univeristy

A more recent effort to build an annotated corpus was started in 2003 at the university of Pennsylvania ([4]). This project focuses on an integration of various forms of annotation, in particular that of syntactic information (Treebank annotation), predicate-argument structure (Propbank annotation), domain entities and (co)reference information. Large sets of both biomedical abstracts and full-text articles are being annotated. The annotation follows a bootstrapping approach to speed up the process. First results are expected in early 2004.

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

4.2 Ontologies and Vocabularies

Another important aid in text mining is domain knowledge, often provided in the form of vocabularies or ontologies. Many efforts are being made around the world to capture information about the entire biomedical domain, or specific subsets of it, in such a structured form. The most important ontologies for the BioMinT project seem to be the following:

- Gene Ontology (<http://www.geneontology.org>) The goal of the Gene Ontology Consortium is to produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO contains ontologies for Molecular Function, Biological Process and Cellular Component.
- MeSH (<http://www.nlm.nih.gov/mesh/meshhome.html>) MeSH is the National Library of Medicine's controlled vocabulary thesaurus. Its scope is very broad, and it is perfectly integrated in the Pubmed/MEDLINE search facility, but not everyone is satisfied with the actual contents, which could do with some better cleaning/structuring in many places.
- ICD (<http://www.cdc.gov/nchs/icd9.htm>) ICD is the International Classification of Diseases, basically designed to promote international comparability in the collection and classification of mortality statistics, and often outdated, but nevertheless a comprehensive system for diseases.
- Enzyme Nomenclature (<http://www.chem.qmul.ac.uk/iubmb/enzyme/>) A list and classification of enzymes, with links to many other information sources.
- UMLS (<http://www.nlm.nih.gov/research/umls/>) The UMLS (Unified Medical Language System) Metathesaurus is a very large metathesaurus built from thesauri, classifications, coding systems, and lists of controlled terms that are developed by different organizations, including e.g. MeSH and ICD. UMLS adds to this the SPECIALIST lexicon containing syntactic, morphological, and orthographic information on medical terms, and various software tools.

Apart from ontologies, many public databases can be used as terminology aids, in particular by providing synonym terms for proteins. For example, for the BioMinT query expansion a synonym database has been compiled from the public sources LocusLink, SwissProt, Flybase, GDB, HUGO, MGD, OMIM, RGD, Ratmap, SGD, TAIR, WormBase, SubtiList and EcoGene. (For more details, see deliverable D3.1.)

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

4.3 Assessments through competitions

BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology, described on <http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html>) is an initiative to set up an assessment of text mining systems in biology, by defining a common, biologically meaningful task, common data sets and a clearly defined evaluation. This initiative follows traditions of competition in NLP and bioinformatics. Results for the first two tasks will become available by the end of march, and it is likely that more similar initiatives will follow.

5. Conclusion

This description of developments in the field of biological text mining should be a good overview, but can of course not be comprehensive. However, the following pointers should lead to lots of additional information.

A late 2003 overview paper on text mining, *Mining the Biomedical Literature in the Genomic Era: An Overview*, was written by Hagit Shatkay and Ronen Feldman ([5]). This paper discusses many issues in more detail, and contains more references to the relevant literature.

Similarly, the web site <http://www.cis.upenn.edu/~mamandel/term.html> contains an enormous overview of resources for biomedical terminology and ontologies.

A global portal site for any news on natural language processing of biology text can be found on <http://www.bionlp.org/>.

References

- [1] G. D. Bader, D. Betel, and C. W. Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Res*, 31(1):248–250, 2003.
- [2] I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G. D. Bader, K. Michalickova, T. Pawson, and C. W. Hogue. Prebind and textomy mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4(11), 2003.
- [3] R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett. Protein structures and information extraction from biological texts: the pasta system. *Bioinformatics*, 19(1):135–143, 2003.

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

- [4] S. Kulick, M. Liberman, M. Palmer, and A. Schein. Shallow semantic annotation of biomedical corpora for information extraction. In *Proceedings of the 2003 ISMB Special Interest Group Meeting on Text Mining (a.k.a. BioLink)*, 2003.
- [5] H. Shatkay and R. Feldman. Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(3):821–855, 2003.