



THE WHITE ROSE GRID

e-Science Centre of Excellence

AMBIT - Acquiring Medical and Biological Information from Text

Providing intelligent access to large and unstructured biomedical data resources

The Problem: Significant results and findings are being generated in the biological and bio-medical fields at an ever increasing rate. In many cases these results are available solely in the scientific literature. However, the volume of text being published and the speed with which new results appear make it impossible for researchers to read and correlate all possible relevant sources. A similar point can be made regarding hospital patient records. These records contain valuable information to support longitudinal and epidemiological patient studies, but due to the large number of records and their textual nature, this information is effectively inaccessible.

The Solution: Information Extraction technology, based on Natural Language Processing (NLP) methodologies, can be used to identify important terms in text and significant relations between them. In this way, an unstructured data source, i.e. text, can yield structured information amenable to computational processing and reporting.

The University of Sheffield NLP group is participating in two e-Science projects, both of which contain biomedical text extraction components. The first of these projects is the EPSRC-funded **myGrid** E-biologist's workbench project, where a text extraction service will be provided as a stand-alone product and a workflow component. The second project, the MRC-funded **CLEF** (Clinical E-Science Framework), aims to provide clinicians and researchers with a structured overview of patient records to help in cancer studies.

The **myGrid** project will provide transparent access to distributed bio-informatics processes and resources. Providing access both to raw text and extracted information is a key part of the biological research process and hence a vital component of the project.

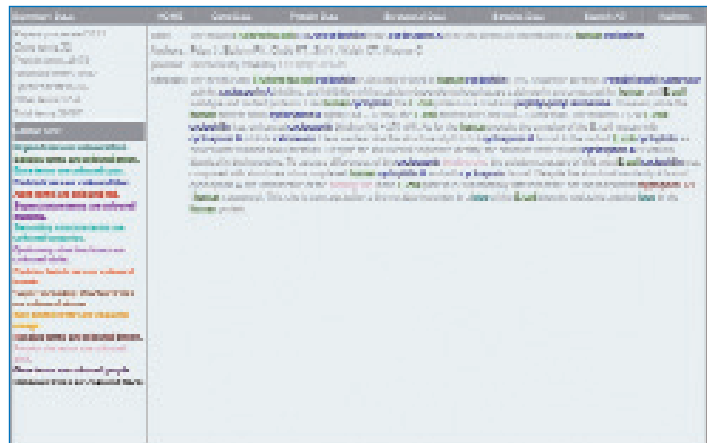


Figure 1: Marked-up entities in a Medline abstract as presented by the AMBIT Results Server

The **CLEF** project specifically addresses the data bottleneck in the clinical domain. The goal of this project is to apply Information Extraction techniques for automatic identification and extraction of key information from cancer patient records. The extracted, structured information is stored in a repository, which can be queried by medical researchers.

The System: Despite the different objectives for text extraction within the myGrid and CLEF projects, many of the challenges they face are the same. To address these commonalities we are building a shared Natural Language Processing infrastructure: **AMBIT**, a system for Acquiring Medical and Biological Information from Text.



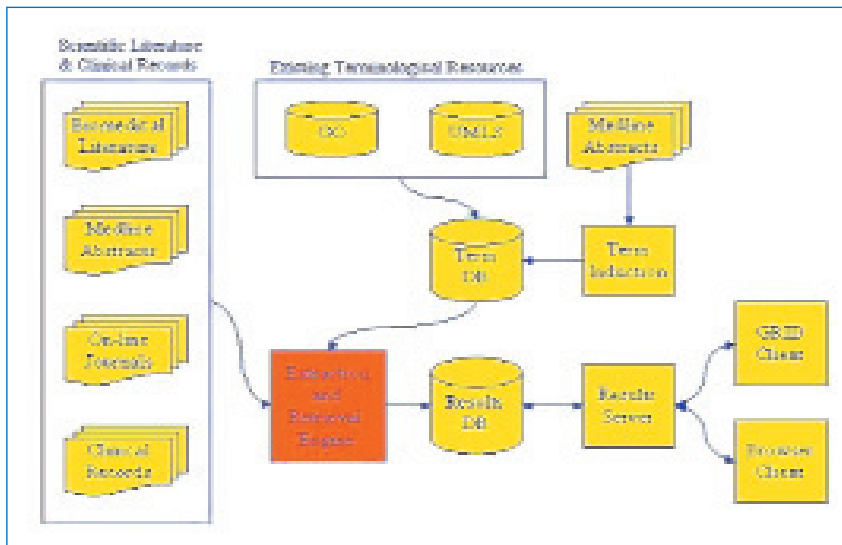


Figure 2: General Architecture of the AMBIT system

The first step in Information Extraction is to identify key entities, such as genes, proteins, diseases etc. In highly technical domains entity identification requires sophisticated terminology recognition capabilities. The recognition task is made difficult by the sheer number of technical terms, the constant introduction of new terms, and the fact that there is no widely accepted nomenclature in the biomedical field. Any given concept can have multiple terms and any given term may appear in different papers referring to different concepts. To address this problem we are building a very large scale terminological repository covering as much of the biomedical domain as possible.

In addition to entity recognition, information extraction applications in the biomedical domain need to identify relations between entities, for example, that a specific residue occurs in a specific protein, or that a cancer occurs in a specific organ. Relations of interest occur across a broad range of biomedical text and techniques for extracting relations, which we are extending from those developed in

the BBSRC-funded PASTA project, apply to a broad range of relations.

Further information

Contacts from the University of Sheffield, Department of Computer Science:

Prof R Gaizauskas,
email: R.Gaizauskas@dcs.shef.ac.uk

Dr M Hepple,
email: M.Hepple@dcs.shef.ac.uk

Dr N Davis,
email: N.Davis@dcs.shef.ac.uk

Dr Y Guo,
email: G.Yikun@dcs.shef.ac.uk

Dr H Harkema,
email: H.Harkema@dcs.shef.ac.uk

Mr A Roberts,
email: A.Roberts@dcs.shef.ac.uk

Mr I Roberts,
email: I.Roberts@dcs.shef.ac.uk

The relevant web pages are at:

1. <http://nlp.shef.ac.uk/mygrid/>
2. <http://nlp.shef.ac.uk/clef/>

