AMBIT: Acquiring Medical and Biological Information from Text

Robert Gaizauskas, Mark Hepple, Neil Davis, Yikun Guo,
Henk Harkema, Angus Roberts, Ian Roberts

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
{initial.surname}@dcs.shef.ac.uk

Abstract

We introduce and motivate the AMBIT system for extracting information from biomedical texts, which is currently under development at the University of Sheffield for use within two ongoing E-Science projects (myGrid and CLEF).

1 Introduction

Significant experimental and clinical results and findings are being reported in the biological and bio-medical fields at an ever increasing rate. In many cases these results are available solely from the scientific literature, and so tend to be unstructured and not amenable to automated processing. The sheer volume of text being published and the speed with which new results appear, make it all but impossible for researchers to read and correlate all of the possible relevant sources in their research area. A similar point can be made regarding information in clinical records. This information could be invaluable to both clinicians and clinical researchers, but is in practice unavailable because of the labour required to identify, extract and collate it from disparate and unstructured sources. If some way were available to generate structured information that is amenable to automated processing from these unstructured data sources, the research biologist/clinician's practice could be made considerably easier and more productive.

Information Extraction technology, based on Natural Language Processing (NLP) methodologies, can be used to identify important entities referred to in text and also significant relations between these entities. The results of this analysis can be stored in a database for subsequent access, or can be used as a basis for intelligent indexing of documents for retrieval. In this way, an unstructured data source, i.e. text, can yield structured information that can be accessed rapidly and effectively by researchers. Furthermore, by processing large numbers of papers it is possible that new relationships will be discovered between entities, both within a single discipline, and perhaps more importantly, between disciplines (cf., e.g, Swanson and Smalheiser [1997]). Such a capability would be a very valuable research aid indeed.

To this end we propose to build a natural language processing infrastructure for processing biomedical text: AMBIT, a system for Acquiring Medical and Biological Information from Text.

2 Biomedical Text Extraction Research at Sheffield

The University of Sheffield NLP group is currently participating in two E-Science projects, both of which contain a significant biomedical text extraction component. Both of these projects are building upon generic information extraction technology specialised for work in bioinformatics applications in a previous BBSRC-supported project which aimed to extract information about protein active sites from the literature on structural molecular biology [Gaizauskas et al., 2003].

2.1 myGrid

The first of these E-Science projects is the EPSRC-funded myGrid project, which is developing an E-biologist's workbench. myGrid uses a workflow architecture, in which individual components are provided as web services communicating via SOAP, presenting the end

user with a single unified workbench through which component services can be accessed using the workflow model. See Goble et al. [2003] for more details of myGrid.

A text extraction service will be provided to myGrid both as a stand-alone product, for browsing the available scientific literature, and a workflow component, to provide extracted information as part of a research protocol. We are also exploring the notion of ambient text, whereby the system will observe the research biologist's activities and make relevant text available in the periphery of her visual workspace.

MyGrid text services comprise an off-line and an on-line component. The off-line component involves pre-processing a large biological sciences corpus, in this case the contents of Medline, to identify various biological entities and relationships and store them in an SQL database. The on-line component supports access to this extracted information, as well as to the raw texts, via a SOAP interface to the SQL database.

2.2 CLEF

The second E-science project, the MRC- funded Clinical E-Science Framework (CLEF) project, is aimed at integrating information for the clinical e-Scientist. The overall goal of the project is to provide a repository of structured and well-organized clinical information which can be queried and summarized both for biomedical research and clinical care (Rector et al. [2003]). Patient notes written by clinicians document the long-term course of patients' illnesses and treatments and therefore contain valuable information to support longitudinal and epidemiological patient studies. These patient notes are, however, in unstructured, textual form. For this information to be included in the repository it must first be extracted from the notes. The large volume of such notes means that information extraction must be automatic rather than manual. Given its use of clinical notes, containing very personal patient information, CLEF must also concern itself with issues security and confidentiality. Regarding confidentiality, exploratory work has been carried out on using NLP methods to provide automatic support for pseudonymisation of clinical records.

3 The AMBIT System

Despite the different objectives for text extraction within the myGrid and CLEF projects, many of the technical challenges they face are the same. For example, both projects require extensive capabilities to recognise and classify biomedical entities as described using complex technical terminology in text. Similarly, both projects need to recognise or acquire information from text about certain types of biomedical relations, e.g. structural or locative relations. Furthermore, both projects need to interface via web services or the Grid with other large systems within which they may form a component. As a consequence we have decided to construct a general framework for the extraction of information from biomedical text: AMBIT, a system for acquiring biological and medical information from text. An overview of the AMBIT architecture is shown in Figure 1. In the following we briefly discuss each of the principal components of this architecture.

Interface Layer The data made available by the system is currently accessed via a SOAP web services interface. There are currently two such interfaces available, one implemented in JAVA (using the Apache AXIS toolkit) and a second in PHP (using the Nusoap toolkit). However the replacement of these with a single PERL SOAP interface (using the PERL::lite toolkit) is being investigated for closer integration with the query engine. An OGSA-DAI compliant Grid interface will be developed as need arises.

Query Engine The AMBIT Query Engine will allow users to access information from the database that contains both raw and annotated text using a number of methods. These methods will include traditional free text search, but will also allow search based on the structured information produced by information extraction, so that queries may refer to specific entities and classes of entities, and specific kinds of relations that are recognised to hold between them. In addition, the Query Engine will interact with the terminology engine, both to provide terminological support for query formulation, and also to allow users (and alse remotely located calling programs) to access information from various lexical resources that is integrated in the terminological database, e.g. such as connections between a given terminological item and various ontologies.

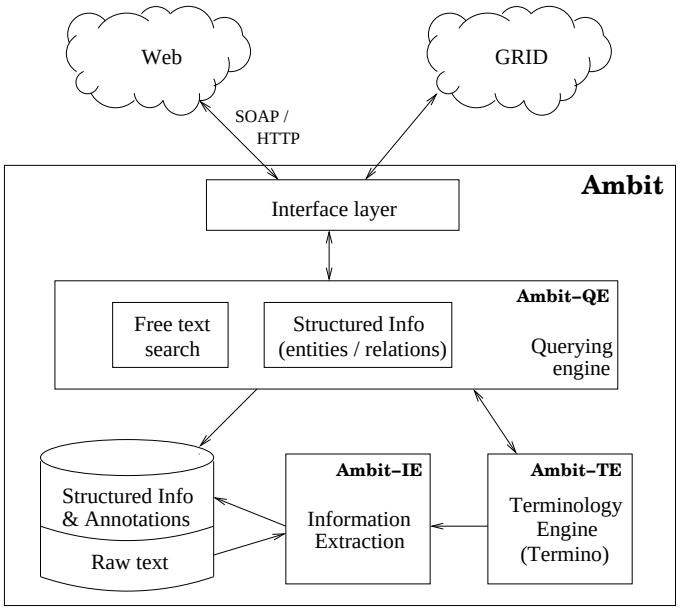

Figure 1: AMBIT Architecture

Raw and Annotated Text DB The raw and annotated texts are held in a set of relational tables in an RDBMS. The AMBIT database is implemented in MySQL version 4.0.13. and currently holds approximately 12 million text records (MedLine abstracts), of which only a portion have been fully processed by the IE engine.

Information Extraction Engine The AMBIT information extraction engine has been adapted from the PASTA IE system described in Gaizauskas et al. [2003]. The system comprises three major stages: lexical and terminological processing, syntactic and semantic processing, and discourse processing. The first stage identifies and classifies bio- medical entities (currently 15 different entity types are recognised) that occur in a given text, such as diseases, drugs, genes etc. using a finite state term recognizer and a term parser. The term recogniser is generated by the Terminology Engine described below. The second stage produces a (partial) syntactic and semantic analysis for each sentence in the text. The third step integrates these results into a discourse model that resolves inter-sentential coreferences and represents the final semantic content of the text. The information is then read from the discourse model and stored in templates. These are structured objects representing the 15 domain specific entities, relevant events (e.g., for CLEF, investigation and intervention) and their relationships (e.g. drug side effect, which relates a drug and a problem). The templates are mapped into the structured information repository, where they can be queried via the query engine or used to generate summaries.

Terminology Engine The first step in automatic information extraction is the identification of entities, such as genes, proteins, conditions, body parts etc. In highly technical domains, such as the biomedical field, entity identification requires sophisticated terminology recognition capabilities. The recognition task is further complicated by the sheer number of technical terms, the constant introduction of new terms, and the fact that there is no widely accepted single nomenclature in the biomedical field. A given concept can have multiple terms referring to it and a given term string 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. The fundamental element of this terminological database is a so-called instance. An instance can be thought of as a (normalized) string associated with various kinds of information, including syntactic information, such as part of speech and morphological class, and semantic information, such as domain of relevance and location in some specified ontology of the

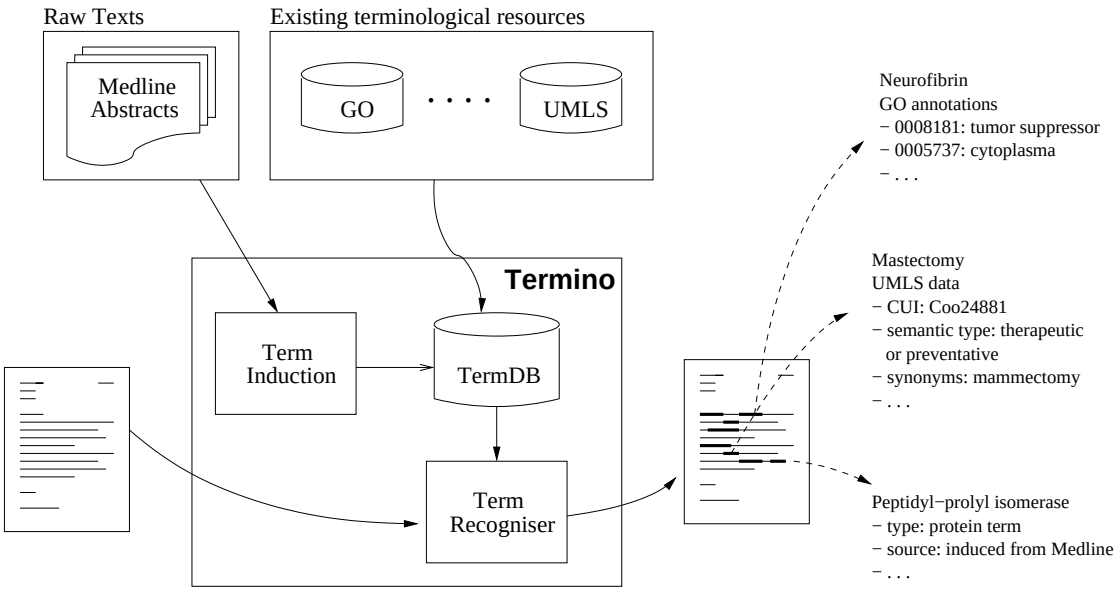

Figure 2: Populating and using the terminological database

concept denoted by the string. The database also contains links to connect synonyms to one another and to connect abbreviations to their full forms. Instances are either imported from existing knowledge sources, e.g. the Unified Medical Language System, or mined from online raw text sources, e.g. Medline. The strings in the database are compiled into a finite state term recognizer, which is run over the texts. A successful match returns a key which gives access to all the information about the string that is stored in the database. In this way, the system has uniform access to terminological knowledge from various sources. Figure 2 illustrates how the terminological database is populated and used The TermDB is also implemented in MySQL (v4.0.13) and currently holds 50 thousand instances in many categories ranging from diseases to proteins.

4 Concluding Remarks

The objective of AMBIT is to provide a powerful, general purpose system for mining information from biomedical texts, be they published abstracts or full texts, or the textual component of clinical records. The results of this system will be made available via a web services or Grid-based interface that will support both interactive user browsing of, and remote progam access to, mined results. To date an initial version of AMBIT has been implemented and functionality and coverage of the system is now being extended. Ultimately AMBIT will be successful if it gives working research biologists or clinicians more effective access to the invaluable information found only in textual data.

References

R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett. Protein structures and information extraction from biological texts: The PASTA system. *Journal of Bioinformatics*, 19(1):135–143, 2003.

C.A. Goble, C.J. Wroe, R. Stevens, and the myGrid consortium. The mygrid project: services, architecture and demonstrator. In *This volume*, 2003.

A. Rector, J. Rogers, A. Taweel, D. Ingram, D. Kalra, J. Milan, R. Gaizauskas, M. Hepple, D. Scott, and R. Power. Joining up healthcare with clinical and post-genomic research. In *This volume*, 2003.

D.R. Swanson and N.R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183–203, 1997.