Information Extraction from Clinical Records

Henk Harkema, Ian Roberts, Robert Gaizauskas, Mark Hepple

Department of Computer Science, University of Sheffield, UK
{h.harkema, i.roberts., r.gaizauskas, m.hepple}@dcs.shef.ac.uk

Abstract

Much of the wealth of information that exists in patient clinical records is difficult or impractical to access, due to it being recorded in unstructured textual formats and the large volume of records available. To facilitate access to this information, we introduce AMBIT: a text analysis system designed to extract key information from clinical and biomedical text. Information derived in this way, and stored in a structured format, can be used in various ways to assist in the provision and development of clinical care. In this paper we discuss the architecture and functionality of AMBIT, and present evaluation results regarding its performance on an information extraction task in the medical domain.

1. Introduction

Clinical records contain a wealth of information, largely in textual form, enhanced access to which could serve a host of purposes. Patient notes written by clinicians document the long-term course of patients' illnesses and treatments and contain information vital to immediate patient care and which could also support longitudinal and epidemiological studies. However, the textual format of these notes and their volume makes it difficult to survey even a single patient's complete record. To aggregate over the records of groups of patients of the size required to carry out clinical research is clearly an even greater task, and is in general not practically feasible. Methods automating access to this content would greatly enhance the capabilities of clinical researchers and clinicians.

To address these concerns and similar issues regarding access to scientific biomedical literature, we introduce AMBIT: a processing framework for Acquiring Medical and Biological Information from Text. Information Extraction (IE) technology, based on natural language processing methodologies, enables AMBIT to identify important entities referred to in a single document and also significant relations between these entities. References to the same entity across documents can also be detected, allowing information to be integrated across multiple sources. 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.

AMBIT is being developed within the context of two e-Science projects: CLEF and myGrid. The myGrid project aims to present research biologists with a single unified workbench through which component bioinformatics services, including services for text mining over biomedical abstracts, can be accessed using a workflow model (see [9] for further details). The goal of the Clinical e-Science Framework (CLEF) project is to provide a repository of structured and well-organized clinical information which can be queried and summarised for biomedical research and clinical care [8]. Specifically, we are addressing the requirement to extract information regarding the treatment of cancer patients. The treatment of such patients may extend over several years and the resulting clinical record may include many documents, such as case notes, lab reports, discharge summaries, etc. We aim to identify a number of significant classes of entities, including patients, drugs, problems (i.e., symptoms and diseases), investigations and interventions, and relationships between such entities, e.g, that an investigation has indicated a particular problem, which, in turn, has been treated with a particular intervention.

The information derived in this way can be integrated in a number of ways, for different purposes. The results of analysing a single patient's notes might be integrated to provide a *chronicle*, a summary of a patient's condition and treatment over time. Such a summary might be of value for direct care, e.g., for a clinician newly taking over a patient's treatment, and in the context of clinical trials, as a means to identify suitable patients for recruitment. The needs of clinical researchers might also be addressed by aggregating information across

```
<LOCUS-3> :=                               <SIGN-6> :=
   NAME:          ‘upper lobe’                NAME:    ‘collapse’
   CODE:          ‘C34.1’
   LATERALITY:    ‘right’                  <LOCATION-5> :=
                                              LOCUS:  <LOCUS-3>
                                              SIGN:   <SIGN-6>
```

Figure 1: Sample template structure

multiple patients, producing data that could be used in epidemiological and other studies. Aggregated results might also be useful to policy makers and healthcare managers in regard to planning and clinical governance. The construction of a chronicle requires an arrangement of the extracted information according to the temporal course of events in the patient’s treatment and condition. This temporal dimension of information extraction in the medical domain is discussed further in [6].

2. Architecture and Functionality

AMBIT consists of various engines and components, which are discussed briefly in the following paragraphs. The *information extraction engine* identifies pre-defined classes of entities and relationships in natural language texts and stores this information in a structured format. The engine has been adapted from the PASTA IE system [3]. The information extraction process comprises three major stages: lexical and terminological processing, syntactic and semantic processing, and discourse processing. The first stage identifies and classifies relevant entities that occur in a text, using a finite state term recogniser and a term parser. The term parser builds longer terms from shorter terms identified by the recogniser according to a given term grammar. The term recogniser is generated by the terminology engine described below. Currently, 15 different entity types are recognised, such as diseases, drugs, genes, etc. The second stage produces a (partial) syntactic and semantic analysis for each sentence in the text. The third step integrates these analyses into a discourse model which represents the semantic content of the text. The information is then read from the discourse model and stored in templates. These are structured objects representing domain-specific entities, their properties, and relationships between them. Templates can be imported into CLEF’s structured information repository, where they can be queried or used to generate summaries. Figure 1 illustrates how the contents of the sentence *There is collapse of the right upper lobe* from a radiology report can be represented a structured way. The Code slot of the Locus template contains an ICD-O-T code indicating a particular area of the lung (see section 3).

Recognizing and classifying relevant entities mentioned in a text is an essential aspect of IE. Highly technical domains, such as the biomedical and clinical field, have expansive and complex terminologies, requiring sophisticated term recognition capabilities. The recognition task is further complicated by the sheer number of technical terms, the constant introduction of new terms, and the absence of a 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 these issues AMBIT contains a *terminology engine*, named Termino [5]. Termino has two main components. Firstly, it contains a database into which very large numbers of single- and multi-word terms can be loaded from heterogeneous resources, including the Unified Medical Language System (UMLS), the HUGO Gene Nomenclature database, and the Gene Ontology (GO), and stored together with relevant information. Secondly, Termino flexibly allows for the compilation of subsets of terms from the database into finite state recognisers, to ensure fast and efficient identification and mark-up of terms within text. To facilitate term recognition in patient notes, Termino includes around 160,000 terms imported from UMLS, drawn from various appropriate semantic types including pharmacologic substances, anatomical structures, therapeutic procedures, diagnostic procedures, and several others.

AMBIT also contains a *query engine*, which allows users to access information through traditional free text search and search based on the structured information produced by IE, so that queries may refer to specific entities and classes of entities, and specific kinds of relations that are recognised to hold between them. The query engine is supported by an *indexing engine*, used to index text and extracted, structured information for the purposes of information retrieval.

AMBIT contains two further components: an *interface layer* providing a web or grid channel to allow user and program access to the system and a *text and annotations database* which holds the data processed by AMBIT. Further discussion of architectural issues involved in the delivery of text processing services in a distributed e-Science environment can be found in [1].

3. IE Evaluation

In this section we present evaluation results concerning AMBIT's performance on an information extraction task in the medical domain. The task under consideration involves mining radiology reports for signs indicative of lung cancer (e.g., *mass*, *density*, *collapse*), locations in the lung (e.g., *upper lobe*, *midzone*, *basal region*), and relationships between these signs and locations expressed in the reports. These two types of entity and the one kind of relationship are structured according to the templates introduced in section 2. The Code slot of a Locus template can take the following ICD-O-T values: C34.1 (upper lobe of the lung), C34.2 (middle lobe of the lung), and C34.3 (lower lobe of the lung), C34.9 (bronchus or lung, not otherwise specified).

The IE task fits within a larger information integration task, the objective of which is to determine, given a set of radiology reports for a lung cancer patient, the specific location of the primary cancer in the lung, i.e., whether it is located in the upper, middle, or lower lobe, or elsewhere (e.g. at the hilum), in the right or left lung. Determining this location requires reasoning over the collection of templates produced for a patient's document set, which potentially includes some templates that are incomplete, duplicated, incorrect or contradictory. This further step is left to the post-IE chronicle builder (see section 1), which will not be discussed here. In the general case, the exact location of a primary cancer in the lung is not recorded in the structured (i.e., non-narrative) parts of an electronic patient record nor explicitly and consistently mentioned in a patient's notes; the scenario sketched above illustrates how IE can be used to enrich patient records with new information.

In order to extract signs and locations from radiology reports we assembled a list of relevant terms to be loaded into Termino by inspecting a set of reports in consultation with a medical expert. Since the scope of the IE task is restricted, the list is rather small; around 40 new items were added to Termino.[1] The list covers various kinds of terms used in the description of signs and locations in the lung associated with lung cancer, including modifiers such as *pulmonary* and *primary*. To enable the term parser to combine these items into larger terms, we constructed a term grammar for signs and locations. Term grammars provide generative capacity, thus greatly reducing the number of terms that need to be stored in Termino. In this particular case, the rules of the term grammar also assign ICD-O-T codes to terms, which will percolate up to the Code slots of the Locus templates.

An analysis of the expressions used in patient notes reveals that there are two general patterns for describing locations in the lung. The first pattern involves nouns such as *zone* or *lobe*, which by themselves do not specify a specific region of the lung. These are combined with modifiers such as *upper*, *mid*, *basal*, etc. to arrive at a complete locational expression. This pattern is captured by the rule

location_np → latitude_adj area_noun

The *location_np* inherits its code from the *latitude_adj*. The second pattern builds on nouns which do specify a specific region of the lung, e.g., *base* and *apex*. These expressions can serve as complete descriptions of locations in the lung on their own or in combination with modifiers such as *lung* or *pulmonary*. This pattern is described by the rule

location_np → (modifier) latitude_noun

In this case, the *location_np* derives its code from the *latitude_noun*. The modifiers mentioned above can also precede terms of the category *area_noun*. Expressions following either kind of pattern can be further modified by (complex) expresssions such as *of the lung*. By a similar process of introspection we have derived a set of term grammar rules for expressions describing signs. The syntactic categories of the various terminal expressions which are combined by the term grammar rules are part of the information stored in Termino and retrieved when a term has been recognized.

Besides creating resources for lexical and terminological processing, we also added some rules to the discourse processing module to adapt it to the given IE task. One set of rules determines the laterality of a location, i.e., whether the location is in the left lung or in the

[1] Interestingly enough, the intersection of this list with the set of terms from UMLS already loaded into Termino was minimal.

Templ.	Sign		Locus			Location			
Slots		Name		Lat.	Code		Sign	Locus	All slots
Precision	64	64	80	92	79	61	53	59	69
Recall	92	92	88	68	87	92	81	90	83

Table 1: Evaluation results

right lung. These rules inspect the semantic structure of sentences and seek out occurrences of the qualifiers *left* and *right* modifying entities that have been recognised as locations. Another set of rules extracts relationships between signs and locations. This set of rules is currently very simple: whenever a sign and a location occur within the same sentence, a relationship between them is assumed.

To evaluate AMBIT's IE performance, a gold standard comprising 83 radiology reports for patients known to have lung cancer was created. Each of the documents in the gold standard was manually annotated with templates capturing the signs, locations and relationships present in the document. The gold standard documents contain a total of 82 Sign templates, 96 Locus templates, and 58 Location templates. Next, AMBIT's terminology and IE engines were run over the documents in the gold standard, after which the resulting templates were compared against the manually constructed templates of the gold standard on a per-document basis. The results are given in table 1. The table provides precision and recall scores for each type of template on the template level and the individual slot level.[2, 3] Recall is a measure of "completeness", defined as the percentage of gold standard templates or slots that has been correctly extracted by the IE engine. Precision is a measure of "cleanness", defined as the percentage of extracted templates or slots that is correct according to the gold standard. The F measure for all slots, defined as (2 × precision × recall) / (precision + recall), is 75.

The results look promising. As is typical for discourse rules based on intra-sentential co-occurrence of entities, recall scores are relatively high (most of the relationships expressed in a text are extracted), while precision scores are lower (some relationships which are not in the text are extracted spuriously). Inspection of some sample documents reveals several issues that will be addressed in future work. With the current set of discourse rules, a sentence mentioning multiple signs and locations will generally give rise to spuriously extracted relationships. Further spurious entities and relationships are extracted when they occur in negated contexts (e.g. "there is no evidence of collapse…"), as the discourse rules do not deal with negation at the moment. Issues like these will be remedied by replacing the current set of simple discourse rules by a set of rules that take full advantage of the contents of the syntactic and semantic analyses produced during IE processing. This change is expected to raise precision without significantly affecting recall. Tuning the co-reference module of the IE engine so that it will be able to recognise that, for example, in certain contexts the terms *mass* and *tumour* may refer to the same sign, will contribute to a further increase in performance.

4. Concluding Remarks

The objective of AMBIT is to provide a powerful, general-purpose system for mining information from the textual component of clinical records and the biomedical literature. Other systems with similar functionality in the medical domain include MENELAS [10], MEDLEE [2], and medSynDiKATe [4]. These systems differ from AMBIT with regard to the scope of their domains of application and the component technologies they use. A detailed

[2] The Name slot of Locus templates is not scored because all information pertinent to a Locus is contained in its Laterality and Code slots.

[3] The figures were obtained using the MUC scorer, which is a piece of software specifically designed to evaluate the performance of template-based IE systems. To score the IE output against the gold standard, the MUC scorer first aligns gold standard templates with IE output templates of the same type, on a per-document basis, taking into account the values of their respective sets of slots. Picking the optimal alignment, the scorer then computes template-level scores and slot-level scores – [5] provides further details.

comparison would go beyond the scope of this paper.

To date a first version of AMBIT has been implemented. The terminology engine is fully functional and web service-accessible. We are evaluating its performance and are investigating methods for automatically deriving term grammars from lists of terms in Termino and the use of contextual patterns to improve term recognition results. Regarding the IE engine, together with a medical expert we have specified a complete set of templates structuring the information to be extracted from our current clinical text corpus (which consists of 332,000 clinic notes for 37,000 cancer patients). We have also defined an initial set of discourse rules, covering the most important entities, events, and relationships of the clinical domain. Substantial effort is being put into extracting information about the temporal properties of events, which is essential for building chronicles. Coverage of the system is now being extended and a comprehensive evaluation is underway. Ultimately AMBIT will be successful if it gives research clinicians and biologists more effective access to the invaluable information found only in textual data.

Acknowledgements CLEF is funded by the UK Medical Research Council (MRC), research grant number MRC 60086. myGrid is funded by the UK Engineering and Physical Sciences Research Council (EPSRC), research grant number EPSRC GR/R67743.

References

1. Davis, N., Demetriou, G., Gaizauskas, R., *et al*. To appear. Web Services Architectures for Text Mining: an Exploration of the Issues via an e-Science Demonstrator. In: *International Journal of Web Services Research.*
2. Friedman, C., Hripcsak, G., DuMouchel, W., *et al*. 1995. Natural Language Processing in an Operational Clinical Information System. In: *Natural Language Engineering*, 1(1).
3. Gaizauskas, R., Demetriou, G., Artymiuk, P., *et al*. 2003. Protein Structures and Information Extraction from Biological Texts: the PASTA System. In: *Bioinformatics*, 19(1).
4. Hahn, U., Romacker, M., Schulz, S. 2002. Creating Knowledge Repositories from Biomedical Reports: the medSynDiKATe Text Mining System. In: *Proceedings Pacific Symposium on Biocomputing*.
5. Harkema, H., Gaizauskas, R., Hepple, M., *et al*. 2004. A Large Scale Terminology Resource for Biomedical Text Processing. In: *Proceedings HLT-NAACL BioLINK Workshop*.
6. Harkema, H., Setzer, A., Gaizauskas, R, *et al*. 2005. Mining and Modelling Clinical Data. In: *this volume.*
7. The Message Understanding Conference Scoring Software User's Manual. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_manual.html.
8. Rector, A., Taweel, A., Rogers, J., *et al*. 2004. Joining up Health and BioInformatics: e-Science meets e-Health. In: *Proceedings UK e-Science All Hands Meeting*.
9. Stevens, R., Tipney, H.J., Wroe, C., *et al*. 2004. Exploring Williams-Beuren Syndrome Using myGrid. In: *Bioinformatics*, 20, Suppl. 1.
10. Zweigenbaum, P. 1994. MENELAS: An Access System for Medical Records Using Natural Language. In: *Computer Methods and Programs in Biomedicine*, 45(1-2).