Journal of Biomedical Informatics 42 (2009) 950–966

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Building a semantically annotated corpus of clinical texts

Angus Roberts *, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, Andrea Setzer

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

ARTICLE INFO

Article history:
Received 6 June 2008
Available online 23 January 2009

Keywords:
Corpora
Semantic annotation
Clinical text
Natural language processing
Gold standards
Evaluation
Information extraction
Text mining
Temporal annotation
Annotation guidelines

ABSTRACT

In this paper, we describe the construction of a semantically annotated corpus of clinical texts for use in the development and evaluation of systems for automatically extracting clinically significant information from the textual component of patient records. The paper details the sampling of textual material from a collection of 20,000 cancer patient records, the development of a semantic annotation scheme, the annotation methodology, the distribution of annotations in the final corpus, and the use of the corpus for development of an adaptive information extraction system. The resulting corpus is the most richly semantically annotated resource for clinical text processing built to date, whose value has been demonstrated through its use in developing an effective information extraction system. The detailed presentation of our corpus construction and annotation methodology will be of value to others seeking to build high-quality semantically annotated corpora in biomedical domains.

1. Introduction

We describe the creation of a semantically annotated corpus of clinical texts. The documents of this corpus are drawn from the free text component of patient records, and the annotations capture clinically significant information communicated by these texts. The corpus is intended for use in developing and evaluating systems that can *automatically* extract this kind of clinically significant information from the textual component of patient records. The corpus has been created within the context of the CLinical E-Science Framework (CLEF) project [1]: a multi-site research project that has been developing the technology and techniques required for a high quality repository of electronic patient records. Such a repository must meet high standards of security and interoperability, and should enable ethical and user-friendly access to patient information, so as to facilitate both clinical care and biomedical research. CLEF has chosen to work in the area of cancer informatics, as one of the project partners—the Royal Marsden Hospital (RMH)—is a large specialist oncology centre.

Although much of the patient information needed to populate such a repository exists as structured data, e.g. database records of drug prescriptions and clinic appointments, free text material still forms an important component of electronic patient records, and contains information that is potentially significant both for day-to-day care and clinical research. For example, letters written from the secondary to the primary care physician (e.g. from specialist consultant to patient GP) form a major component of any UK medical record, and free text plays a key role in the reporting of imaging and pathology findings. Clinical narratives may record, for instance, why drugs were given or discontinued, the results of physical examination, and issues considered important when discussing patient care but which are not coded for audit. Such information, when combined with that from the structured record, and suitably presented, could contribute to individual patient care, e.g. providing a consultant with a concise summary of their patient's clinical history, or access to concise histories for patients with similar conditions elsewhere. Aggregation of information across all the records in a large repository could bring benefits for clinical research. For example, being able to get answers to questions such as "*How many patients with stage 2 adenocarcinoma who were treated with tamoxifen were symptom-free after 5 years?*" could assist a researcher in formulating hypotheses that could be later explored in clinical trials.

The need to make the information that exists in clinical texts available for integration with the structured record, for subsequent use in clinical care and research, has been addressed within CLEF through the use of *information extraction* (IE) technology [2,3]. Although some IE research has focused on unsupervised methods of developing systems, as in the earlier work of Riloff [4], most practical modern IE work requires data that have been manually annotated with the events, entities and relationships that are considered to express key content for the given domain. These data serve three purposes. First, the analysis of data that is required to

* Corresponding author. Fax: +44 114 222 1810.
E-mail address: a.roberts@dcs.shef.ac.uk (A. Roberts).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.12.013

create the annotation scheme serves to focus and clarify the information requirements of the task and domain. Second, the annotated data provide a *gold standard* against which to assess the performance of systems designed to automatically identify this information in texts. Third, it serves as a resource for system development: extraction rules may be created either automatically or by hand, and statistical models of the text may be built by machine learning algorithms.

This paper reports on the work done within CLEF to create an annotated corpus, to aid the development and evaluation of the CLEF IE system. To the best of our knowledge, no one else has explored the problem of producing a corpus annotated for clinical IE to the depth and extent reported here, and the resulting corpus is the most richly semantically annotated resource for clinical text processing built to date. Our annotation exercise draws its texts from a large background corpus of clinical narratives, covers multiple text types, and involves over 20 annotators. Results are encouraging, and suggest that a rich corpus to support IE in the medical domain can be created.

We reported the early development of the CLEF corpus in [5]. The current paper elaborates quantitative results from this development process, giving a much greater level of detail. Quantitative results have also previously been given, for the partially complete corpus, in [6]. The results in the current paper are final, reflecting the finished corpus. In addition, the current paper provides results and descriptions not previously published, including: annotation with UMLS CUIs; annotation of temporal expressions; the summary results of an annotator difference analysis; a discussion of time taken to annotate; detailed descriptions of the annotation guidelines, their development and application; and greater detail of our annotation methodology. We also summarise work on the corpus in use, to train and evaluate a working IE system. We believe that this detailed account of our methodology, corpus, and its use will be of benefit to other groups contemplating similar exercises.

The paper is organised as follows: in the next section, we summarise previous efforts to create annotated corpora in biomedical domains. Section 3 describes how material was selected for inclusion in our corpus, and then in Section 4, we describe the semantic annotation schema, the annotation methodology, the development of the annotation guidelines, as well as the measures for assessing the consistency of human annotations. Section 5 presents an analysis of aspects of the annotation process and Section 6 presents inter annotator agreement scores for the finished corpus, and figures on the distribution of entity and relation types by document type across the corpus. The next section describes work carried out subsequent to the initial corpus construction work, to add a layer of temporal annotation. Finally, in Section 8, we mention on-going use of the corpus for training and evaluation of our supervised machine learning IE system.

2. Annotated corpora for biomedical research

Annotated corpora, or text collections, are now recognised as resources of central importance in biomedical language processing research. They may be taxonomized in various ways. For example, they can be grouped by domain (e.g. protein–protein interactions and oncology), document type or genre (e.g. research article, clinical narrative, and radiology report), type of annotation (e.g. semantic—entities, relations and/or syntactic—part-of-speech, parse structure), intended language processing application (e.g. information extraction, text classification), intended mode of use (e.g. for training adaptive systems, for specific system evaluation, for community wide shared task evaluation), or availability (e.g. publicly available or not publicly available). It is not our intention to attempt a complete characterisation and review of all annotated corpus resources that have been used in biomedical language processing research. Instead we focus on a few that enable us to show where the CLEF corpus fits in the context of prior research and what novel contribution it makes.

The CLEF corpus may be characterised as a semantically annotated corpus of clinical documents of mixed type (clinic letters, radiology, and histopathology reports) which is designed to support both automated training and evaluation of information extraction systems. While it is not publicly available at time of writing we are working towards its release (see below) and reusability has been an important consideration informing its design.

There are now a significant number of publicly available semantically annotated corpora designed to support information extraction research comprising texts drawn from the biomedical research literature. For example, the GENIA corpus is a collection of ~200 MEDLINE abstracts in the area of molecular biology that has had mentions of specific biological entities and events annotated within it [7,8]. The PennBioIE corpus [9] consists of ~2300 MEDLINE abstracts, in the domains of molecular genetics of oncology and inhibition of enzymes of the CYP450 class and is annotated for biomedical entity types (it is also annotated syntactically for parts-of-speech amd some portion of it has been annotated for Penn Treebank style syntactic structure). The Yapex corpus contains 200 MEDLINE abstracts annotated for protein names [10]. The BioText project has made several semantically annotated corpora available, including one for disease–treatment relation classification consisting of ~3500 sentences drawn from MEDLINE abstracts labelled for DISEASE and TREATMENT and seven types of relation holding between them [11], and one for protein–protein interaction classification consisting of ~800 sentences drawn from full-text journal papers, where each sentence contains mentions of an interacting protein pair [12]. The ITI TXM corpus [13] has annotated tissue expressions in 238 full-text documents drawn from PubMed and protein–protein interactions in 217 documents obtained from PubMed Central and PubMed.

While these corpora have been developed in the contexts of specific research projects they have been developed with a view to reusability and have been released to the wider research community. Other semantically annotated corpora drawn from the biomedical research literature have been developed specifically for the purpose of shared task evaluations of information extraction systems. These evaluations include the Biocreative challenge, which utilised the GENETAG corpus containing 20,000 sentences with gene/protein names annotated [14]), the LLL05 challenge task, which supplied training and test data for the task of identifying protein/gene interactions in sentences from MEDLINE abstracts [15], and the TREC Genomics Track, which, while focussed on information retrieval rather than information extraction, did yield some datasets which could be viewed as semantically annotated, e.g. the TREC 2007 task for which human relevance judgements include lists of domain-specific entities associated with relevant passages [16].

The corpora mentioned so far consist of texts drawn from the research literature. Corpora consisting of clinical texts, e.g. clinic letters, radiology, and histopatholgy reports, are much rarer—getting access to clinical text for research purposes is difficult due to issues of patient confidentiality and getting permission to release them to the wider research community is even more challenging. To our knowledge the only annotated corpora intended to support research in clinical information retrieval and extraction that have been released to the wider research community are those developed in the context of several recent shared task challenges. For example, the corpus prepared and released for the Computational Medicine Challenge [17] consists of 1954 (978 training and 976 test) radiology reports annotated with ICD-9-CM codes, where the challenge is to automatically code the unseen test data. The ImageCLEFmed 2005 and 2006 image test collections consist of ~50,000 images with associated textual annotations (case descrip-