Conference Report Editor: Samantha Barton
s.barton@elsevier.com

conference report

Spot the dog, see Spot jump, watch spot denature

David Bousfield, david@ganesha-associates.com

First International Symposium in Semantic Mining in Biomedicine
European Bioinformatics Institute (EBI), Hinxton, UK
10–13 April 2005

The *First International Symposium in Semantic Mining in Biomedicine*, which was held on 10–13 April 2005 at the European Bioinformatics Institute (EBI; Hinxton, UK), was designed to bring together scientists and information technology (IT) experts in the field of ontologies, text mining and database design and to discuss the latest developments in the integration of terminological resources into retrieval and extraction methods to further advance IT solutions in the biomedical domain.

Setting the scene

Udo Hahn (Jena University, Jena, Germany), one of the programme chairs, and Stefan Schulz (University Hospital, Freiburg, Germany) and Dietrich Rebholz-Schuhmann (EMBL-EBI), the organizers of the meeting, are members of the Network of Excellence (NoE) for 'Semantic Mining' (Semantic Interoperability and Data Mining in Biomedicine) – an NoE funded by the European Union as part of the Framework 6 program Integrating and Strengthening the European Research Area (www.cordis.lu/indicators/projects_era.htm). This NoE gathers together clinicians, molecular biologists, linguists, computer scientists and engineers to develop new methods in ontology research, natural language, information retrieval and text-mining technologies, and accelerates the creation of essential resources, such as thesauri, lexicons and ontologies, running the gamut from biology and medicine to health care statistics (www.semanticmining.org).

As keynote speaker Carol Friedman (Columbia University, USA) remarked in the opening plenary session, the future of medicine depends on our ability to create technologies that enable the automatic integration, organization and distillation of literally astronomical volumes of data from, for example, clinical records, primary literature and structured databanks. The objective is to provide better clinical decision-support, quality assurance and patient management, as well as insight into basic disease mechanisms. To achieve this goal, it will be necessary to develop natural language processing (NLP) software that integrates biological and clinical domain knowledge (e.g. taxonomies, thesauri and ontologies). The vision is that this software will be able to identify biological entities and relationships within unstructured text and build this knowledge into sophisticated tools capable of, for example, displaying all known protein–protein interactions, predicting drug–disease contra-indications or identifying the beginning of a new disease outbreak.

Friedman went on to describe some of the progress she and her colleagues have made in developing robust practical systems for everyday use. The Medical Language Extraction and Encoding System (MedLEE; http://lucid.cpmc.columbia.edu/medlee) was developed in the 1990s at The New York Presbyterian Hospital and is designed to convert the free text reports generated by the Radiology Department into a machine readable format that can then be used to alert physicians automatically to the need for follow-up, existence of co-morbidities, and so on. MedLEE is a working proof-of-concept, and in that respect is unique. However, the success of MedLEE has stimulated the builders of other systems to incorporate MedLEE-like features. A second system, GENIES, has been developed to acquire knowledge automatically by extracting and structuring biomolecular relations from the literature. Friedman's next goal is to construct tools that bring together knowledge from the clinical and biomedical domains – BioMedLEE.

Creating new tools

Gert-Jan van Ommen (Leiden University, The Netherlands) and Steve Foord (GlaxoSmithKline, Stevenage, UK) brought their perspectives from biomedical research and the pharmaceutical industry and extended the challenge to include the mining and integration of experimental and clinical data. van Ommen explained how distinctive data patterns in differential expression studies (gene, protein and metabolite) could be used to dissect out clinically distinct sub-populations. Studies highlighted by van Ommen as benefiting significantly from the use of pathway tools to speed comparisons between the experimental data and the consensus views

1359-6446/05/$ – see front matter PII: S1359-6446(05)03503-8

conference report

buried in the primary literature were the apolipoprotein E3 Leiden mouse model for atherosclerosis and research conducted at the Erasmus University on the clinical identification of acute myeloid leukaemia subtypes.

By contrast, Foord commented that rigid compartmentalising of data, for example, having purely text-based or ontology-based biomedical knowledge integration, could compromise our abilities to infer unanticipated biological relationships from experimental data.

Jun'ichi Tsujii (University of Tokyo, Japan) presented a tour of tools and resources being developed by his group and collaborators in Singapore, France and elsewhere. These comprised a realm of linguistics- or statistics-based software components for the identification and extraction of, for example, named entities, noun phrases and sentence boundaries, plus additional tools for parsing complete sentences to extract these entities. A typical application for these tools is the automatic extraction of information on protein–protein interactions or the function of a gene or protein.

These tools can only operate efficiently if they also have access to domain knowledge represented in controlled vocabularies, thesauri and ontologies. Such resources keep track of known synonyms, homonyms and abbreviations such that the meaning of words can be identified unambiguously. Ontologies, such as the gene ontologies [GOs (www.geneontology.org)] and KEGG (www.genome.jp/kegg), provide additional conceptual information that captures knowledge about an entity, such as its molecular mechanism of action, role in a biological process or cellular location. This information can greatly improve the correct identification of relationships between entities in text. With the current generation of tools and resources available sentences such as 'Few known genes [Interleukin-2 (IL-2), members of the IL-8 family and interferon gamma] are induced in T-cells only although the combined effect of phorbol myristic acetate (PMA) and a Ca^{2+}-ionophore, and the expression of only these genes can be fully expressed by cyclosporine a (CyA)' can be tackled, separated into pieces and then reassembled into a set of machine-readable biological relationships.

To achieve this level of sophistication requires a painstaking amount of manual work creating a fully annotated reference corpus that can then be used to train the software, a process that can also need a significant amount of human supervision. Tsujii's group have created the GENIA corpus, which comprises over 4000 Medline abstracts selected on the basis of the medical subject headings 'human', 'blood cells' and 'transcription factors'. Each abstract has been annotated so that the results extracted from tokenization, part-of-speech tagging, named entity recognition, sentence analysis and co-reference resolution (processes that inspect punctuation marks to identify where sentences begin and end, identify where words begin and end, which is particularly important for abbreviations and technical terms, and recognize whether words are nouns, verbs and conjunctions) are all captured as a reference database. GENIA is used to train and test NLP tools, but it does have its limitations – it represents only a fraction of the 15 million abstracts available on Medline and is not optimized for the broader scope of discourse styles found in full text papers. The growth of open access sources of content, such as PubMed Central, should help to alleviate this problem.

So, the performance of these systems is still highly domain and application dependent. Consequently, there is much interest in developing practical systems that can generalize, albeit to a limited extent. Jasmin Saric (EML, Heidelberg) and his colleagues described their research on the large-scale extraction of gene–protein relationships in a series of model organisms. Building on previous research on yeast, they showed that from a dataset of almost one million PubMed abstracts they could extract information about regulatory networks and phosphorylation events from three additional species (*Bacillus subtilis*, *Escherichia coli* and mouse), with precision ratings in the range of 83–90% and 86–95%, respectively. Next, Saric and colleagues plan to tackle glycosylation and methylation events.

Extending ontologies

Anita Burgun (University of Rennes, France) described research using the Chemical Entities of Biological Interest (ChEBI; www.ebi.ac.uk/chebi) ontology to extend the functionality of the GO of molecular functions, biological processes and cellular components by identifying new dependencies within GO. Such tools could have a particularly useful role in annotating the new PubChem database (http://pubchem.ncbi.nlm.nih.gov) that is being developed as part of the National Institutes of Health Roadmap initiative. Jong C. Park (KAIST, Korea) described the automatic extension of GO with newly published information about gene products, and Ted Sandler (University of Pennsylvania, USA) showed how automated techniques could be used to create and extend term lists used for entity extraction.

Joachim Wermter (Jena University) described how the performance of laboratory prototype and commercial (TEMIS) natural language tools could be improved over and above that achieved from training with Penn Treebank corpus (which is based on text from the Wall Street Journal) by addition of a few biomedical domain-specific adaptations provided by GENIA. Jian Su (National University of Singapore) showed how statistical techniques (Maximum Entropy Model) could be used to combine lexical, syntactic and semantic features to maximize the extraction of protein–protein interactions without the need for state-of-the art co-occurrence- or rule-based approaches.

On a less theoretical note, Jörg Hakenberg (Humboldt University, Berlin) came to the fascinating conclusion that the performance of their text classification system for hereditary diseases could be improved simply by varying the degree of emphasis placed on the contributions made by different sections from the full-text papers – abstract, introduction and materials and methods sections proved most useful. Christine Chichester (GeneBio, Switzerland) and Agnes Sandor (Xerox, France) described how linguistic tools could be used to identify articles that describe paradigm shifts in the field of neurodegenerative diseases.

Future challenges

The meeting reported some sound developments in the fields of entity and relationship extraction – research that will provide great assistance to the systems

conference report

biologist building pathway maps and so forth – but it also clearly delineated the huge gap between our current capabilities and the goals and needs set out by the plenary talks given by Friedman, van Ommen and Foord.

Near the conclusion of the meeting, Hahn listed and summarized seven of the most proximal challenges for the field:

(i) Taming lexical 'hyper-ambiguity' – the propensity of biologists for giving different names and acronyms to the same entity or the same name to many entities.

(ii) The development of a bio-version of the unified medical language system (UMLS) developed by the National Library of Medicine and used to facilitate the retrieval and integration of machine readable biomedical information from, for example, clinical records, literature and organizational databanks. Although GO is a start, conceptual coverage of the whole of the biological-clinical domain is required.

(iii) The problems associated with relying on Medline as a source of test data. Better access to large quantities of full text is required, giving access to more knowledge, as well as a greater range of discourse styles. Commercial publishers could play an important role in this area.

(iv) The creation of a corpus that would allow access to significant quantities of full text would enable the development of a BIO Treebank with GENIA at its nucleus, analogous to the Penn Treebank.

(v) The field already hosts several consensus conferences where the state of the field is chronicled and competitions are held based around carefully devised benchmark problems – for example, BioCreative 2004 (www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html) and KDD Cup 2002 (www.biostat.wisc.edu/~craven/kddcup) – but more needs to be done to bring a more formal structuring to the directional priorities of research, pooling of resources and the involvement of other key players outside of academia, such as biotechnology companies.

(vi) To be useful, semantic mining tools must be capable of joining a variety of data types ranging from structured (genome) databases, quantitative (proprietary) data collections and unstructured textual data (technical reports and patents).

(vii) Entity and relationship mining tools are already in widespread use by database curators, but how do we best move semantic mining technologies from this experimental stage onto the desktop of the biomedical researcher and clinician?

Clearly there are plenty of topics for the next symposium in the series, but this author shares Hahn's call for much clearer links to be made between these research activities and the practical needs of clinicians and researchers at the bench.

David Bousfield
Ganesha Associates,
73 de Freville Avenue,
Cambridge,
UK, CB4 1HP
e-mail: david@ganesha-associates.com