Camtology: Intelligent Information Access for Science

Ted Briscoe[1,2], **Karl Harrison**[5], **Andrew Naish-Guzman**[4], **Andy Parker**[1], **Advaith Siddharthan**[3], **David Sinclair**[4], **Mark Slater**[5] **and Rebecca Watson**[2]

[1]University of Cambridge
ejb1@cl.cam.ac.uk,
parker@hep.phy.cam.ac.uk,
[2]iLexIR Ltd
rfw@ilexir.co.uk
[3]University of Aberdeen
advaith@abdn.ac.uk
[4]Camtology Ltd
david.sinclair@imense.co.uk,
a.naish@gmail.com
[5]University of Birmingham
kh@hep.ph.bham.ac.uk,
mws@hep.ph.bham.ac.uk

Abstract

We describe a novel semantic search engine for scientific literature. The Camtology system allows for sentence-level searches of PDF files and combines text and image searches, thus facilitating the retrieval of information present in tables and figures. It allows the user to generate complex queries for search terms that are related through particular grammatical/semantic relations in an intuitive manner. The system uses Grid processing to parallelise the analysis of large numbers of papers.

1 Introduction

Scientific, technological, engineering and medical (STEM) research is entering the so-called 4th Paradigm of "data-intensive scientific discovery", in which advanced data mining and pattern discovery techniques need to be applied to vast datasets in order to drive further discoveries. A key component of this process is efficient search and exploitation of the huge repository of information that only exists in textual or visual form within the "bibliome", which itself continues to grow exponentially.

Today's computationally driven research methods have outgrown traditional methods of searching for scientific data, creating a widespread and unfulfilled need for advanced search and information extraction. Camtology combines text and image processing to create a unique solution to this problem.

2 Status

Camtology has developed a search and information extraction system which is currently undergoing usability testing with the curation team for FlyBase[1], a $1m/year NIH-funded curated database covering the functional genomics of the fruit fly. To provide a scalable solution capable of analysing the entire STEM bibliome of over 20m electronic journal and conference papers, we have developed a robust system that can be used with a grid of computers running distributed job management software.

This system has been deployed and tested using a subset of the resources provided by the UK Grid for Particle Physics (Britton et al., 2009), part of the worldwide Grid of around 200000 CPU cores assembled to allow analysis of the petabyte-scale data volumes to be recorded each year by experiments at the Large Hadron Collider in Geneva. Processing of the FlyBase archive of around 15000 papers required about 8000 hours of CPU time, and has been successfully completed in about 3 days, with up to a few hundred jobs run in parallel. A distributed spider for collecting open-source PDF documents has also been developed. This has been run concurrently on over 2000 cores cores, and has been used to retrieve over 350000 subject-specific papers, but these are not considered in the present demo.

3 Functionality

Camtology's search and extraction engine is the first to integrate a full structural analysis of a scientific paper in PDF format (identifying headings, sections, captions and associated figures, citations and references) with a sentence-by-sentence grammatical analysis of the text and direct visual search over figures. Combining these capabilities allows us to transform paper search from keyword based paper retrieval, where the end result is a set of putatively relevant PDF files which need to be read, to information extraction based on the ability to interactively specify a rich variety of linguistic patterns which return sentences in specific document locales, and which combine text with image-based constraints; for instance:

> "all sentences in figure captions which contain any gene name as the theme of the action 'express' where the figure is a picture of an eye"

[1]http://flybase.org/

Proceedings of the NAACL HLT 2010: Demonstration Session, pages 1–4,
Los Angeles, California, June 2010.

Camtology allows the user to build up such complex queries quickly though an intuitive process of query refinement.

Figures often convey information crucial to the understanding of the content of a paper and are typically not available to search. Camtology's search engine integrates text search to the figure and caption level with the ability to re-rank search returns on the basis of visual similarity to a chosen archetype (ambiguities in textual relevance are often resolved by visual appearance). Figure 1 provides a compact overview of the search functionality supported by our current demonstrator. Interactively, constructing and running such complex queries takes a few seconds in our intuitive user interface, and allows the user to quickly browse and then aggregate information across the entire collection of papers indexed by the system. For instance, saving the search result from the example above would yield a computer-readable list of gene names involved in eye development (in fruit flies in our demonstrator) in a second or so. With existing web portals and keyword based selection of PDF files (for example, Google Scholar, ScienceDirect, DeepDyve or PubGet), a query like this would typically take many hours to open and read each one, using cut and paste to extract gene names (and excludes the possibility of ordering results on a visual basis). The only other alternative would require expensive bespoke adaptation of a text mining system by IT professionals using licensed software (such as Ariadne Genomics, Temis or Linguamatics). This option is only available to a tiny minority of researchers working for large well-funded corporations.

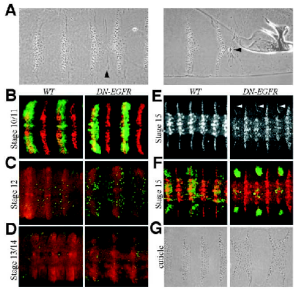
4 Summary of Technology

4.1 PDF to SciXML

The PDF format represents a document in a manner designed to facilitate printing. In short, it provides information on font and position for textual and graphical units. To enable information retrieval and extraction, we need to convert this typographic representation into a logical one that reflects the structure of scientific documents. We use an XML schema called SciXML (first introduced in Teufel et al. (1999)) that we extend to include images. We linearise the textual elements in the PDF, representing these as elements in XML and classify these divisions as {Title|Author|Affiliation|Abstract|Footnote|Caption| Heading|Citation| References|Text} in a constraint satisfaction framework.

In addition, we identify all graphics in the PDF, including lines and images. We then identify tables by looking for specific patterns of text and lines. A bounding box is identified for a table and an image is generated that overlays the text on the lines. Similarly we overlay text onto images that have been identified and identify bounding boxes for figures. This representation allows us to retrieve figures and tables that consist of text and graphics. Once bounding boxes for tables or figures have been identified, we identify a one-to-one association between captions and boxes that minimises the total distance between captions and their associated figures or tables. The image is then referenced from the caption using a "SRC" attribute; for example, in (abbreviated for space constraints):

<CAPTION SRC= "FBrf0174566_fig_6_o.png"> Fig. 6. Phenotypic analysis of denticle belt fusions during embryogenesis. (A) The denticle belt fusion phenotype resulted in folds around the surrounding fused... ...(G) ...the only cuticle phenotype of the DN-EGFR-expressing embryos was strong denticle belt fusions in alternating parasegments (paired domains).</CAPTION>



Note how informative the caption is, and the value of being able to search this caption in conjunction with the corresponding image (also shown above).

4.2 Natural Language Processing

Every sentence, including those in abstracts, titles and captions, is run through our named-entity recogniser and syntactic parser. The output of these systems is then indexed, enabling semantic search.

Named Entity Recognition

NER in the biomedical domain was implemented as described in Vlachos (2007). Gene Mention tagging was performed using Conditional Random Fields and syntactic parsing, using features derived from grammatical relations to augment the tagging. We also use a probabilistic model for resolution of non-pronominal anaphora in biomedical texts. The model focuses on biomedical entities and seeks to find the antecedents of anaphora, both coreferent and associative ones, and also to identify discourse-new expressions (Gasperin and Briscoe, 2008).

Parsing

The RASP toolkit (Briscoe et al., 2006) is used for sentence boundary detection, tokenisation, PoS tagging and finding grammatical relations (GR) between words in the text. GRs are triplets consisting of a relation-type and two arguments and also encode morphology, word position and part-of-speech; for example, parsing "John likes Mary." gives us a subject relation and a direct object relation:

(|ncsubj| |like+s:2_VVZ| |John:1_NP1|)
(|dobj| |like+s:2_VVZ| |Mary:3_NP1|)

Representing a parse as a set of flat triplets allows us to index on grammatical relations, thus enabling complex relational queries.

4.3 Image Processing

We build a low-dimensional feature vector to summarise the content of each extracted image. Colour and intensity histograms are encoded in a short bit string which describes the image globally; this is concatenated with a description of the image derived from a wavelet decomposition (Jacobs et al., 1995) that captures finer-scale edge information. Efficient similar image search is achieved by projecting these feature vectors onto a small number of randomly-generated hyperplanes and using the signs of the projections as a key for locality-sensitive hashing (Gionis et al., 1999).

4.4 Indexing and Search

We use Lucene (Goetz, 2002) for indexing and retrieving sentences and images. Lucene is an open source indexing and information retrieval library that has been shown to scale up efficiently and handle large numbers of queries. We index using fields derived from word-lemmas, grammatical relations and named entities. At the same time, these complex representations are hidden from the user, who, as a first step, performs a simple keyword search; for example "express Vnd". This returns all sentences that contain the words "express" and "Vnd" (search is on lemmatised words, so morphological variants of "express" will be retrieved). Different colours represent different types of biological entities and processes (green for a gene), and blue shows the entered search terms in the result sentences. An example sentence retrieved for the above query follows:

> It is possible that like ac , sc and l'sc , vnd is expressed initially in cell clusters and then restricted to single cells .

Next, the user can select specific words in the returned sentences to indirectly specify a relation. Clicking on a word will select it, indicated by underlining of the word. In the example above, the words "vnd" and "expressed" have been selected by the user. This creates a new query that returns sentences where "vnd" is the subject of "express" and the clause is in passive voice. This retrieval is based on a sophisticated grammatical analysis of the text, and can retrieve sentences where the words in the relation are far apart. An example of a sentence retrieved for the refined query is shown below:

> First , vnd might be spatially regulated in a manner similar to ac and sc and selectively expressed in these clusters .

Camtology offers two other functionalities. The user can browse the MeSH (Medical Subject Headings) ontology and retrieve papers relevant to a MeSH term. Also, for both search and MeSH browsing, retrieved papers are plotted on a world map; this is done by converting the affiliations of the authors into geospatial coordinates. The user can then directly access papers from a particular site.

5 Script Outline

I Quick overview of existing means of searching science (PubMed, FlyBase, Google Scholar).

II Walk through the functionality of Camtology (these are numbered in Figure 1:

- (1) Initial query through textual search box; (2) Retrieval of relevant sentences; (3) Query refinement by clicking on words; (4) Using implicit grammatical relations for new search;
- Alternative to search: (5) Browse MeSH Ontology to retrieve papers with MeSH terms.
- (6) Specifically searching for tables/figures
- (7) Viewing the affiliation of the authors of retrieved papers on a world map.
- (8) Image search using similarity of image.

6 Acknowledgements

This work was supported in part by a STFC miniPIPSS grant to the University of Cambridge and iLexIR Ltd.

References

T. Briscoe, J. Carroll, and R. Watson. 2006. The second release of the RASP system. In *Proc. ACL 2006*.

D. Britton, AJ Cass, PEL Clarke, et al. 2009. GridPP: the UK grid for particle physics. *Philosophical Transactions A*, 367(1897):2447.

C. Gasperin and T. Briscoe. 2008. Statistical anaphora resolution in biomedical texts. In *Proc. COLING'08*.

A. Gionis, P. Indyk, and R. Motwani. 1999. Similarity search in high dimensions via hashing. In *Proc. 25th ACM Internat. Conf. on Very Large Data Bases*.

B. Goetz. 2002. The Lucene search engine: Powerful, flexible, and free. *Javaworld http://www. javaworld. com/javaworld/jw-09-2000/jw-0915-lucene. html*.

C.E. Jacobs, A. Finkelstein, and D.H. Salesin. 1995. Fast multiresolution image querying. In *Proc. 22nd ACM annual conference on Computer graphics and interactive techniques*.

S. Teufel, J. Carletta, and M. Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proc. EACL'99*.

A. Vlachos. 2007. Tackling the BioCreative2 gene mention task with CRFs and syntactic parsing. In *Proc. 2nd BioCreative Challenge Evaluation Workshop*.

Figure 1: Screenshots showing functionality of the Camtology search engine.