WeScience$_0$

Super-Charging the National Archives of Scholarly Literature

Stephan Oepen (IFI), Jan Tore Lønning, (IFI), Morten Erlandsen (USIT)

1 General Information

The following is a project outline for an initial feasibility study into the use of *language technology* for improved use of open access archives. The goals of this initiative are two-fold: viz. (a) to improve accessibility and utilization of the NORA collection, as the main on-line entry point to Norwegian research literature; and (b) to prepare the technological infrastructure and expertise for a follow-up project—applied to larger collections of scholarly literature and providing more advanced services.

This is a joint initiative of the Department of Informatics (IFI) at UiO and the UiO Center for Information Technology (USIT). We propose to initiate the implementation of a next-generation search infrastructure for scholarly literature, pairing bibliographic database and search expertise with competency in language technology, to be made available as an overlay to the existing NORA interfaces.

The project parallels on-going work in the international Association for Computational Linguistics (ACL) and its Digital Anthology (Bird et al., 2008; Schäfer, Uszkoreit, Federmann, Marek, & Zhang, 2008) and will exchange tools and data with the ACL. Another closely related role model is the national Intute Repository Search in the UK (see `http://www.intute.ac.uk/irs/`).

2 The Research Infrastructure

Scholarly digitial archives already take a central role in both the dissemination and (impact) assessment of scientific results, and ongoing strong trends towards digital publication will only increase academic dependence on immediate and effective access to digital content—across all scientific disciplines. The national Norwegian Open Research Archives (NORA) today provide a uniform 'umbrella' interface to various scholarly archives, large and small, at Norwegian institutions of higher education and research. At present, NORA is an on-line catalogue of about 22,000 scientific publications[1]—most with common metadata (author(s), title, type of publication, et al.), a brief abstract, classification by discipline, and an electronic copy of the document itself (mostly in the form of PDF files, for about two thirds of the collection).

The existing web-based search interface for NORA allows browsing the collection by metadata properties or keyword search on the abstracts, but *not* on the documents proper. Besides this lack of textual search on the actual documents (which could in principle be accomplished by having the NORA site indexed by a service like Google Scholar), there is no support for *content-based* access—for example search based on bibliographic references[2] or disambiguation of terms (e.g. the distinct usages of *motion* in law vs. in mechanics), let alone the recognition of abstract semantic properties and relations instantiated in a document.

To motivate such extensions to NORA, consider the following use scenarios. We believe that textual search on the full documents in NORA would already contribute to making the archive more attractive; when searching for documents that mention a specific approach, for example, there is no guarantee that search on document abstracts only will yield satisfatory retrieval results. The ability

[1] A majority of entries today are graduate student theses, but the counts for doctoral theses and self-archived scholarly articles are increasing steadily.

[2] We use the term *reference* to mean full bibliographic information, typically given at the end of an article, and the term *citation* to refer to an inline 'pointer' to a complete reference, as part of the running text.

to parse and track bibliographic references in scholarly publications paves the way to novel, more advanced search and quantitative analysis: documents can be retrieved on the basis of what they cite, say, or according to how often they are cited. Based on references among documents, the whole collection can be treated as a linked graph, which allows for alternate ways of browsing, visualization, or link analysis. A more advanced use scenario might include analyzing the argumentative function of each citation (Siddharthan & Teufel, 2007), such that users could search for documents, for example, which contradict or criticize an earlier publication. Finally, as a more long-term vision, search beyond the limitations of string-level keywords should take into account morphological variation (*parsing*, *parse*, and *parsed*, for example, all are morphological variants); so-called word sense disambiguation (the different, domain-specific meanings of *motion* in our earlier example); taxonomic relations among words (a *motion* in law is a kind of *proposal*; in mechanics, *movement* and *motility* are synonymous with *motion*); and ultimately semantic relations among concepts, for example making it possible to search for documents that specifically discuss *motions presented to the Norwegian parliament*.

WeScience$_0$ proposes to prepare the foundations for such extensions to NORA (and similar archives) through the application of basic language technology. This pre-project will conduct a technological feasibility study. Starting from the current collection of PDF files in NORA and existing open-source tools (see below), the project will establish the following fully-automated document processing infrastructure: (a) text extraction from PDF; (b) text correction;[3] (c) text segmentation (into paragraphs and sentences); (d) reference parsing, (e) citation identification and matching, and (f) full-text search.

On the one hand, the results of processes (a) through (f) will be made available as an experimental extension to the current NORA search interface, hosted at UiO. The extra functionality will add to the recognition and use of the existing collection, and may over time further the adaptation of scholarly digitial libraries as a community resource. On the other hand, the pilot project will build up local expertise in combining and using existing tools on NORA documents, and it will help 'map out' typical obstacles and challenges, for example correlations of text extraction errors to the original process used in PDF creation. Finally, both the resulting collection of pure text files and structured bibliographic information, as well as the extraction and post-processing pipeline itself provide a re-usable platform for subsequent R&D on next-generation digital archives.

To a certain degree, steps (a) through (f) above require the replication of functionality that services like Google Scholar and CiteSeer use internally. However, for the adaptation to conventions used in Norwegian academic publishing, for the integration with existing national services, as well as for the development of an enabling infrastructure for follow-up eScience R&D in Norway, it is important that both technology and expertise are freely available. Although the project will survey and test a variety of text extraction tools, we plan to establish the complete document processing and retrieval infrastructure as an open-source repository. Re-use of existing tools (specifically ones in use already in the ACL community) and speedy release of extensions and new technology developed by the project will aid international recognition and exchange with the larger scientific community.

3 Technology & Work Packages

Following is a brief discussion of existing technology that we expect to put to use, combined with an indication of possible 'work packages' for the proposed pilot study.

[3] Even for PDF files that were 'born digitally', custom font encodings or glyphs, column and page breaks, displayed elements (equations, tables, figures), and similar idiosyncrasies present challenges to reliable text extraction. Also, a process of 'un-hyphenation' is called for, to eliminate spurious (but obviously not all) intra-word hyphens occuring at line breaks.

Text Extraction Many packages exist for text extraction from PDF, some based on OCR-like techniques (primarily for scanned documents), others working as limited PDF interpreters, reading out a pure text stream from 'digitally born' documents. One of the more widely used packages appears to be Apache PDFBox (`http://incubator.apache.org/pdfbox/`), which we will evaluate as our baseline—parallel to much ongoing work in the international ACL community. For a smaller sample of NORA documents, we also plan to contrastively look at tools like PDFtoHTML (`http://pdftohtml.sourceforge.net/`), A-PDF Text Extractor (`http://a-pdf.com/text/index.htm`), and Adobe Acrobat.

Text Correction and Segmentation While quite a number of open-source text segmentation tools exists (where we have good experience with a package called `tokenizer`, see `http://www.cis.uni-muenchen.de/~wastl/misc/`, we are not aware of existing technology to deal with idiosyncrasies that are typical of text extracted from PDF (or, more generally, OCR), as for example page breaks—where page headers and footers may seem to occur as part of the main text stream—or hyphen elimination at line breaks. One primary goal of the pilot project is to obtain a better understanding of the nature and distribution of such idiosyncrasies; in this initial phase, we expect to devise a collection of relatively simple text correction scripts (combining pattern matching and the use of spelling dictionaries) and determine their efficacy.

Reference Tracking Tracking of bibliographic references encompasses two sub-tasks, viz. citation identification in running text and reference parsing, i.e. breaking a bibliographic reference into its component pieces (author(s), title, year of publication, publisher, and others). The former task we will address as a classification problem (combining off-the-shelf supervised classifiers and a semi-automated procedure for the creation of training data), while for the latter task we plan to adapt the open-source ParsCit software package (`http://aye.comp.nus.edu.sg/parsCit/`). Unlike the tasks discussed so far, we anticipate a certain degree of variation in the formatting of bibliographic references according to the choice of language, scientific discipline, and possibly also type of publication (say an MSc thesis vs. an article published in a peer-reviewed and copy-edited journal). This initial feasibility study will help determine the degree of such variation and need for customization or parametrization of ParsCit.

Full-Text Search Finally, for the full-text search component of the proposed pre-project, we plan to build on the mature Apache Lucene infrastructure (`http://lucene.apache.org/java/docs/`), which appears to provide a de-facto open-source standard for full-text search and document retrieval tasks. One of the doctoral students associated to our research group, Gisle Ytrestøl, has detailed knowledge of Lucene and will assist in this work package. We expect to arrive at a functional (if experimental) full-text search interface for the full NORA collection by the end of 2009.

4 Estimates of Effort — Budget

The project will be a collaboration between professors Stephan Oepen and Jan Tore Lønning, both at the UiO Department of Informatics, on the one hand, and senior engineer Morten Erlandsen, at the UiO Center for Information Technology (USIT), on the other hand.

The project will employ a research assistant for the equivalent of three months of full-time employment. It is unlikely that this work can commence before the summer, hence we expect to conduct the project between August 1 and December 31, 2009.

The assistant will carry out the practical work under supervision. She or he will

- test out the different tools for text extraction;
- compare results both quantitatively and qualitatively;
- consider what types of clean-up need to be done and
- apply and adapt available techniques for clean-up;
- create a Lucene-based interface for full-text search.

The assistant will write up the results from these feasability studies in a project-final report.

The cost of the research assistant is estimated as NOK 52,400 per month, all included. To allow for some minor budget adjustments, we apply for a total amount of NOK 160,000.

References

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., & Tan, Y. F. (2008). The ACL Anthology Reference Corpus. A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech, Morocco.

Schäfer, U., Uszkoreit, H., Federmann, C., Marek, T., & Zhang, Y. (2008). Extracting and querying relations in scientific papers. In *Proceedings of the 31st Annual German Conference on Artificial Intelligence* (pp. 127 – 134). Kaiser-slautern, Germany: Springer.

Siddharthan, A., & Teufel, S. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 316,– 323). Rochester, NY.