

Computational Linguistics meets Metadata, or the automatic extraction of key words from full text content

Marilyn Deegan (marilyn.deegan@kcl.ac.uk)

Harold Short (harold.short@kcl.ac.uk)

King's College London

Dawn Archer (d.archer@lancaster.ac.uk)

Paul Baker (p.baker@lancaster.ac.uk)

Tony McEnery (eiaamme@exchange.lancs.ac.uk)

Paul Rayson (paul@comp.lancs.ac.uk)

Lancaster University

Introduction

For the past year, the Centre for Computing in the Humanities (CCH), at King's College London (www.kcl.ac.uk/cch) and the Forced Migration Online (www.forcedmigration.org) team at the Refugee Studies Centre, University of Oxford (www.rsc.ox.ac.uk) have been working together to investigate the use of computational linguistics techniques for extraction of keywords from full-text content in a pilot project funded by the Andrew W Mellon Foundation.

The starting point for this project was the realization that it is easier to digitize large volumes of textual data than it is to create bibliographic records, and that it is particularly time consuming and expensive to add intellectual data such as keywords and abstracts. Our engagement with these issues grew out of several years work on the development of and investigation into hybrid and digital libraries through the Malibu project (Managing the hybrid Library for the Benefit of Users, www.kcl.ac.uk/cch/malibu) and the Forced Migration Online digital library, in both of which the Centre for Computing in the Humanities at Kings College London and the Refugee Studies Centre at Oxford University were centrally involved. The Malibu project ran from 1998 to 2001, Forced Migration Online has been ongoing since 1997.

A great deal of progress has been made in automating the capture of full text from printed documents by the production scanning of print originals or surrogates followed by the application of advanced optical character recognition (OCR) algorithms. Once text is produced, there are also sophisticated systems for full-text search using pattern matching or fuzzy matching which offer excellent retrieval. However, the use of bibliographic descriptions, the addition of keywords to a document and the application of topics trees and other taxonomic devices are still needed to improve precision and recall, and these meta activities generally still

need a great deal of human time, effort and skill. Some elements of bibliographic description are relatively easy to add to a documentary source, but the addition of keywords and other classificatory information still generally needs expert work, which is time-consuming and costly, and can also be highly subjective. Taxonomy is an intellectually demanding activity, and there have been many aids to this developed over the last century in the form of classification schemes and subject thesauri but assigning the terms is still a manual process. An ancillary problem is that as subject areas grow and change, the classification schemes need continual updating, and a circular process exists where thesauri inform classification and new classifications in turn inform the further development of the thesauri.

Some argue that adding value to content is unnecessary as search engine will always find it, but a) that depends on knowing what you are searching for and b) can result in over-retrieval. As a recent commentator has remarked:

Is it time to detach from our reliance on search engines? Consider the reality of relying on your favorite search engine. You're applying a pretty dumb technology (search algorithms) against a huge, undifferentiated pile of randomly selected, unorganized content; then adding billions of dollars of keyword-matched ads to the sorted output. Moreover, the effect over time of persistent ad placement in search results is to push those web resources that lack the capacity or interest in placing ads further down the search results list and out of sight of most searchers.

www.workingfaster.com/sitelines/archives/2004_03.html#000176

What we were interested in is using intelligent algorithms that have been developed according to some statistical and/or linguistic principles to aid, not in the searching, but in the classification and keyword extraction processes to gain the benefits of automation with the precision of human-generated work. Were we successful? Read on ...

The subject matter: grey literature about forced migration

FMO (www.forcedmigration.org) is a portal which provides access to a wide variety of online resources dealing with the situation of refugees and forced migrants worldwide. Designed for use by practitioners, policy makers, researchers, students or anyone interested in the field, FMO aims to give comprehensive information in an impartial environment and to promote increased awareness of human displacement issues to an international community of users. There is a great deal of content, with some 80,000 pages of full text content in the digital library and journals, as well as several thousand records in a web catalogue and organizations directory. This content forms an excellent test set of diverse types of information based on one particular domain. What is particularly interesting here is that the content derives from many different kinds of agencies and individuals: the academic sphere, governmental organizations, non-governmental organizations, the press, and this has significant effects on the results of the

various trials to extract keywords from them. Content on forced migration and refugee issues outside of Forced Migration Online has also been used in some of the trials described below: up-to-date news from the UNHCR web site as well as current newspaper content has also been analysed.

The full text content on Forced Migration Online outside of the journals is largely grey literature which presents particular problems for bibliographic description, classification, and assignment of subject terms and keywords. Organizations such as the Refugee Studies Centre (www.rsc.ox.ac.uk) have built collections of grey literature explicitly because it is rare, difficult to get hold of, and difficult to find in major library collections. Given the particular nature of this growing field, too, classification and thesaurus support are weak in the major classification schemas and thesauri, and it is generally not possible to obtain records from the major suppliers.

Thesaurus development

A key input for the trials on keyword extraction described below is the UNHCR Thesaurus of Refugee Terminology (*ITRT*), which was designed to facilitate information retrieval and exchange. In print since 1988, this has hitherto only been available in paper form. In 2003, in parallel with the Keyword Project, the UNHCR Library and FMO began discussing how to create a web-based version of the Thesaurus that a) would be more responsive to the needs of its users and b) could be used in the course of the keyword extraction trials.

It was decided to move very rapidly towards the development of the online version of the *Thesaurus* and so FMO and UNHCR commissioned Oxford ArchDigital, an Oxford University spin-out company, to develop the resource using their ToadHMS product, a customizable content management system (www.oxarchdigital.com).

The Thesaurus is now available as an interactive and searchable tool online, in English, French, and Spanish (<http://refugeethesaurus.org/>). Launched in December 2003, this new version is already serving as a more efficient medium for identifying relevant indexing terminology and as a value-added mechanism for managing refugee- and forced migration-related information. The Thesaurus was ready just in time for the trials undertaken by the Lancaster teams. These trials are discussed below.

Pilot project research methods

With a focus on forced migration generally and the FMO collections in particular, we started out to investigate the following questions:

1. How might key terms be extracted from bodies of digital library materials in order to provide rich metadata that does not need to be human-generated?
2. How would these terms relate to the semantic environment that a thesaurus provides?
3. How could term extraction be 'improved' by use with thesauri, and thesauri 'improved' by term extraction?
4. What other work is being carried out that might inform developments in this area?
5. Who are the key players?
6. What commercial products are available?

The activities by which we carried out this study have been desk research; testing of data to prove concepts; extensive consulting in the community; an expert workshop to discuss findings and the way forward.

Desk research and consultation have yielded a great deal of information about this field. This is presented elsewhere (see the project web site at <http://www.kcl.ac.uk/humanities/cch/ake/final/content/index.html>). The rest of this article reports on the remarkable results of the testing of data by the Department of Linguistics and Modern English Language and the Department of Computing at Lancaster University.

The one-month challenge

At the end of December 2003, and after some initial discussions, the Departments of Linguistics and Computing, Lancaster University, agreed to carry out two separate investigations on refugee material in time for a workshop to be held at the beginning of February 2004 workshop.ⁱⁱ The short timescales for this led to it becoming known as the 'one-month challenge'! Given the emphasis on keyword analysis, these were carried out by members of UCRELⁱⁱⁱ, a cross-departmental research centre that specializes in the automatic/computer-aided analysis of large bodies of naturally occurring language. In one experiment, Archer and Rayson semantically annotated material provided by the Forced Migration Online team, using the UCREL Semantic Annotation System (USAS), a software package for automatic dictionary-based content analysis. In the other, Baker and McEnery collected their own refugee data from the news section of the UNHCR web site and from online newspapers, and performed keyword analyses on that data, using Wordsmith^{iv} and similar tools.

Each team agreed to work separately, and to keep their findings secret until they presented their respective results at the Keywords workshop. The results were remarkable, as was the occasion as, when the first presenters (Archer and Rayson) were speaking, they did not know what the other team were going to say. What surprised everyone (the presenters included) was the close correspondence between the results of the two experiments.

The Archer and Rayson trial

What Archer and Rayson did was investigate in detail the benefits of semantically annotating refugee material, using the UCREL (University Centre for Computer Corpus Research on Language) system (henceforth USAS) and the feasibility of mapping the semantic domains of USAS to the classes used in the UNHCR Refugee Thesaurus. The USAS system is designed to undertake the automatic semantic analysis of present-day English texts (spoken and written), and this involves two stages:

(i) A part-of-speech tag is assigned to every lexical item or multi-word expression (MWE), using probabilistic Markov models of likely part-of-speech sequences (- 97% accuracy)

(ii) Output is fed into SEMTAG, which assigns semantic field tags on the basis of pattern matching between the text and two computer dictionaries developed for use with the program, and then applies a set of disambiguation techniques intended to select the correct semantic tag on each item given its context (- 92% accuracy)

The present applications of the system include: linguistic analysis, market research, content analysis, information extraction, assistance for translation. USAS (via its web interface called Wmatrix) is a quantitative content analysis tool which can automatically (i) measure/compare the frequency of occurrence of different domains; (ii) provide statistical information regarding key concepts; (iii) Provide a record of the vocabulary resources for those domains. It offers therefore a useful means of assessing the differing themes, concerns, attitudes (and mindsets/world views) of various texts/authors/institutions.

The trial on the Forced Migration Online data involved mapping of top level categories of the *ITRT* onto the USAS categories, and then analysing a number of documents provided by FMO. These documents were drawn from different domains and agencies, categorized as UNHCR, Federation of the Red Cross, Government agencies (general), NGOs (general) and Academic (mostly FMO grey literature). The total number of words in the document set was 432,317.

The results of this were most promising. USAS was able to map onto the Thesaurus top level categories and to categorize the documents from the different sectors with ease. One consequence of the trial was the identification of a number of terms that were not represented in the Thesaurus that represent important topics in the documents, thus proving the value of automated techniques for 'improving' thesauri, one of the stated intentions of the pilot project. Another was the analysis of attitudinal factors represented by the different agencies, as well as the topic categorizations of the documents. The system is able to categorize data within Forced Migration Online with a considerable degree of accuracy, and to assign keywords to it at least as well as human cataloguers. What is very promising for future work is how rapidly large volumes of data can be processed.

The Baker and McEnergy trial

In their trial, Paul Baker and Tony McEnergy made comparisons between news on refugees as reported by UNHCR on their web site (<http://www.unhcr.ch/cgi-bin/texis/vtx/news>) throughout 2003 and news on refugees as reported in a wide range of British newspapers during 2003. The analyses were carried out using the corpus analysis software package, WordSmith Tools (<http://www.lexically.net/wordsmith/>). The results were again most impressive. Themes and ideas could be extracted readily from the texts, as with the Archer and Rayson trial, but what became apparent here was the difference in tone between UNHCR (an intergovernmental agency) and the press. UNHCR overall used a neutral tone of reporting, while the press used highly emotive, persuasive and manipulative terminology. Baker and McEnergy found very different discourses in different representations of reality, and observed that data matters when constructing resources to reflect the world, for different discourses represent different worlds.

What was most illuminating for those present at the initial presentations of these two papers was a) the uncanny correspondence between the themes, ideas, tones and keywords extracted by the different methods on different corpora representing the same subject domain, and b) the ability to extract from the texts matters which are of urgent and current concern to those dealing with forced migration. Those of us who were familiar with the field were struck forcibly by the accuracy with which the corpus linguists who had hitherto had little exposure to this area could present the issues. It was a graphic demonstration of what we had hoped for in the project, which is that computational linguistics techniques could be used to extract accurate keywords from digital library content in a meaningful way.

Conclusion

What has been particularly interesting and productive about the project is that it has involved the collaboration of individuals and research groups from a number of different domains who have rather different methodological perspectives, and who do not normally engage closely with each other, viz computational linguists, information specialists, digital library specialists, humanities computing specialists, system designers and specialists in forced migration. This led to some terminological misunderstandings, and there was much discussion during the course of the work about differences in terminology between different domains and approaches. For instance, the word ‘thesaurus’ can mean a list of controlled terms in a library cataloguing environment and a list of synonyms and antonyms (eg Roget’s Thesaurus). See (<http://www.kcl.ac.uk/humanities/cch/ake/final/content/about/about.html>) for a full list of project participants.^v

The success of the experiments has surprised all those who have taken part in them. The various propositions we started out with were confirmed, and the tagging and analysis systems worked with remarkably little trouble. The teams are planning a number of follow-up projects, including mapping the USAS system to the *ITRT* and using this as a browse tool for Forced Migration Online, and testing the systems on different domains of grey literature.

ⁱ See <http://www.unhcr.ch/cgi-bin/texis/vtx/home?page=research&id=3d884d834> for details about the UNHCR Library

ⁱⁱ See www.kcl.ac.uk/humanities/cch/ake/final/content/events/workshop/index.html for the workshop programme, participants’ list, report and all workshop presentations.

ⁱⁱⁱ UCREL is the acronym for University Centre for Computer Corpus Research on Language.

^{iv} See Scott, M. (1999) *WordSmith Tools Help Manual*. Version 3.0. Mike Scott and Oxford University Press for details about WordSmith.

^v A major publication is being produced by the project. Current versions of the essays are available at www.kcl.ac.uk/humanities/cch/ake/final/content/pubs/pub01.html.