Developments in information retrieval:
Part 2

In this, the second part of his article, Itamar Kastner reviews the most topical research papers presented at the European Chapter of the Association for Computational Linguistics conference.

AT THE NATIONAL Library of Medicine at Bethesda, Maryland, US, researchers have been trying to automatically assign tags to images in their collection, a task which is both costly to have done manually and potentially inconsistent. Although their system takes a standard approach of learning from existing training data comprised of captions and existing indexing terms, they also had a number of rather original features: expanding medical terms to their synonyms using the medical database MeSH and the Unified Medical Language System was one such feature. This eventually led to between 30-50 indexing terms per document.

While the work in Bethesda is typical of the field, the European Poeticon project has different ambitions. Part of this project, Cosmoroe, focuses on video annotation, both literal and figurative, in an attempt to model relations between language, video and human actions. In what they call ‘multimedia dialectics’, the Athenian researchers annotate a video with both literal and figurative actions. Of particular interest are the figurative ‘metaphorical’ and metonymic entities. In one scene, a group of musicians is playing while a tour guide says ‘We enjoy the melodies while walking through the old streets’. The annotation relates ‘enjoy melody’ to listen on the textual level, and the movement of the musicians to music on the physical level. The combination is then stored and can easily be found using a simple search interface. The goal is to automatically learn these annotations and recognise similar dialectics. Current research is however still in early stages. A number of other interesting studies are highly recommended:

● Staying on the cultural side of things, German researchers are trying to use NLP methods in order to parse and catalogue museum exhibits. Their current case study is an exhibition documenting the goldsmith’s art.

● A paper on the importance and prospects of digital preservation.

● At this point, I’m going to shamelessly plug my own paper. This study considers the ‘Story Highlights’ on the CNN.com website and uses relatively simple linguistic features in order to automatically extract them from the story text.

● In a similar vein, other work tries to extract biographical facts about a person from an article (usually a Wikipedia article).

● Work done in EML Research Heidelberg aims to be even more useful for real-world applications, seeking out news items regarding a given company by automatically going through a large collection of financial news (the Yahoo! Finance website) and extracting only the sentences discussing the said company.

● Extending search queries intelligently has always

Their system takes a standard approach of learning from existing training data comprised of captions and existing indexing terms; they also had a number of rather original features: expanding medical terms to their synonyms using MeSH and the Unified Medical Language System.

Itamar Kastner (itk1@dcs.qmul.ac.uk) is a Linguistics and Computer Science student at the University of London. His travel to EACL was made possible in part by the John Campbell Trust.
was a taste of the sort of relevant research presented at EACL 2009, its relation to the two complementary fields of term indexing and metadata, and the NLP concepts in between. Having said that, the computational linguistics community is quite industrious: EACL 2009 took place in early April, but NAACL 2009 (the conference of the North American Chapter) took place in June and ACL 2009 (the main international conference) landed in Singapore for the 2009 meeting. So there is a great deal of research going on. Hopefully information scientists can find ways in which it is applicable for their work as well, and vice versa.

On the cultural side of things, German researchers are trying to use Natural Language Processing methods in order to parse and catalogue museum exhibits. Their current case study is an exhibition documenting the goldsmith’s art.

References
1 Simpson et al. 'Using Non-lexical Features to Identify Effective Indexing Terms for Biomedical Illustrations.' In EACL 2009.
2 Pastra and Balta. 'A Text-based Search Interface for Multimedia Dialectics.' In EACL 2009 Demo Session.
3 Goerz and Scholz. 'Content Analysis of Museum Documentation with a Transdisciplinary Perspective.' In LaTeCH-SHELT&R 2009 Workshop.
4 Dobreva and Ikonomov. 'The Role of Metadata in the Longevity of Cultural Heritage Resources.' In LaTeCH-SHELT&R 2009 Workshop.
5 Kastner and Monz. 'Automatic Single-Document Key Fact Extraction from Newswire Articles.' In EACL 2009.
7 Filippova et al. 'Company-Oriented Extractive Summarization of Financial News.' In EACL 2009.
8 Davidov and Rappoport. 'Translation and Extension of Concepts across Languages.' In EACL 2009.
9 Wennerberg. 'Aligning Medical Domain Ontologies for Clinical Query Extraction.' In EACL 2009 Student Workshop.

been an important part of search engines and question-answering mechanisms. Research in the Hebrew University of Jerusalem extends concepts across more than 30 languages, expanding the search accordingly and fetching additional relevant documents.●And if we can search for words, why not search for images once again. Part of the German Theseus Medico image search project concerns retrieval of appropriate images from multiple medical ontologies. The practitioner would query a certain illness and receive X-rays, health records, etc. relating to that illness.●