Multilingual Retrieval in Twente.

Franciska de Jong and Dennis Reidsma and Djoerd Hiemstra
TKI research group, dept. of Computer Science, University of Twente
{*fdejong@cs.utwente.nl*}

Abstract

The participation of the University of Twente in CLEF 2003 was very low profile. The eternal pressure on resources known to researchers world-wide caused the focus of activities to be on other projects for some time. Therefore this paper will not go into the techniques with which the results of Twente were produced but rather gives an overview of current work in retrieval at the University of Twente.

1 Current Work in Twente.

The TKI reserch group has participated and is participating in several past and ongoing Dutch and European retrieval projects concerned with the disclosure of multimedia archives. Besides tackling the problem of processing the multimedia content, many of these projects focus on using parallel textual content as an extra source of information. This content can help searching effectively on information that cannot be derived from the multimedia content in itself.

Within the Dutch project DRUID[1] a lot of work has been done on robust speech recognition for Dutch[2], video segmentation and information extraction and filtering[3]. The IST project ECHO[2] aimed at disclosing *historic film materials*. The major research themes were content-based processing of the video material and speech processing. Both projects had a strong focus on *cross lingual retrieval* of the content, expanding the results of previous projects.

The “Waterland project” (an ongoing Dutch project) does not directly concern multilingual retrieval, but it investigates a flexible and generic architecture for distribution of multimedia content.

In Pidgin[3] retrieval is not the major theme. However, crosslanguage aspects figure heavily in this project: the resulting demonstrator should be able to generate automatic “translations” of user utterances, to enable conversations between persons using different languages. The translations need not be grammatically perfect, but should be good enough to convey the correct meaning.

MUMIS[4], a recently completed IST project, was aimed at disclosing *video recordings of soccer matches*. The use of parallel textual reports on those matches played a major role in this project. Furthermore a first start with TKI research on Information Extraction was made in the MUMIS project. The most interesting development was automatic alignment and merging of the extracted information from separate sources, which should improve the quality of retrieval [1].

The above can be summarised as follows: the major part of the TKI research on retrieval focuses on *multilingual* access to *multimedia* content and in the future more work will be done on Information Extraction for Dutch and especially on the promising theme of cross document information extraction, using information from one source to improve the extraction for another source.

[1] http://dis.tpd.tno.nl/druid/
[2] http://pc-erato2.iei.pi.cnr.it/echo/
[3] http://www.pidgin.nl/
[4] http://parlevink.cs.utwente.nl/projects/mumis/

References

[1] Jan Kuper, Horacio Saggion, Hamish Cunningham, Thierry Declerck, Peter Wittenburg, and Dennis Reidsma. Intelligent multimedia indexing and retrieval through multi-source information extraction and merging. In *Proceedings of the IJCAI'03, Acapulco*, August 2003. to appear.

[2] Roeland Ordelman, Arjan van Hessen, Franciska de Jong, and David van Leeuwen. Speech recognition for dutch spoken document retrieval. In David Bearman and Franca Garzotto, editors, *Proceedings from the ICHIM01 meeting, Volume 2: Short Papers/Posters and Demos, Milan*, September 2001.

[3] Martijn Spitters and Wessel Kraaij. A language modeling approach to tracking news events. In *CBMI'01, Brescia*, 2001.