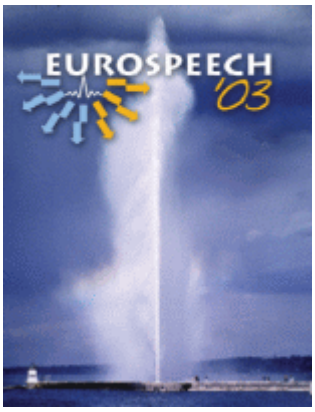
ISCA Archive
http://www.isca-speech.org/archive

EUROSPEECH 2003 - INTERSPEECH 2003
8th European Conference on Speech Communication and Technology

Geneva, Switzerland
September 1-4, 2003



Multi-Scale Document Expansion in English-Mandarin Cross-Language Spoken Document Retrieval

Wai-Kit Lo (1), Yuk-Chi Li (1), Gina Levow (2), Hsin-Min Wang (3), Helen M. Meng (1)

(1) Chinese University of Hong Kong, China
(2) University of Chicago, USA
(3) Academia Sinica, Taiwan

This paper presents the application of document expansion using a side collection to a cross-language spoken document retrieval (CL-SDR) task to improve retrieval performance. Document expansion is applied to a series of English-Mandarin CL-SDR experiments using selected retrieval models (probabilistic belief network, vector space model, and HMM-based retrieval model). English textual queries are used to retrieve relevant documents from an archive of Mandarin radio broadcast news. We have devised a multi-scale approach for document expansion - a process that enriches the Mandarin spoken document collection in order to improve overall retrieval performance. A document is expanded by (i) first retrieving related documents on a character bigram scale, (ii) then extracting word units from such related documents as expansion terms to augment the original document and (iii) finally indexing all documents in the collection by means of character bigrams and those expanded terms by within-word character bigrams to prepare for future retrieval. Hence the document expansion approach is multi-scale as it involves both word and subword scales. Experimental results show that this approach achieves performance improvements up to 14% across several retrieval models.

Full Paper

Bibliographic reference. Lo, Wai-Kit / Li, Yuk-Chi / Levow, Gina / Wang, Hsin-Min / Meng, Helen M. (2003): "Multi-scale document expansion in English-Mandarin cross-language spoken document retrieval", In *EUROSPEECH-2003*, 2337-2340.