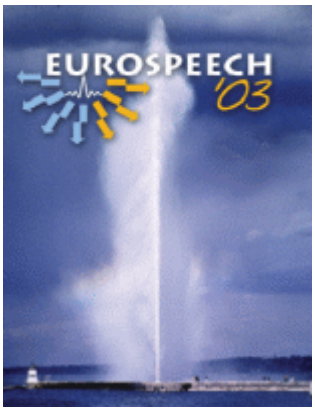
ISCA Archive
http://www.isca-speech.org/archive

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

Geneva, Switzerland
September 1-4, 2003



Automatic Extraction of Bilingual Chunk Lexicon for Spoken Language Translation

Limin Du, Boxing Chen

Chinese Academy of Sciences, China

In language communication, an utterance may be segmented as a concatenation of chunks that are reasonable in syntax, meaningful in semantics, and composed of several words. Usually, the order of words within chunks is fixed, and the order of chunks within an utterance is rather flexible. The improvement of spoken language translation could benefit from using bilingual chunks. This paper presents a statistical algorithm to build the bilingual chunk-lexicon automatically from spoken language corpora. Several association measurements are set up as the criteria of the extraction. And local best algorithm, length ratio filtration and stop-word filtration are also incorporated to improve the performance. A bilingual chunk-lexicon was extracted from a corpus with precision of 86.0% and recall of 86.7%. The usability of the chunk-lexicon was then tested with an innovative framework for English-to-Chinese Spoken Language translation, resulted in translation accuracy of 81.83% and 78.69% for training and test sets respectively, measured with Levenshtein distance based similarity score.

Full Paper

Bibliographic reference. Du, Limin / Chen, Boxing (2003): "Automatic extraction of bilingual chunk lexicon for spoken language translation", In *EUROSPEECH-2003*, 2333-2336.