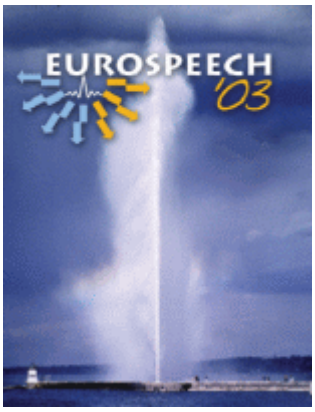
ISCA Archive
http://www.isca-speech.org/archive

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

Geneva, Switzerland
September 1-4, 2003



Using the Web for Fast Language Model Construction in Minority Languages

Viet Bac Le (1), Brigitte Bigi (1), Laurent Besacier (1), Eric Castelli (2)

(1) CLIPS-IMAG Laboratory, France
(2) MICA Center, Vietnam

The design and construction of a language model for minority languages is a hard task. By minority language, we mean a language with small available resources, especially for the statistical learning problem. In this paper, a new methodology for fast language model construction in minority languages is proposed. It is based on the use of Web resources to collect and make efficient textual corpora. By using some filtering techniques, this methodology allows a quick and efficient construction of a language model with a small cost in term of computational and human resources. Our primary experiments have shown excellent performance of the Web language models vs newspaper language models using the proposed filtering methods on a majority language (French). Following the same way for a minority language (Vietnamese), a valuable language model was constructed in 3 month with only 15% new development to modify some filtering tools.

Full Paper

Bibliographic reference. Le, Viet Bac / Bigi, Brigitte / Besacier, Laurent / Castelli, Eric (2003): "Using the web for fast language model construction in minority languages", In *EUROSPEECH-2003*, 3117-3120.