INTERSPEECH 2006 - ICSLP
Ninth International Conference on Spoken Language Processing

Pittsburgh, PA, USA
September 17-21, 2006

Linguistic Tuple Segmentation in N-gram-Based Statistical Machine Translation

Adri□ de Gispert, Jos□ B. Mari□o

Universitat Polit□cnica de Catalunya, Spain

Ngram-based Statistical Machine Translation relies on a standard Ngram language model of tuples to estimate the translation process. In training, this translation model requires a segmentation of each parallel sentence, which involves taking a hard decision on tuple segmentation when a word is not linked during word alignment. This is especially critical when this word appears in the target language, as this hard decision is compulsory.

In this paper we present a thorough study of this situation, comparing for the first time each of the proposed techniques in two independent tasks, namely English-Spanish European Parliament Proceedings large-vocabulary task and Arabic-English Basic Travel Expressions small-data task. In the face of this comparison, we present a novel segmentation technique which incorporates linguistic information. Results obtained in both tasks outperform all previous techniques.

Full Paper

Bibliographic reference. Gispert, Adri□ de / Mari□o, Jos□ B. (2006): "Linguistic tuple segmentation in n-gram-based statistical machine translation", In *INTERSPEECH-2006*, paper 1049-Tue2CaP.1.