[CICLING 2010: 11th International Conference on Intelligent Text Processing and Computational Linguistics, March 21-27, 2010, Iaşi, Romania]

Abstract

Peng Li, Maosong Sun and Ping Xue. Fast-Champollion : A Fast and Robust Sentence Alignment Algorithm

Aligned parallel texts are important resources to many natural language processing tasks including statistical machine translation, etc. With the rapid growth of online parallel texts, efficient and robust sentence alignment methods become increasingly important. In this paper, we propose a fast and robust sentence alignment algorithm, i.e., Fast-Champollion, which employs a combination of both length-based method and lexicon-based method. By optimizing the process of splitting the input bilingual texts into small segments for alignment, Fast-Champollion, as our extensive experiments show, is 3.9 to 9.0 times as fast as the baseline method Champollion on short texts and about 50.6 times long texts, and Fast-Champollion is as robust as Champollion.