ACM Transactions on Speech and Language Processing 5 (2), 2008, art.4.

Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita:
Chinese word segmentation and statistical machine translation.

Abstract

Chinese word segmentation (CWS) is a necessary step in Chinese-English statistical machine translation (SMT) and its performance has an impact on the results of SMT. However, there are many choices involved in creating a CWS system such as various specifications and CWS methods. The choices made will create a new CWS scheme, but whether it will produce a superior or inferior translation has remained unknown to date. This article examines the relationship between CWS and SMT. The effects of CWS on SMT were investigated using different specifications and CWS methods. Four specifications were selected for investigation: Beijing University (PKU), Hong Kong City University (CITYU), Microsoft Research (MSR), and Academia SINICA (AS). We created 16 CWS schemes under different settings to examine the relationship between CWS and SMT. Our experimental results showed that the MSR's specifications produced the lowest quality translations. In examining the effects of CWS methods, we tested dictionary-based and CRF-based approaches and found there was no significant difference between the two in the quality of the resulting translations. We also found the correlation between the CWS F-score and SMT BLEU score was very weak. We analyzed CWS errors and their effect on SMT by evaluating systems trained with and without these errors. This article also proposes two methods for combining advantages of different specifications: a simple concatenation of training data and a feature interpolation approach in which the same types of features of translation models from various CWS schemes are linearly interpolated. We found these approaches were very effective in improving the quality of translations.