Abstract

Translating from English into a morphologically richer language like Arabic is a challenge in statistical machine translation. Segmentation of Arabic text was introduced to bridge the inflection morphology gap. In this work, we investigate the impact of supporting Arabic morphologically segmented training corpus in a phrase-based statistical machine translation system with one to one dictionary and examine the effects on system performance. The results show that the dictionary improves the quality of the translation output especially when the corpus used is normalized and fully segmented excluding the determiner. The dictionary also decreases the out of vocabulary rate. The effect of the dictionary support with different baseline and factored models using data ranging from full word form to fully segmented forms are also demonstrated.

References

A Statistical Method for English to Arabic Machine Translation


- Stolcke, Andreas. SRILM-an extensible language modeling toolkit. in Proceedings of the international conference on spoken language processing, pp. 901-904, 2002.


Index Terms

Computer Science

Keywords

Statistical machine translation Factored phrase based Natural language processing.