
Abstract

This desertation presents a new classification model for constructing English Persian parallel bilingual corpus. The results have shown that both of linguistic and non-linguistic features will improve performance of sentence aligning at an acceptable level. The proposed multi classifier model has crucial role to recognize aligning types. Also we studied the effects of the linguistics characteristics, literal translation and grammatical similarities independently or in combination on the quality results. The proposed model supports adding new features to amplify the power of aligner engine. The challenge of the proposed and other methods relate to variety of translation ways and the translators favorites especially when the similarity between two phrases is only semantic , usually it can be recognized by only human.

Keywords: English Persian corpus, SMT, Sentence Aligning, Machine translation