Abstract

The availability of Sundanese-Indonesian parallel corpus are few in number. Parallel corpus is important and could be used as a training data source for machine translation system or natural language processing system. This work is trying to collect parallel sentences extracted from pairs of Wikipedia articles using interlanguage links facility. A bilingual lexicon and a series of filters based on word occurrence, sentence length and word overlap between sentences were used to collect parallel sentence. The bootstrapping method was used to improve the quality of the parallel sentence by updating the bilingual lexicon using IBM Model 4 EM learner implemented in GIZA++ tool. GIZA++ was run on parallel sentence candidate which was generated in each iteration until the system reached convergence state. Manual evaluation result using human judgement shows that 79,5% of parallel corpus built by the system had proven to be parallel.

Keywords: parallel corpus, Wikipedia, bootstrapping, expectation maximization