

Abstract

This thesis proposes an approach to improve Thai word segmentation by utilizing Named Entity Recognition (NER). The Conditional Random Fields (CRFs) is applied to learn a Thai Named Entity Recognition (NER) model. Prefixes and suffixes of named entities are used as main the features for learning the model. Several models are constructed and compared by based on three word-level grams: 3, 5 and 7. The corpus used for training and evaluating the models is the BEST 2009 (Benchmark for Enhancing the Standard of Thai language processing), which consists of 5 million words. The study finds that the amount of 7 gram provides the best efficiency of NER Model. The NER Model is used to improve the efficiency of Thai word segmentation. The experiment finds word segmentation in the program. Thai Lexeme Analyser (Tlex), approximately 37 articles with 72,000 words, has F1 as equal to 92.23%. And improved Thai word segmentation with named entity recognition with NER Model, approximately 6 articles, it results in F1 as equal to 93.80% , 1.57% increase