Abstract

Syllabication or syllabification is an activity to detect syllable boundaries in a word. There are two main ways for automatic syl-labification, namely rule-based and data-driven. The rule-based approach is based on the general principle of syllabification, while the data-driven method uses a set of syllabified words to create a syllabification of unknown words. Research on syllabification of words has been done a lot. However, most of these studies only deal with the formal words but still a few studies for named entities. Besides, named entities tend to be more complicated than the regular words. In this research, a syntactic n-Gram is proposed and investigated to syllabify the named entities since it is developed based on the n-gram that has an excellent accuracy and tends to be consistent with various languages. Evaluation on 20 k named-entities based on 4-fold cross-validation show that the proposed model gives a competitive syllable error rate (SER) compare to another similar n-gram-based model.

Keywords: syllabification; named-entities; syntactic n-gram;