論 文 内 容 の 要 旨

博士論文題目
Word Segmentation and Lexical Normalization for Unsegmented Languages
(分かち書きされない言語における単語分割と語彙正規化)

氏　　名 東山 翔平

(論文内容の要旨)

Word segmentation is a fundamental technology in natural language processing in unsegmented languages, such as Japanese and Chinese. Although accurate segmentation is required to prevent error propagation to downstream tasks, proper approaches can be different according to the characteristics of and available linguistic resources in text domains. For general (and any) domains, a disambiguation method of ambiguous word boundaries is necessary. For specialized domains, the use of unlabeled data and lexicons is effective to recognize domain-specific words. For user-generated text domains, lexical normalization of nonstandard words is important because those words can degrade performance of word segmentation and downstream tasks. This research focuses on our four studies to achieve accurate Japanese and Chinese word segmentation in various domains and Japanese lexical normalization in user-generated text.

In the first study, a neural character-based model is investigated that learns the importance of each character in multiple candidate words in order to improve the performance on in-domain semi-supervised word segmentation by better disambiguation of word boundaries. In the second study, a model for cross-domain word segmentation is proposed that learns the occurrences of lexical words in unlabeled target sentences together with segmentation label information from labeled source sentences, in which little labeled data is available in target domains. The third and fourth studies tackle two problems in Japanese word segmentation and lexical normalization of user-generated text: the lack of public evaluation data and that of training data. In the third study for the former problem, a public Japanese evaluation dataset is constructed with morphological normalization and word category information for detailed evaluation of segmentation and normalization systems. In the last study for the latter problem, pseudo-label

氏　名	東山 翔平

（論文審査結果の要旨）

Word segmentation is a fundamental technology in natural language processing in unsegmented languages, such as Japanese and Chinese. Although accurate segmentation is required to prevent error propagation to downstream tasks, proper approaches can be different according to the characteristics of and available linguistic resources in text domains. For general domains, a disambiguation method of ambiguous word boundaries is necessary. For specialized domains, the use of unlabeled data and lexicons is effective to recognize domain-specific words. For user-generated text domains, lexical normalization of nonstandard words is important because those words can degrade performance of word segmentation and downstream tasks. This research focuses on our four studies to achieve accurate Japanese and Chinese word segmentation in various domains and Japanese lexical normalization in user-generated text.

In the first study, a neural character-based model is investigated that learns the importance of each character in multiple candidate words by disambiguating word boundaries. In the second study, a model for cross-domain word segmentation is proposed that learns the occurrences of lexical words in unlabeled target sentences together with segmentation label information from labeled source sentences. In the third study, a public Japanese evaluation dataset is constructed with morphological normalization and word category information for detailed evaluation of segmentation and normalization systems. In the last study, training data augmentation methods are investigated, and a neural text editing model is proposed for joint Japanese word segmentation, part-of-speech tagging, and lexical normalization.

Through these studies in this thesis, the research demonstrated that the proposed segmentation methods for different domain types achieved accurate performance in various domains. In addition, the Japanese lexical normalization method could be employed in practical systems in the future, and the evaluation dataset would be a useful benchmark for fair comparisons. The four studies are published in two high quality peer-reviewed journal papers, two peer-reviewed top international conference papers, and one peer-reviewed international workshop paper. The research would have an influence for further investigations, and would have an impact to the end applications, such as information extraction, demanding accurate word segmentation. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.