論 文 内 容 の 要 旨

博士論文題目

Towards Morphological And Syntactic Analyses For The Khmer Language

(クメール語の形態素および統語解析に関する研究)

氏　名　　KAING HOUR

(論文内容の要旨)

Language is a structural system that humans use for communication. Humans analyze structures of sentences to understand meanings and to generate grammatical sentences. In natural language processing (NLP), the same structural analysis is necessary to enhance computers' understanding and generation of natural language sentences. Its usefulness has been demonstrated for many downstream systems e.g., language model, machine translation, text summarization. Even though the structural analysis for computers has been studied for a long time, many languages have not yet been investigated because of their resource scarcity and a requirement of language-specific knowledge. This problem is crucial, particularly in this multilingual NLP era. This thesis concerns morphological and syntactic analyses for a low-resource language---Khmer. We organize our contributions into two parts.

Part one is about morphological analysis, which is the most fundamental task in NLP. The Khmer language is highly analytic and has no word boundary indicators. Previously, tokenization and part-of speech (POS) tagging have been studied for words and have been considered as two separate tasks. For the Khmer language, the definition of tokens and the classification of the POS is intrinsically ambiguous, which makes the tasks challenging. This thesis focuses on morphology to investigate joint processing for tokenization and POS-tagging. Our contribution is a preparation of the

largest tokenized and POS-tagged corpus that contains annotation for more linguistic phenomena than a previous corpus, and an investigation of the automatic processing of tokenization and POS tagging on this corpus using various state-of-the-art approaches. We present analyses and discussions based on our experimental results to show the achievements and limitations of the data and approaches.

Part two is about syntactic analysis for Khmer. We focus on constituency parsing, which is another core NLP task. Because the development of constituency treebank is costly, in this thesis, we investigate cross-lingual transfer learning techniques for constituency parsing. We conduct two experiments for many diverse languages in addition to Khmer. The first experiment discusses the single-source transfer performance of a POS-based delexicalized model and the effectiveness of various source-language selection techniques. Then, the second experiment examines a multi-source transfer using a pretrained multilingual language model as a cross-lingual channel. Because a multilingual constituency treebank consists of diverse structures and label sets, we propose a treebank preprocessing step and typological features integration with smooth sampling and dropout to improve the cross-lingual performance.

氏　名	KAING HOUR

（論文審査結果の要旨）

This thesis developed fundamental natural language processing modules for Khmer language, such as tokenization, part-of-speech (POS) tagging, and constituency parsing. This thesis contributed in two parts. Part one is about tokenization and POS tagging tasks for Khmer language, which utilizes a complex writing system and lacks word boundary indicator. The contribution of the thesis is a development of a two-layers—morpheme and word level—tokenization and POS tagged data for Khmer, and an investigation of their automatic processing using various state-of-the-art sequence labeling approaches. Part two is about constituency parsing, which is another core NLP task. The thesis investigates cross-lingual transfer learning techniques. First, the thesis proposes a delexicalized constituency parser for single source transfer. Second, the thesis examines a multi-source transfer using a pretrained multilingual language model and confirmed its effectiveness for Khmer language.

The proposed two contributions provide better fundamental natural language processing modules for Khmer and can be applied to the various application in natural language processing applications in Khmer. The thesis research brought a significant contribution to Khmer language. A series of his research resulted in two high-quality peer-reviewed journal papers, one peer-reviewed international conference papers. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.