論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Incorporating Multi-granularity Linguistic Units in Character-based Word Segmentation
著者(和文)	CHAY-INTR Thodsaporn
Author(English)	Thodsaporn Chay-intr
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12542号, 授与年月日:2023年9月22日, 学位の種別:課程博士, 審査員:奥村 学,熊澤 逸夫,中山 実,篠崎 隆宏,船越 孝太郎
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12542号, Conferred date:2023/9/22, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 系 情報通信 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of （ 工学 ）
学生氏名： Student's Name	CHAY-INTR Thodsaporn	審査員主査： Chief Examiner	奥村 学 教授

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Word segmentation is a fundamental task for an understanding of natural language for most Asian languages, such as Japanese, Chinese, and Thai. The task is to determine word boundaries from a running text. Incorrect segmentation can lead to error propagation in subsequent tasks, such as Named Entity Recognition (NER), part-of-speech (POS) tagging, and parsing, emphasizing the importance of accurate word information.

Characters, in various combinations, can form new words with different roles, meanings, or grammatical properties. Consequently, a character sequence inherently contains segmentation alternatives. which gives rise to segmentation ambiguity. This ambiguity comes from the fact that a character sequence can be segmented into words in multiple valid ways. This type of ambiguity poses a significant challenge in word segmentation, as it can lead to incorrect segmentation results if not handled properly. Character-based models attempt to resolve this ambiguity implicitly through segmentation alternatives by learning underlying patterns and relationships between character units. Although these models could lessen ambiguity issues, they rely solely on character units, which may lack the inherent meaning often found in larger units such as words. Given this consideration, the inclusion of additional linguistic units, such as word units, alongside character units, may enhance the effectiveness of character-based models in handling segmentation alternatives, ultimately contributing to improved segmentation performance.

Previous studies have successfully utilized linguistic units either subwords or words, in addition to character units, to alleviate the ambiguity problem in character-based word segmentation. They focus on constructing a set of either potential subwords or words from a character sequence, with the aim of implicitly deriving multiple different segmentation alternatives. Despite the progress made by these studies, there are still limitations to be addressed. First, their approaches explore only one fundamental unit in addition to a character unit. Second, they do not jointly utilize multi-granularity linguistic units such as subwords and words together. Thus, further handling segmentation alternatives using a broader range of multi-granularity structures jointly may not be fully exploited and becomes a research aspect.

A set of possible segmentation alternatives can be represented in a graph structure, specifically a multi-path lattice. Previous works attempt to capture these alternatives by constructing a lattice based on character and word units along with word-boundary nodes. They utilize pre-trained models (PTMs) such as bidirectional encoder representations from transformers (BERT) and employ graph neural networks (GNNs) to encode the lattice. Despite its potential, this approach's segmentation performance is mostly on par with methods using multi-criteria (MC) segmentation across multiple datasets, which are based on PTMs, specifically BERT.

This may be due to two factors. First the method mainly focuses on constructing lattices to represent potential segmentation results. However, it only uses word boundary nodes to enhance character representations by a concatenation operation, rather than attentively using nodes from character and word units. Second, the method does not utilize multiple datasets. Thus, to further improve the segmentation performance, it remains a challenge to effectively leverage segmentation alternatives based multi-granularity linguistic units through the use of lattices for complementing character representations.

In this thesis, we explore two main aspects that serve as our goals. The first aspect aims to jointly utilize a broader range of multi-granularity linguistic units together in a character sequence using multiple attentions. This strategy is inspired by previous work that successfully employed multiple attentions in multi-task scenarios to estimate relationships between multiple types of knowledge. To the best of our knowledge, such a strategy has not been exploited in word segmentation. Thus, we introduce multiple attentions to word segmentation, which jointly consider representations at different granularity levels, thereby enabling more effective handling of possible segmentation alternatives and improving segmentation performance.

Moreover, most studies on Asian Languages, such as Japanese and Chinese languages, rely only on subwords or words. In contrast, the Thai language offers a unique opportunity for a broader exploration due to the presence of character clusters (CCs), which are indivisible units derived from predefined rules in the Thai writing system and have proven effective in Thai word segmentation. Given this distinct characteristic, we introduce a method, particularly for the Thai language, that leverages the joint use of CCs, subwords, and words with multiple attentions. Experimental results regarding this method demonstrates that applying word attention followed by smaller units, either CC or subword attention, effectively improves segmentation performance on BEST2010, TNHC, and VISTEC datasets, outperforming previous Thai word segmentation methods.

The second aspect aims to incorporate multi-granularity linguistic units through the use of lattices in character-based word segmentation. We propose a method, called Lattice ATTentive Encoding (LATTE), that effectively leverages possible segmentation alternatives based on multi-granularity linguistic units, including character and word units, using a lattice structure. Our experimental results regarding this method show improved segmentation performance on the BCCWJ, CTB6, and BEST2010 datasets for Japanese, Chinese, and Thai languages.