論文 / 著書情報
Article / Book Information

題目(和文)	タスクに応じた単語分割
Title(English)	Task-Oriented Word Segmentation
著者(和文)	平岡達也
Author(English)	Tatsuya Hiraoka
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11829号, 授与年月日:2022年3月26日, 学位の種別:課程博士, 審査員:岡崎 直観,徳永 健伸,篠田 浩一,宮崎 純,井上 中順
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11829号, Conferred date:2022/3/26, 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	情報工学系 知能情報コース	申請学位（専攻分野）： Academic Degree Requested	博士（工学） Doctor of
学生氏名： Student's Name	平岡 達也	指導教員（主）： Academic Supervisor(main)	岡崎 直観 教授
		指導教員（副）： Academic Supervisor(sub)	

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

Word segmentation or tokenization is a fundamental process in natural language processing (NLP). A sentence is split into small units such as words, subwords, or other tokens to process natural language on a computer for NLP tasks such as text classification. Because the downstream model is trained and evaluated with a tokenized sentence, the performance of the downstream model depends on the tokenization strategy. Therefore, exploring the proper tokenization method is a fundamental issue to improve NLP performance.

In general architectures of NLP, word segmentation or tokenization is considered a preprocessing task. In other words, sentences can be tokenized into tokens in advance of training the downstream model. This means that the tokenization strategy is not changed after preprocessing. However, recent studies have shown that the appropriate tokenization depends on the downstream task and model. This implies that a gap exists between tokenization as preprocessing and the training of the downstream model. In other words, the tokenization strategy must be determined in isolation from the downstream model, where the appropriate tokenization depends on the downstream task and the architecture of the downstream model. Determining the tokenization strategy without information about the downstream task and model is not recommended, even though information can be accessed when choosing the tokenization strategy. To bridge this gap, a novel method is proposed herein to train both the tokenization module and downstream model simultaneously. In contrast to the conventional tokenization method in NLP, the proposed method improves the tokenization strategy during the training of the downstream model and enables the tokenization module to generate a more appropriate tokenization for the downstream model, thereby improving the performance of the model.

This study introduces two approaches to optimize tokenization. The first approach embeds the tokenization module into the architecture of the downstream model and exploits the sentence representation calculated in the downstream model to select better tokenization during the training of the model. This first approach is specialized to the downstream model using sentence vectors to solve a task such as text classification. The second approach exploits loss values of the downstream model calculated to optimize the tokenization module. This approach is applicable to various downstream models, as it uses only loss values for the update, and this implies that it can be used with various NLP tasks, including generation tasks such as machine translations. Both approaches employ neural networks for the tokenization module, known as a neural unigram language model, and the downstream model and tokenization module are trained simultaneously as combined neural networks.

This study evaluates the proposed method on two famous NLP tasks, namely, text classification and machine translation, on multiple languages. For text classification, sentiment analysis in Chinese, Japanese, and English is employed for a task using a single sentence for the input. The rating and genre prediction tasks are also exploited using reviews on E-commerce services in Chinese, Japanese, and English. In addition, natural language inference in English is employed for a task using multiple inputs. For machine translation, seven language pairs are used, where one side of the translation pair is English, and the other side uses German, Vietnamese, Chinese, Arabic, French, Hungarian, and Romanian. The experimental results demonstrate that the proposed method improves the performance of the downstream model by optimizing tokenization on both text classification and machine translation as compared with the conventional tokenization strategy. The experimental results also show that the proposed method improves the performance of the downstream task even when the already trained downstream model is used and its trainable parameters are frozen. These results demonstrate that the proposed method can improve the downstream performance only by finding more appropriate tokenization for the downstream model. The experimental results on text classification demonstrate that the proposed method can be applied to various downstream models such as

classifiers with the self-attention mechanism, bi-directional long short-term memory (BiLSTM) encoders, and logistic regression. Finally, the results show that the proposed method are applicable to the downstream model, including BERT, which is a well-known large pre-trained language model.

Analysis of the acquired tokenization by the proposed method shows that the optimized tokenization differs depending on the downstream task and model. For example, the proposed method acquires different tokenizations for different text classification tasks even when the input text is the same. This study also provides the observation that the number of tokens in the acquired tokenization differs depending on the downstream tasks and languages. For example, the number of tokens in the tokenization for text classification is much greater than that for a target side corpus of machine translation. (759 words)