論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Efficient and Robust Methods for Korean Tokenization
著者(和文)	文翔煥
Author(English)	Moon Sangwhan
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第395号, 授与年月日:2025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,宮崎 純,村田 剛志,井上 中順
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第395号, Conferred date:2025/3/26, 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 (学術) Doctor of
学生氏名： Student's Name	文 翔煥	審査員主査： Chief Examiner	岡﨑 直観

要旨（英文 300 語程度）
Thesis Summary (approx.300 English Words)

This thesis examines two key challenges in Korean natural language processing: handling out-of-vocabulary tokens and efficient tokenization in multilingual models.

The research begins with a comprehensive review of recent developments in tokenization methods, particularly focusing on subword approaches like WordPiece, BPE, and SentencePiece. These methods have largely addressed tokenization challenges for alphabetic languages but face unique difficulties with character-diverse languages like Korean. The review examines both vocabulary-free approaches and methods specifically designed for handling CJK (Chinese, Japanese, Korean) languages.
The first major investigation addresses out-of-vocabulary (OOV) challenges in multilingual BERT. Through analysis of vocabulary distributions, the research demonstrates that Korean receives disproportionately small vocabulary allocations compared to other languages. Analysis revealed that while Wikipedia shows a CJK language allocation of 22%, multilingual BERT's vocabulary dedicates only 15% to these languages. This imbalance directly impacts model performance through information loss when tokenizing Korean text.

To address this challenge, the work introduces a post-training recovery method using surrogate token mapping. The approach maps OOV tokens to existing vocabulary items through three strategies: character distance matching, masked language model prediction, and unseen subword assignment. Character distance matching leverages the structure of Unicode CJK blocks, where adjacent codepoints often share semantic or phonetic properties. Masked language model prediction uses BERT's pre-trained knowledge to suggest contextually appropriate replacements. The unseen subword strategy repurposes tokens that never appear in downstream tasks. When combined with continued pre-training, this method improved accuracy by 1-2% across multiple downstream tasks, including sentiment analysis and question answering.

The research validates these improvements through controlled experiments with artificially degraded models. Models were systematically degraded by removing common words, rare words, or random vocabulary selections. Even under extreme conditions with 50% OOV rates, the method achieved 86.04% accuracy on the NSMC dataset, outperforming traditional baselines. This demonstrates the approach's effectiveness in recovering model performance without requiring complete retraining.
The second major contribution introduces Jamo Pair Encoding, a novel tokenization method for Korean that operates at the sub-character level. Rather than treating Korean syllable blocks as atomic units, the method decomposes them into constituent Jamo (phonetic characters). This reduces the required vocabulary size from over 11,000 characters to approximately 100 sub-characters while maintaining complete coverage of the Korean writing system. The method leverages unused Unicode code points as processing hints, ensuring compatibility with existing tokenizer implementations while preserving reconstruction capability.

The method offers two variants: an aligned approach using a fixed three-character grid, and an automaton method employing a flexible state machine. The aligned approach maintains a consistent three-character structure through explicit padding, simplifying reconstruction but potentially constraining subword learning. The automaton method implements a simplified variant of Korean IME architectures, eliminating bidirectional processing requirements and compound validation to optimize for the unidirectional nature of model generation. Both guarantee round-trip consistency between original text and tokenized forms, making them suitable for both understanding and generation tasks. Experimental results demonstrate significant reductions in vocabulary requirements while maintaining

or improving sequence length efficiency.
The thesis examines these methods through recently proposed evaluation frameworks for tokenization quality. This analysis includes metrics like corpus token count, fertility, compression ratios, and entropy measures. The results demonstrate that improvements in both robustness and efficiency are achievable through careful consideration of writing system characteristics. Particularly notable is the method's impact on sequence length efficiency - a critical factor in transformer-based architectures where computational complexity scales quadratically with sequence length.

While byte-level approaches have largely resolved OOV challenges in current large language models, the proposed methods maintain relevance for resource-constrained scenarios and for improving efficiency in multilingual model vocabulary allocation. The success of both approaches demonstrates that current limitations in processing character-diverse languages represent addressable challenges rather than fundamental constraints.

Finally, the methods are retrospectively evaluated with current state-of-the-art tokenizer quality evaluation methods. This analysis examines their validity alongside modern approaches like byte-level tokenization used in large-scale language models. The work opens several promising directions for future investigation, including integration with byte-level tokenization approaches, extension to other writing systems, and development of hybrid architectures. Most importantly, it shows that the seemingly competing goals of tokenization robustness and efficiency can be simultaneously improved through principled approaches to writing system analysis and vocabulary management.