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)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		文 翔煥	
		氏　名	職　名		氏　名	職　名
論文審査 審 査 員	主査	岡崎 直観	教授	審査員	井上 中順	准教授
	審査員	徳永 健伸	教授			
		宮崎 純	教授			
		村田 剛志	教授			

論文審査の要旨（2000 字程度）

本論文は、「Efficient and Robust Methods for Korean Tokenization」（韓国語のトークン化のための高効率かつ頑健な手法）と題し、英文 7 章から構成されている。具体的には、多言語モデルにおける未登録語（OOV; Out-Of-Vocabulary）問題の緩和と、新しいサブワード・トークン化手法を提案している。

第 1 章「Introduction」（序論）では、現代の自然言語処理において、多様な文字体系を持つ言語を処理する際の基礎的な課題を説明している。特に、韓国語は中国語や日本語と一緒に扱われることが多いが、その表音文字システムと現在のユニコード表現の制約により、独自の課題を抱えていることを指摘している。さらに、大規模言語モデルの学習では、しばしば語彙の不均衡な割り当てを引き起こし、CJK 言語（中国語・日本語・韓国語）が不当に少ない語彙配分を受けることを説明している。この問題意識のもと、本論文の 2 つの主要な貢献である、学習済みのモデルを活用しながら OOV 問題へ対処する手法と、韓国語テキストのためのより効率的なトークン化手法を提案する。

第 2 章「Backgronud」（背景）では、自然言語処理の発展の流れを、ルールベースの手法から統計的手法、そして現代のニューラルアーキテクチャまで概観する。また、トークン化、埋め込み表現、転移学習といった基本概念について詳細に説明している。CJK 言語の中でも、特に韓国語の形態音節的な構造と、そのテキスト処理における影響に焦点を当てる。さらに、トークン化の品質と効率を評価するための主要な指標を紹介し、本論文の貢献を理解するための基礎を提供している。

第 3 章「Related Word」（関連研究）では、トークン化の課題に対する既存のアプローチを包括的に検討している。語彙非依存の手法、語彙適応戦略、各種の OOV 緩和手法について調査・議論している。また、WordPiece、バイト対符号化（BPE; Byte-Pair-Encoding)、SentencePiece などの異なるサブワード・トークン化手法を比較し、文字体系の異なる多様な言語を処理する際の強みと限界を分析している。特に、サブ・キャラクター（文字未満のレベルでのトークン化）アーキテクチャや情報損失を伴う手法について詳しく議論し、本論文の貢献を多言語処理研究の文脈に位置付けている。

第 4 章「The Effects and Mitigation of Out-of-vocabulary in Universal Language Models」（OOV の影響と緩和策）では、多言語言語モデルにおける OOV の影響を体系的に調査している。語彙分布の分析を通じて、韓国語は多言語 BERT において期待される語彙配分のわずか 20%しか割り当てられていないことを示した。これに対処するため、本章では、OOV 緩和戦略の新しい手法として、代替トークンへの対応付けを提案する。この手法は、文字距離マッチング、マスク付き言語モデル予測、および未知サブワード割り当ての 3 つの手法を組み合わせるものである。広範な実験を通じて、この手法がモデルの再学習を行わずに OOV の影響を緩和できることを示し、例えば、人工的に 50%の OOV が存在する条件を作り出した場合でも、NSMC データセットで 86.04%の精度を達成したことを報告している。

第 5 章「Jamo Pair Encoding: Subcharacter-tokenization of Korean」（Jamo ペアエンコーディング：韓国語のサブキャラクタートークン化）では、韓国語のテキスト処理を革新する新規のサブキャラクタートークン化手法を提案している。本手法では、韓国語の音節ブロックを単一の単位として扱うのではなく、それを構成する Jamo（音素文字）に分解することで、必要な語彙量を 11,172 文字から約 100 のサブ・キャラクターに削減している。固定された 3 文字のグリッドを使用する手法と、状態機械を用いたオートマトン手法の 2 つを提案している。どちらの手法も、情報の損失がないため、往復の変換が保証されているが、韓国語のトークン化の効率を大幅に向上させる。複数のフレームワークにわたる包括的な評価を行い、トークン化品質と系列長の効率性の向上を実証した。

第6章「Discussion」（考察）では、最新のトークン化品質評価フレームワークを用いて、提案手法の有効性を検証している。コーパスのトークン数、受容度（fertility）、圧縮率、エントロピー測定といった指標を用いて、各手法の長所と短所を詳しく分析している。特に、近年の大規模言語モデルにおけるバイト単位のトークン化技術の発展を踏まえ、提案手法の統合可能性について検討している。さらに、他の文字体系への応用やハイブリッドアーキテクチャの開発など、今後の研究の有望な方向性を検討している。

第7章「Conclusion」（結論）では、本論文の主要な貢献を総括し、文字体系の特性と語彙の管理方法を慎重に考慮することで、ロバスト性と効率性の両方において大幅な改善が可能であることを述べている。また、本研究が多言語モデルの開発に与える広範な影響について考察し、今後の研究におけるいくつかの有望な方向性を提示している。さらに、これらの知見を急速に発展する自然言語処理分野の中で位置付け、多言語処理研究の今後の指針を述べている。

トークン化は、大規模言語モデルの入口と出口で用いられているが、大規模言語モデルの性能や挙動をトークン化のレベルで分析することは稀である。ただ、大規模言語モデルを実際に利用するときには、学習データに無かった単語や文字（OOV）の取り扱いに悩まされるので、OOV に遭遇した時にも頑健に対処する方法や、文字の構成を考慮して OOV が起こらないようにしつつ、語彙をコンパクトにする手法は有用性が高い。また、本論文で提案されている手法は韓国語のみならず、その他の言語や、画像処理や音声処理など、基盤モデルへの入出力においてトークン化が行われる分野への展開も考えられる。よって，本論文は工学の発展に寄与し，博士（学術）の学位論文として十分価値があるものと認める.