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-INTRThodsaporn
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)	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		CHAY-INTR Thodsaporn	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	奥村学	教授	審査員	船越孝太郎	准教授
	審査員	熊澤 逸夫	教授			
		中山実	教授			
		篠崎隆宏	准教授			

論文審査の要旨（2000 字程度）

本論文は「Incorporating Multi-granularity Linguistic Units in Character-based Word Segmentation」と題し，英文全5章より構成されている．

第1章「Introduction」では，本研究の背景を説明するとともに，本論文の目的を述べている．まず，日本語，中国語，タイ語などアジア言語では，入力文字列を単語列に分割する単語分割が言語理解のための最初の処理として重要であること，文字ベースの手法は，単語ベースの手法と比べ，語彙に含まれない単語への対処などで有効であること，従来の文字ベースの手法は，subwordや単語などの言語単位を解析に利用することに成功しているが，依然様々な言語単位を十分に活用しきれていないため，性能改善の余地があることを述べている．そして，本研究の2つの目的についてその概要を説明している．1つ目の研究では，複数のアテンション機構を用いて様々な粒度の言語単位を同時に導入すること，特にタイ語では，独自の言語単位であるcharacter clusters (CCs)も導入することで単語分割の性能を向上させることが，2つ目の研究では，ラティスを用いて様々な粒度の言語単位を同時に導入することで単語分割の性能を向上させることが，それぞれ目的であることを述べている．

第2章「Related Work」では，本研究に関連する，文字ベース，単語ベースの単語分割手法，様々な言語単位を導入した文字ベースの単語分割手法について説明している．

第3章「Incorporating Multi-granularity Linguistic Units with Multiple Attentions」ではまず，提案モデルが，文字ベースのBiLSTM-CRF構造のニューラルモデルに，単語，CCs，subwordという，様々な粒度の言語単位に対応する複数のアテンション機構を同時に導入したものであることを述べている．そして，提案手法の評価実験では，タイ語の3つの評価データセットを用いて，従来の最高性能の手法を上回る性能を発揮することを確認している．

第4章「Incorporating Multi-granularity Linguistic Units through the Use of Lattices」では，文字，単語単位での単語分割結果である複数のラティスを用いるLattice ATTentive Encoding (LATTE)という手法を提案し説明している．また，そのための具体的なモデルとして，事前学習済みモデルBERT，ラティス構造を表現したグラフニューラルネットワーク，アテンション機構，CRF層からなるモデルを示している．提案手法の評価実験では，日本語，中国語，タイ語のデータセットを用いて，従来の手法を上回る性能を発揮することを確認している．

第5章「Conclusion and Future Work」では，本研究の結論と，第3章，第4章で提案した手法に対する課題と今後の展望について述べている．

以上を要するに，本論文は，日本語，中国語，タイ語などにおける文字ベースの単語分割手法の高度化のため，2つの方向性で新しい手法を提案し，その有効性を検証しており，工学上貢献するところが大きい．よって博士（工学）の学位を授与するに十分な価値を持つものと認められる．