論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Information Extraction Beyond Sentence Boundary
著者(和文)	MAYoumi
Author(English)	Youmi Ma
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12916号, 授与年月日:2024年9月20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,金崎 朝子
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12916号, Conferred date:2024/9/20, 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		MA Youmi	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	岡崎 直観	教授	審査員	金崎 朝子	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，「Information Extraction Beyond Sentence Boundary」と題し，英文6章から構成されている．自然言語処理において，テキストに記述された知識を自動抽出し，表やグラフなどの構造化データに整形することを情報抽出（Information Extraction）と呼ぶ．構造化データでは，知識が事前に定義された形で格納されるため，検索がしやすく，人間の意思決定やコンピュータによる質問応答などに役立つ．テキストから構造化データを自動構築する方法を探求すること，すなわち，高精度の情報抽出器を構築することは，自然言語処理分野の重要な課題の一つである．情報抽出のうち，関係知識を（subject，relation，object）の三つ組として抽出する設定を関係抽出と呼ぶ．多くの関係抽出は文に閉じた設定で行われてきたが，関係知識は文の境界をまたがって書かれることがある．文境界をまたがって記述された関係知識もテキストから抽出するタスク設定として，文書レベル関係抽出がある．本研究では，文書レベル関係抽出の実用上の課題である教師信号の不足の問題に取り組む．まず，文書レベル関係抽出器を学習するための人手ラベル付けデータを十分に活用できていなかったという先行研究の課題に対し，根拠情報を教師信号として活用できる関係抽出器の構築法を提案した．次に，データセットの新規構築コストが高いという課題に対し，機械翻訳を活用したデータセットの効率的な構築法を提案した．

第１章「Introduction」では，自然言語処理分野における情報抽出の位置づけ，情報抽出の歴史，文書レベル関係抽出が提案された経緯について説明した上で，文書レベル関係抽出の課題を述べている．その後，本論文により提案された課題の解決策の概要を紹介し，本研究がもたらす貢献を説明している．最後に，本論文の構成と章ごとの要約をまとめている．

第２章「Background Knowledge」では，本論文の基礎である自然言語処理技術について説明している．具体的には，自然言語処理に極めて重要な深層学習モデルであるTransformer，およびその核心である注意機構と，言語モデルと大規模言語モデルの紹介から構成されている．

第３章「Preliminaries and Related Work」では，本論文を理解するための予備知識と既存研究を詳説している．具体的には，文書レベル関係抽出の定式化と本論文で用いるデータセット，およびその構築方法を紹介し，第４章で述べる関係抽出器構築法の既存研究と，第５章で述べるデータセット構築法の既存研究を，それぞれ説明している．本論文で用いるデータセットでは，関係ラベル以外にも，根拠ラベルが人手により付与されており，関係推定に必要な文の集合を示している．関連研究の説明では，提案手法のベースラインとなる技術と，比較対象となる技術の両方を詳述している．また，既存研究では，根拠を認識するために，関係抽出器とは独立の根拠認識器を学習していることを説明している．

第４章「Model Construction: DREEAM」では，根拠情報を用いて関係抽出器のエンコード過程を誘導することにより，根拠の教師信号を関係抽出器の学習に統合し，根拠認識器を撤廃する手法を提案している．本研究の狙いは，根拠情報を教師信号として活用し，より高精度な関係抽出器を構築することである．具体的には，人手で付与された根拠ラベルから導出された正解根拠分布を，関係抽出器のエンコーダにおける注意機構の教師信号とし，根拠に高い重みを付与するように誘導する．その後，根拠ラベルのないデータに対して疑似根拠を付与し，関係抽出器の学習に用いる手法も提案している．提案手法により構築された関係抽出器は，関係抽出と根拠認識両方で高い性能を示し，2024年7月現在でも，複数のベンチマークで世界最高精度を維持している．また，注意機構の誘導により，関係抽出器の解釈性向上にも貢献できたことを分析により示している．これにより，人手で付与された根拠ラベルを関係抽出器の学習に活用することができ，提案手法は学習データの利用率向上に貢献していることが実証された．

第５章「Dataset Construction: JacRED」では，機械翻訳をデータセット構築作業に組み込むこ

とにより，文書レベル関係抽出データセットの新規構築における人手作業コストを削減する手法を提案している．この狙いは，なるべく人手を用いずに，英語以外の言語（ここでは，日本語を対象とする）で文書レベル関係抽出データセットを構築することである．そのために，まず機械翻訳を用いて英語データセットを日本語に翻訳する全自動構築法を試み，得られた翻訳データセットで関係抽出器を学習し，その精度を評価している．その結果，翻訳データセットで学習した関係抽出器は精度が低く，実用に耐えないことが明らかになった．次に，翻訳データセットで学習した関係抽出器から予測された関係知識事例を人手ラベル付けの起点とし，人間作業者が予測事例を添削することでラベル付けを行う手法を提案している．既存研究と比較した結果，提案手法は人手作業コストを半減しつつ，良質なデータセットを構築できることが実証された．また，提案手法により，日本語初の文書レベル関係抽出データセットが構築された．

第６章「Conclusion」では，本論文のまとめと今後の展望を述べている．

本論文では，文書レベル関係抽出における二つの課題に対して解決策を提案した．本研究により得られた知見は，文書レベル関係抽出だけでなく，自然言語処理の幅広い分野に有用である．まず，根拠ラベルを用いて関係抽出器の注意箇所を誘導した結果，その精度向上に寄与できたことから，人間の振る舞いを AI に模倣させるアプローチの成功事例と言える．また，自然言語処理分野においてデータセットの他言語適用は自動翻訳に留まることが多いのに対し，本研究は自動翻訳によるデータセット構築の欠陥を分析し，それを補う方法まで提案している．さらに，本研究の成果は，構造化データの自動構築精度を向上し，非構造化データの管理やコンピュータによる質問応答などの応用にも繋がる．よって，本論文は工学の発展に寄与し，博士（工学）の学位論文として十分価値があるものと認める．