日本学術振興会に紙媒体で提出する必要はありません。

様　式　Ｆ－７－１

科学研究費助成事業（学術研究助成基金助成金）実施状況報告書（研究実施状況報告書）（令和元年度）

機関番号	１４６０３

所属研究機関名称		奈良先端科学技術大学院大学
研究代表者	部局	先端科学技術研究科
	職	助教
	氏名	進藤　裕之

１．研究種目名　若手研究　　２．課題番号　18K18109

３．研究課題名　科学技術論文からの統合的な構造解析に関する研究

４．補助事業期間　平成３０年度～令和２年度

５．研究実績の概要

科学技術論文の出版数は加速度的に増大しており，個人が必要な論文を検索し，その全てに目を通すことは極めて困難な状況である．科学技術論文は，概要，本文，数式，図表などで構成される構造化文章であり，PDFを構造化する技術が確立されれば，従来では困難であった論文の高度な検索や情報抽出が可能となる．
そこで本研究では，論文を解析してXMLなどの構造化フォーマットへ変換するためのモデルおよびアルゴリズム構築を目指している．
平成３１年度は，論文に含まれる表，数式，本文それぞれの構造化モデルを相互に組み合わせることにより，実際の論文を解析するモデルおよびアルゴリズムの構築を行った．具体的には，PDFから取得した文字とその位置情報に基づいて，機械学習により文字列にタグを付与することでセクションや段落などの構造を決定する．また，セクションや段落同士の関係性も機械学習により推定することで，PDF全体の木構造を決定することができる．
実際の専門分野への応用として，材料科学分野の文献を対象として，論文の構造化に関する実験を行った．材料科学文献では，物質名や物性値に関する情報の多くが表に含まれるため，表の構造化を正しく行うことが情報抽出にとって重要である．実験の結果，本文に関しては，非常に高い性能で構造化を実現することができた，また，おおよそ90%の表に関しては正しく解析を行うことができたが，行や列が連結された複雑な表については解析誤りが多く含まれることがわかった．今後の課題として，複雑な表の解析性能を向上させるためのモデル改善やアルゴリズム改善を実施する必要がある．

６．キーワード

論文解析　構文解析　PDF

７．現在までの進捗状況

区分　(2) おおむね順調に進展している。

理由

おおむね順調に進展している．平成３１年度は，予定通り，PDFを構造化するためのモデル構築およびアルゴリズム考案を進捗通りに進めることができた．次年度は，構造解析の更なる性能向上と，いくつかの専門分野の論文へ本技術を適用して評価することを中心として作業を進める予定である．

日本学術振興会に紙媒体で提出する必要はありません。

８．今後の研究の推進方策

令和２年度は，平成３１年度の研究で明らかになった構造化性能の向上，特に複雑な表の解析に関して，モデルおよびアルゴリズムの両側面から改善を目指す．また，様々な専門分野（バイオロジー，材料科学など）で適用しても性能が下がらないように，分野適応や教師無し学習の知見を取り込んでいく予定である．また，PDFを構造化および情報抽出について，バイオロジーや材料科学分野の研究者と協調し，どのような構造や情報が抽出できると実際に役に立つかということを考慮して技術開発に反映させていく．

９．次年度使用が生じた理由と使用計画

データ作成に関する謝金・外注費に関して，一部を次年度に回す方が効率的に研究開発を進めることができるため．

１０．研究発表（令和元年度の研究成果）

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Kato Akihiko、Shindo Hiroyuki、Matsumoto Yuji	４．巻 26
２．論文標題 Construction and Analysis of Multiword Expression-aware Dependency Corpus	５．発行年 2019年
３．雑誌名 Journal of Natural Language Processing	６．最初と最後の頁 663～688
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.5715/jnlp.26.663	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Liu, J., Shindo, H. and Matsumoto, Y	４．巻 67
２．論文標題 Development of a computer-assisted Japanese functional expression learning system for Chinese-speaking learners	５．発行年 2019年
３．雑誌名 Educational Technology Research and Development	６．最初と最後の頁 1307～1331
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1007/s11423-019-09669-0	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Hiroki Teranishi, Hiroyuki Shindo, Yuji Matsumoto
２．発表標題 Decomposed Local Models for Coordinate Structure Parsing
３．学会等名 In Proceedings of NAACL（国際学会）
４．発表年 2019年

１．発表者名 Tatsuya Hiraoka, Hiroyuki Shindo, Yuji Matsumoto
２．発表標題 Stochastic Tokenization with a Language Model for Neural Text Classification
３．学会等名 In Proceedings of ACL, 2019（国際学会）
４．発表年 2019年

１．発表者名 Van-Hien Tran, Hiroyuki Shindo, Yuji Matsumoto
２．発表標題 Relation Classification Using Segment-Level Attention-based CNN and Dependency-based RNN
３．学会等名 In Proceedings of NAACL, 2019（国際学会）
４．発表年 2019年

〔図書〕　計0件

１１．研究成果による産業財産権の出願・取得状況

計0件（うち出願0件／うち取得0件）

１２．科研費を使用して開催した国際研究集会

計0件

１３．本研究に関連して実施した国際共同研究の実施状況

―

１４．備考

―