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２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３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を構造化でき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．
令和２年度は，材料科学分野の論文を対象として，図表，数式，本文の構造化を行うアルゴリズムの改善および評価を行った．前年度で問題となっていた表の連結セルの解析誤りについては，教師データの開発およびアルゴリズムの改善によって解析誤りを緩和させられることを検証した．
また，モデルが広範囲な文脈を捉えられるために，自己注意型の機構を取り入れ，文書のレイアウトや潜在的な構造を特徴量として上手く抽出できることを確認した．また，従来のボトムアップ型の解析と比較して，トップダウン型のアルゴリズムによる解析の方が同等の性能かつ低コストで動作することを検証した．今後の課題として，シミュレーション技術による多様なデータ拡張を導入することにより，モデルの性能向上を実現することが挙げられる．
上記と並行して，構造情報の教師データを作成するためのツール開発も行っている．材料科学やバイオロジーなど，各分野の専門家が直感的に文献へアノテーションを行い，情報抽出を行うことをサポートできる環境を構築できた．

６．キーワード

論文解析　PDF　構文解析

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．令和２年度は，材料科学文献を対象として，PDFを構造化するためのモデルやアルゴリズムの検証を進めることができた．また，従来より問題となっていた表の連結セルの解析誤りに対して，広い文脈を捉えるモデルや教師データの増強によって性能改善を実現することができた．今後の課題として，図表，数式，本文などの個別の解析に留まらず，それらの多様な情報を含む文書全体を統一的に解析するモデル構築を進める必要がある．

【研究代表者・所属研究機関控】

日本学術振興会に紙媒体で提出する必要はありません。

1版

８．今後の研究の推進方策

令和３年度は，材料科学文献だけでなく，バイオロジーや医療などの専門分野へ対象を拡大して，これまでのモデルが適用可能であるか検証を行う．その際に，あらゆる専門分野で教師データを大量に用意することは非現実的であるため，シミュレーションによるデータ増強や転移学習を併用することにより，少量データでも頑健に動作する仕組みの構築を目指す．材料科学分野に関しては，情報抽出の形式やシステムのユーザーインターフェースを改善し，専門家が教師データを作成しやすい環境構築を行っていく．特に，テキスト中に出現する並列関係や照応関係をどのようにアノテーションするかということが課題で，専門家がなるべく直感的に教師データを作成できるようにシステムがサポートできれば，各専門分野へ本研究のモデルを適応する際に大幅な効率化が見込める．

９．次年度使用が生じた理由と使用計画

教師データ作成やツール開発に関する謝金・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２年度の研究成果）

〔雑誌論文〕　計0件

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　2件）

１．発表者名
Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, Yuji Matsumoto

２．発表標題
LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention

３．学会等名
In Proceedings of EMNLP（国際学会）

４．発表年
2020年

１．発表者名
Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, Yuji Matsumoto

２．発表標題
Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia

３．学会等名
In Proceedings of EMNLP (demo)（国際学会）

４．発表年
2020年

１．発表者名 山口泰弘，進藤裕之，渡辺太郎
２．発表標題 ラベルの不均衡を考慮したEnd-to-End情報抽出モデルの学習
３．学会等名 言語処理学会第27回年次大会(NLP2021)
４．発表年 2021年

１．発表者名 平野颯，野村航，進藤裕之，渡辺太郎
２．発表標題 遺伝子二重欠失研究のための関連論文検索手法
３．学会等名 言語処理学会第27回年次大会(NLP2021)
４．発表年 2021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

―