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）（平成３０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

科学技術論文を対象とした情報抽出では，「ある特定のデータやパラメータを用いて実験を行った論文」といった高度な検索を行うことが難しい．これは，論文データから，セクション，段落，数式，図表などの基本的な構造を解析できていないことが根本的な原因の一つである．本研究では，様々な分野の論文を構造化する技術の確立を目指す．
平成３０年度は，化学・材料分野，情報分野，バイオロジー分野の３つの専門分野の論文データを収集し，それらに共通の構造について調査を行った．また，それらを集約し，一貫性のあるXMLの仕様を定義する作業を行った．主に，タイトル，セクション，数式，段落，図，表といった基本要素によって，分野によらない論文フォーマットを定義することが可能であることがわかった．また，JATS（既存の科学技術論文フォーマット）は，細かいタグや定義が曖昧なタグが多く，実際の論文では使用されていないものも多い．そこで，JATSを大幅に簡略化したタグ仕様を定義し，それに基づいてPDFを構造化することとした．
次に，上記のXML仕様に基づく学習データ（PDFをXML化するための学習データ）を構築した．具体的には，PubMedのJATS形式の論文とPDFのペアを大量に収集し，それらを変換して，XMLとPDFと対応付ける作業を行った．単純な文字列マッチングでは上手く対応が取れないケースがあり，いくつかの近似文字列マッチングアルゴリズムを考案し，それに基づいて評価実験を行った．

６．キーワード

知識獲得　情報抽出　科学技術論文　構文解析　意味解析　自然言語処理

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３０年度は，予定通り，データ収集や仕様定義などの準備作業・環境構築作業が中心であり，来年度からは本格的に技術的な作業へ取り掛かることができる．

８．今後の研究の推進方策

平成３１年度は，平成３０年度に構築した学習データを用いて，PDFをXML化する技術の確立を目指す．特に，図表の位置認識や段落の認識が重要であり，この部分に関して変換技術の性能評価を行う予定である．まずは，個々の要素（図，表，数式など）ごとに別々のモデルを考案し，最終的にはそれらを統合して一つのPDF変換プログラムとすることを計画している．また，実際の化学・材料・バイオ研究者と協調し，XML化された論文からどういった情報を抽出するかについて議論し，データフォーマットの定義やアノテーションガイドラインの作成に目途をつける予定である．

９．次年度使用が生じた理由と使用計画

使用物品を予定よりも安価で調達することができたため，若干の次年度使用額が生じた．こちらは次年度の物品費として使用する予定である．

１０．研究発表（平成３０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　1件）

１．著者名 Phan Duc-Anh、Matsumoto Yuji、Shindo Hiroyuki	４．巻 1
２．論文標題 Autoencoder for Semisupervised Multiple Emotion Detection of Conversation Transcripts	５．発行年 2018年
３．雑誌名 IEEE Transactions on Affective Computing	６．最初と最後の頁 1～11
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TAFFC.2018.2885304	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Teranishi Hiroki、Shindo Hiroyuki、Matsumoto Yuji	４．巻 25
２．論文標題 Similarity and Replaceability Feature Representations of Word Sequences for Identifying Coordination Boundaries	５．発行年 2018年
３．雑誌名 Journal of Natural Language Processing	６．最初と最後の頁 441～462
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.25.441	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5件（うち招待講演　0件／うち国際学会　5件）

1．発表者名
Hiroyuki Oka, Hiroyuki Shindo, Keisuke Goto, Yuji Matsumoto, Atsushi Yoshizawa, Isao Kuwajima and Masashi Ishii

2．発表標題
Automatic extraction of polymer data from tables in xml

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

4．発表年
2018年

1．発表者名
Keisuke Goto, Hiroyuki Shindo and Yuji Matsumoto

2．発表標題
Line Detection Considering Spatial Context for Reading Line Charts

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

4．発表年
2018年

1．発表者名
Shuhei Kondo, Yuji Matsumoto and Hiroyuki Shindo

2．発表標題
Translating Chemical Substance Names using Attentional Encoder-Decoder

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

4．発表年
2018年

1．発表者名
Hiroki Ouchi, Hiroyuki Shindo and Yuji Matsumoto

2．発表標題
A Span Selection Model for Semantic Role Labeling

3．学会等名
In Proceedings of EMNLP, 2018（国際学会）

4．発表年
2018年

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．発表者名 Ikuya Yamada and Hiroyuki Shindo
２．発表標題 Representation Learning of Entities and Documents from Knowledge Base Descriptions
３．学会等名 In Proceedings of COLING, 2018（国際学会）
４．発表年 2018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

―