様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 ２ 日現在

機関番号：１４６０３
研究種目：若手研究
研究期間：2018～2021
課題番号：１８Ｋ１８１０９

研究課題名（和文）科学技術論文からの統合的な構造解析に関する研究

研究課題名（英文）Research on Integrated Structural Parsing from Scientific Literature

研究代表者
進藤　裕之（Hiroyuki, Shindo）

奈良先端科学技術大学院大学・データ駆動型サイエンス創造センター・特任准教授

研究者番号：２０７３４７８４

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：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困難である．そこで本研究では，論文に含まれる本文，数式，図表などの領域や構造を解析し，XMLなどの構造化フォーマットへ自動変換するためのモデルおよびアルゴリズム構築を行った．特に，材料科学の文献を主なターゲットとして，図表や数式の領域は画像処理によって検出し，本文や表の構造は自然言語処理の構造解析技術を援用した統合的な解析手法を実現した．また，モデルの訓練や評価に必要な図表領域のデータセットや，本文および表の構造に関するデータセットなどのリソースを構築し，論文の統合的な構造解析を行う技術を確立した．

研究成果の学術的意義や社会的意義
本研究により，PDF形式の論文データを入力として，図表，数式，段落などのオブジェクトを抽出することや，表の内部構造（ヘッダや行列）を取得することができるようになった．そのため，ある分野における論文の実験データを網羅的に収集することや，図表に記述されている情報の細かい分析や検索が可能になると考えられる．また，本技術を用いて様々な分野の論文を構造化して知識データベースを構築し，ユーザーが閲覧できるようなサービスの実現も可能となる．

研究成果の概要（英文）：The number of published scientific papers is increasing at an accelerating rate, and it is difficult for individuals to search and read all the necessary papers.
In this study, we developed a model to automatically detect the objects such as figures and tables, and analyze the structure of the text and tables in a paper to convert them into structured formats such as XML. Our integrated parser mainly targets materials science literature, using image processing to detect the regions of figures and tables, and natural language processing to analyze the structures of text and tables. In addition, we developed resources for training and evaluating our model such as datasets for the region of tables and figures, as well as the structure of the text and tables in a paper.

研究分野：自然言語処理，画像処理

キーワード：論文解析　自然言語処理　構文解析　オブジェクト検出

１．研究開始当初の背景

科学技術論文の出版数は加速度的に増大している．そのため，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になっており，この状況は今後益々深刻なものとなる．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された文章である．本文は，さらに背景，手法，結果といったセクションや段落構造を持ち，数式や表もそれぞれ独自の文法構造を持つ．論文から有用な知識を自動抽出するためには，上記の構造を解析した上で，手法，実験設定，結果などの具体的な情報を精度良く同定する必要がある．

しかしながら， Web 上で配布される PDF 形式の論文データは，「文字や図形を画面上のどこへ表示するか？」というレイアウトの保持を目的としており，論文の持つ構造情報は失われている． PDF から単にテキストのみを抽出することや，タイトルや著者などの限定的な情報を取り出すことは現時点でも可能であるが，特定の書式やスタイルに依存せず，数式や表などを含む論文の完全な構造を解析できる技術は未だに確立されていない．これが，論文中の表や数式も含めた高度な情報検索を実現するための大きなボトルネックとなっている．

そこで，本研究では，以下の二点の問いを明らかとするための研究を行う．

1. 分野横断的な科学技術論文の共通構造（背景，目的，結果などのセクションや，数式，表など）をどのように形式的に定義するか．
2. PDF のような構造情報を持たない論文データから，上記の論文構造をどのように精度良く解析するか．

自然言語処理の観点からは，数式，表などの異種構造を含むテキストをどのように構造解析するかという問題はこれまでにあまり扱われておらず，論文データ特有の問題と言える．申請者は，これまでにテキストの構文解析（文法的な構造解析）に取り組んできた．本研究ではこれをさらに発展させ，論文データからの知識抽出の基盤となる技術の確立を目指す．

図１　PDF に含まれる文字や図形と座標情報の例

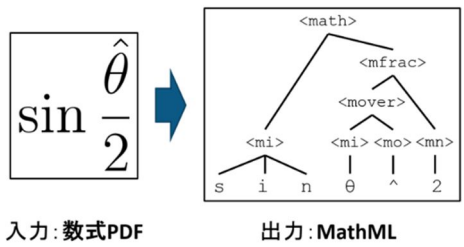


図２　PDF 中の数式を構造解析した例

２．研究の目的

本研究は，分野横断的な科学技術論文の構造を形式的に定義し，PDF のような構造情報を持たない論文データから，その構造を統計的に推定する手法の確立を目的とする．具体的には，

① 幅広い分野の論文を網羅的に調査し，段落，セクション，数式，表といった分野に依らない基本的な構造を統計的に明らかにして，それを XML 形式として定義する．学術論文のために，JATS と呼ばれる XML 形式が既に標準化されているが，これは XML から PDF を生成するためのフォーマットであり冗長性が高い．そのため，本研究のような PDF から XML を導出する目的とは異なる．
② 上記で定義した論文の構造を解析するための統計モデルおよびアルゴリズムを考案し，幅広い分野の論文に頑健で，かつ高精度な手法を確立する．
③ 材料科学とバイオロジー分野の論文を対象として，本研究で得られた構造化テキストから情報抽出を行い，本研究の有用性を評価する．
④ 本手法によって幅広い分野の構造化論文コーパスを構築し，論文からの知識抽出のための言語資源として広く公開する．

３．研究の方法

本研究では，分野に依らず，論文全体の構造（本文，数式，表の内部構造含む）を高精度に解析する統計的手法を確立し，成果をツールおよび言語資源として公開する．具体的には，以下の項目を実施する．

（１）分野共通の論文構造の形式化

まず，あらゆる分野の論文データを収集し，それらに共通の構造を統計的に明らかにする必要がある．また，それらを集約し，一貫性のある XML の仕様として定義する．

（２）論文 PDF と XML のデータ収集と自動対応付け

次に，上記の XML 仕様に基づく学習データを構築する．これは，論文 PDF と LaTex ソースのペアが公開されている arXiv などの文献データベースから論文を収集し，ソースと PDF を対応付けすることによって得られる．PDF と LaTex をどのように対応付けるかというアルゴリズムは自明でないため，手法を考案し，評価を行う．申請者は，既に arXiv から 100 万件程度の文献を収集済である．

（３）数式，表，本文の個別モデルの構築

上記の学習データを用いて，数式や表，本文などに特化した構造解析モデルの考案と実装を行う．これは，従来の自然言語処理における構文解析技術をベースラインとして，さらにモデルやアルゴリズムを工夫することにより高精度化を図る．数式に関しては，既に申請者らのグループによって構造解析手法が提案されており，まずはそれらを表や本文へ適用するところから始め，個別のモデルごとに性能向上を目指す．

（４）大域的な統合モデルの考案と性能評価

実際の論文では，本文，表，数式など様々な要素が独自の文法構造を持ち，それらが互いに組み合わさっているため，３で構築した数式や表の各モデルを統合して論文全体の解析を行うためのモデルとアルゴリズムの提案を行う．

（５）材料科学文献による知識抽出

材料科学分野の文献の構造解析を行い，構造化テキストから，材料と物性値の関係を抽出する．これは，申請者らのグループがこれまでに開発した学習データおよび情報抽出（関係抽出）手法を用いる．実際に各分野の研究者に抽出された情報の精度や再現率を評価してもらい，現実的な観点から構造解析手法の必要性能を評価する．

４．研究成果

（１）PDF 論文の構造解析

PDF 形式の論文を入力として，ヘッダー，段落，セクション，数式，図表の構造を解析するモデルを構築した．図表の領域については，深層学習に基づく画像処理によってオブジェクトを検出するモデルとなっている．それ以外の要素については，PDF から取得されるテキストの自然言語解析によって構造化を行うモデルを構築した．

モデルの訓練に用いるデータセットは，入手の容易性から，まずは材料科学以外の分野から大量に収集された PDF データを用いて事前学習を行った．次に，材料科学の文献について図表領域のアノテーションを行うためのガイドラインを整備し，人手で約 3000 件のデータセットを構築した．図表以外の要素，具体的にはヘッダー，段落，セクション，数式についても同様に，アノテーションガイドラインを整備して人手で約 1000 ページ分のデータセットを構築した．図表の解析の概要を図３に示す．

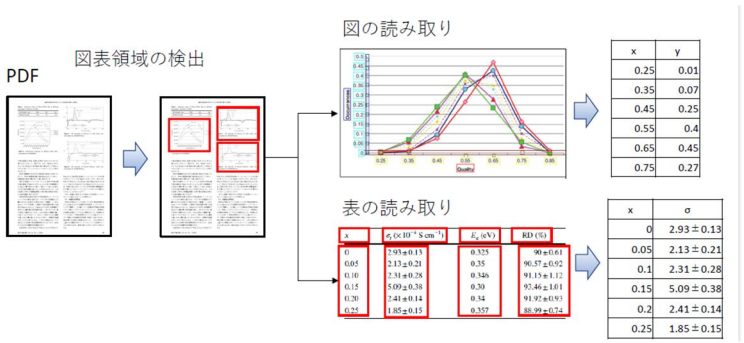


図３　PDF 論文を解析して図表の読み取る処理

実験結果を図４に示す．ヘッダー，段落，セクション，数式，図の要素については，0.9 以上の精度と再現率で解析することができた．表については，他の要素と比較してやや性能が低いが，一つの原因として，材料科学文献では，図と比較して表の出現頻度が低いため，学習データの量が十分でなかったと考えられる．ヘッダーについては，出現位置がページの上下に固定されているため，エラー無しで解析を行うことができた．

	header	段落	セクション	数式	図	表
precision	1.0	0.96	0.99	1.0	0.90	0.83
recall	1.0	0.98	0.96	0.91	0.91	0.85
F-value	1.0	0.97	0.97	0.95	0.90	0.84

図４　論文の構造解析の実験結果

（２）材料科学文献の表の意味解析

表については，構造を解析した後，セルにどのような内容が書いてあるかを解析し，中身を読み取ることが必要である．材料科学文献では，主に物質と物性値の関係が表として記述されているため，表のヘッダーセルを同定し，ヘッダーに書かれている情報や，コンテンツセルに書かれている数値を読み取ることが必要となる．このような解析を意味解析と呼ぶ．本研究では，高分子に関する材料科学文献を対象として，表の構造を（１）のモデルで解析済であるという前提のもと，表内部のセルの意味解析を行った．図５に読み取り対象となるヘッダーセルの例を示す．また，図６は，意味解析の入力と出力の例を示す．ここでは，“ ΔHi ”がエンタルピー変化を表す物性名で，“ 10^2 mJ g^{-1} ”が単位を表している．これを解析して，図６の下にあるような構造化データへ変換することが目的である．

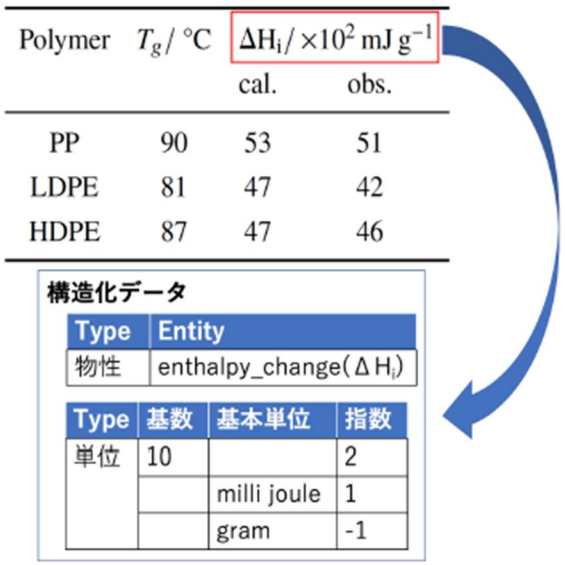
Table 1. Synthesis of porous resin monoliths[a]

Run No.	Feed[b]					Porous resin monoliths			
	Styrene (mmol)	CMS[c] (mmol)	DVB[d] (mmol)	AIBN[e] (mmol)	SMO[f] (mmol)	Yield (%)	Pore volume[g] (mL/g)	Modulus[h] (Mpa)	Strength[i] (Mpa)
S-1	186	0	6.4	0.9	2.6	84	8.7	29	0.87
S-2	176	0	12.8	1.2	2.6	84	8.2	32	1.01
S-3	156	0	26.4	1.5	2.6	88	8.4	34	1.09
S-4	127	0	45.4	2.3	2.6	92	8.3	33	1.12
S-5[j]	186	0	6.4	0.9	2.6	84	7.9	27	0.61
V-1	0	120	12.8	1.7	5.4	88	8.2	—	—

[a] Polymerized for 24 h at 60℃. [b] The other component of the emulsion was 180 g of pure water. [c] Chloromethylstyrene. [d] Divinylbenzene. [e] α, α′-Azobisisobutyronitrile. [f] Sorbitan monooleate. [g] Measured by mercury intrusion method. [h] Compressive modulus. [i] Compressive strength. [j] The emulsion was prepared by using a conventional mixer.

図５　表の構造解析例

表は“ 高分子論文集，Vol．61，No．1，pp．12―21（Jan.，2004）”より引用



Polymer	T_g / ℃	$\Delta H_i / \times 10^2$ mJ g^{-1}	
		cal.	obs.
PP	90	53	51
LDPE	81	47	42
HDPE	87	47	46

構造化データ

Type	Entity
物性	enthalpy_change(ΔHi)

Type	基数	基本単位	指数
単位	10		2
		milli joule	1
		gram	-1

図６　表のヘッダーセルの意味解析

図は，“ 加藤ら，材料科学論文の表の意味解釈データセットの構築，言語処理学会第 28 回年次大会 ”より引用

このために，我々は，物性や単位の辞書を人手で構築し，約 3000 事例についてアノテーションを行った．解析手法は，入力文字列を系列ラベリングの問題として捉えて，各文字ごとに BIOE タギングを行い，物性や単位の先頭文字は B，中間文字は I，終了文字は E，それ以外は O のタグを付与して，それを学習データとした．解析モデルは，，畳み込みニューラルネットワークにスキップ結合を入れたブロックを３層積層し，最後に線形層と CRF 層を持つニューラルネットワークを構築した．実験結果を図に示す．

Type	Entity	Precision	Recall	F1
基本単位	Å	1.000	0.923	0.960
	MPa	1.000	1.000	1.000
その他の数量表現	指数	0.989	0.966	0.977
	基数	1.000	0.966	0.982
物性	熱分解温度	0.500	0.500	0.500
	ガラス転移温度	0.842	0.842	0.842
高次構造情報	数平均分子量	0.522	0.800	0.632
実験手法		0.667	0.400	0.500
Material		0.837	0.772	0.803
Sample		0.862	0.581	0.694
溶媒		0.833	0.333	0.476

図 7　表の意味解析の実験結果
図は，“ 加藤ら，材料科学論文の表の意味解釈データセットの構築，言語処理学会第 28 回年次大会 ” より引用

図 7 に示す通り，基本単位や数量表現については 0.96 以上の高い F 値で解析を行うことができた．物性や実験手法などの項目については，学習データの量が十分に多いものについては比較的高精度で，そうでないものは 0.5 程度の F 値に留まるという結果であった．

以上のように，本研究では，PDF 論文に含まれる図表や段落などの基本構造を認識するモデルを構築し，表については，内部のヘッダーセルについて意味解析を行う手法を実現した．また， モデルの訓練や評価に必要な図表領域のデータセットや，本文および表の構造に関するデータセットなどのリソースを構築し，論文の統合的な構造解析を行う技術を確立した．本研究の成果を活用することにより，ある分野における論文の実験データを網羅的に収集することや，図表に記述されている情報の細かい分析や検索が可能になると考えられる．また，本技術を用いて様々な分野の論文を構造化して知識データベースを構築し，ユーザーが閲覧できるようなサービスの実現も可能になると考えられる．

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4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 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） 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） 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 －

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） 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12件（うち招待講演　0件／うち国際学会　10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

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

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

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

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関