材料科学論文の表の意味解釈データセットの構築

加藤明彦 [1] 近藤修平 [1] 進藤裕之 [1,2] 渡辺太郎 [2]
[1] 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
[2] 奈良先端科学技術大学院大学 情報科学研究科
{kato.akihiko.ju6,shuhei-k,shindo,taro}@is.naist.jp

概要

本研究では材料科学論文の表の意味解釈，即ち，セルに記載されている物性や単位の検出とリンキングを行うタスクに取り組む．そこでモデル構築の基盤となる言語資源として，近年の高分子論文集の論文約 200 本に含まれる全ての表に対して注釈を行った網羅的なデータセットを構築した．また，文字レベルの系列ラベリングと辞書マッチングを組み合わせた解析を行ったので合わせて報告する．

1 はじめに

日々多数出版される科学技術論文の全てを人手で読解する事は物理的に不可能であるため，論文から自動で情報抽出を行う技術の研究開発が求められる．特に，論文の表には実験に関する情報が整理されているため，表からの情報抽出は重要な研究課題である．たとえば高分子化学の論文では，実験によって合成されたポリマー（重合体）の物性値や実験条件が表に記載されているため，これらの情報を表から抽出したいというニーズがある（図 1）．ここでポリマーの例としては，ポリアセチレンやポリエステルが挙げられ，物性値とは，融点や引張弾性率など，ポリマーの持つ特性情報を指す．

上記を踏まえ，本研究では，材料科学論文の表からの情報抽出において最も重要なタスクの 1 つである，表中のセルの意味解釈に取り組む．このタスクでは，主にヘッダ部分のセルに記載されている物性や単位といったエンティティの表層文字列（メンション）を検出した上で，材料科学分野に特化したエンティティ辞書への紐付けを行う．たとえば図 1 の表中の枠線で囲ったセルには，エンタルピー変化（エンティティ名: enthalpy_change）という物性の表層文字列である ΔH_i と，その単位の表層文字列である 10^2 mJ g^{-1} が出現している．表の意味解釈ではこれらの各表層文字列が物性および単位のメンションであることを予測し，かつ，前者を enthalpy_change エンティティに紐付け，後者を基本単位のべき乗の積に分解する必要がある（図 1）．

Polymer	T_g / °C	ΔH_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(ΔH_i)

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

図 1 表の意味解釈の例．

しかし，材料科学分野において，表の意味解釈のための言語資源はあまり存在しておらず，一定以上の規模のアノテーションを付与したデータセットの入手は困難である．そこで本研究では，材料科学分野に現れる単位や物性の辞

書を構築した上で，近年の高分子論文集[1]の論文約 200 本に含まれる全ての表に対して意味解釈アノテーションを行った網羅的なデータセットを構築した．また，文字レベルの系列ラベリングと辞書マッチングを組み合わせた解析を行ったのでその結果も合わせて報告する．

2 関連研究

Web ページ中の表の意味解析に取り組んだ研究として [1] が挙げられる．彼らは表の列のタイプ（アルバムのリリース日などの属性名）と，列のペアの関係を予測するという課題に取り組んでいる．また，予測するラベルの候補としては，知識ベースのオントロジーに登録されているものを用いている．彼らはこの課題に取り組むために，テーブル内およびテーブル間の文脈情報の両方を考慮した新しいテーブル表現の学習手法を提案している．本研究は，Web ページではなく論文中の表を扱っている点，また，列ではなく，表のセル内に出現しているエンティティのリンキングを行っている点において [1] とは異なる．

材料科学論文からの情報抽出に関する研究としては [2] が挙げられる．彼らは XML 形式の論文中の表からルールベースで構造化データを抽出している．一方，本研究では PDF 形式の論文中の表を解析対象としており，また，エンティティ・メンションの検出を系列ラベリングとして定式化して機械学習ベースのモデルを用いている点において [2] とは異なる．

材料科学分野に限らず，学術文書一般からの情報抽出については，論文間の関係性を予測する研究 [3] や，個々の論文に記載されている知見を抽出する事を目的とした研究がある．後者についてはキーフレーズ抽出とキーフレーズ間の関係抽出に関する Shared task [4] も実施されている．また，PDF や画像など，非構造化文書中の表の構造解析については，テーブル画像と対応する構造化 HTML 表現を収録したデータセットを [5] が公開している．なお，本研究で取り組んでいる表の意味解釈は，表の構造解析の下流タスクとして位置づけられる．

3 材料科学論文の表のセルの意味解釈

本タスクでは，入力として受け取ったセル文字列中に出現している物性や単位（以下，組立単位）の表層文字列の範囲を予測し（メンション検出），辞書中の特定のエンティティへの紐付けを行う．ここで組立単位とは，複数の基本単位のべき乗の積で構成される単位のことであり，たとえば図 1 の $10^2\,\mathrm{mJ\,g^{-1}}$ という組立単位は 10^2, $(\mathrm{milli\,joule})^1$, g^{-1} という 3 つの構成要素の積に分解できる[2]．図 1 からも分かるように，単位の意味解釈とは，基数，基本単位，指数などのカテゴリに情報を埋めていくタスクである．特に基本単位については，あらかじめ材料科学分野の辞書に登録してある単位一覧から特定の単位を選択する問題となる．

4 データセットの構築

4.1 アノテーション仕様の策定

材料科学論文の表には，合成した物質（高分子など）の物性（融点など）や高次構造情報（数平均分子量など），基本単位や指数などの各種数量表現の他にも，実験手法や，物質の合成時に用いる溶媒や触媒などがしばしば記載される．そこでまず PoLyInfo[3] をひな形として，材料科学論文から収集したセルを整理して，本研究でアノテーション対象とする材料科学分野のエンティティ・タイプの一覧を作成した（表 1）．

4.2 各タイプのエンティティ辞書の作成

次に，表から検出したエンティティを紐付けるために，材料科学分野に特化したエンティティの辞書を各タイプについて構築した．辞書の構築にあたっては，まず PoLyInfo から，単位，物性，実験手法等，各タイプに属するエン

1） https://main.spsj.or.jp/c5/koron/koron.php

2） 基本単位には M (mega), m (milli) 等の接頭辞が付与されることもある．

3） https://polymer.nims.go.jp/

表 1　検出対象とするエンティティタイプ一覧と本研究で構築したコーパスにおける出現数.

大分類	Entity type	Entity の例	出現数
数量表現	数値	$\mathbf{10}^{3}$	125
	接頭辞	$\mathbf{m}g$	1665
	基本単位	g, Pa	
	指数	10^{-3}	394
	演算記号	/ · × ( )	442
物性		T_g, T_m	1346
高次構造情報		M_n, M_w	257
Material		PMMA	1039
分子構造		-CH-	51
Sample		First layer	232
実験手法		NMR	67
溶媒		toluene	53
添加剤		VGCF	39
開始剤		DCP	13
触媒		Na-Diph	19
収率		Yield	83

ティティの一覧を収集した．これだけでは網羅率が十分ではないと考えられるため，表の意味解釈アノテーション（4.3 節）において，各タイプの辞書に紐付けることができないエンティティの内，Wikidata [4] にリンク可能なものを辞書に追加した．辞書中の各エンティティについては，ID，正式名，同義語リスト（当該エンティティが取りうる表層文字列のリスト）を収録している．

4.3　高分子化学論文の表に対する意味解釈アノテーション

第 3 に，高分子論文集 [5] の 2004 年 1 月～2013 年 3 月（Vol.61～Vol.70）の論文約 200 本に含まれる全ての表に対して以下のアノテーションを行った．まず各セルにおいて，検出対象とするいずれかのタイプ（表 1）を有するエンティティが出現している場合，セル文字列中のスパンと，当該エンティティの辞書 ID を注釈する．単位についてはスパンを注釈した後に，基本単位や基数のべき乗の積への分解アノテーション（3 節）を行った．この分解作業の目的は以下の 2 点である．第 1 の目的は，組立単位間の相互変換を可能にすることによって，各物性に紐づく単位の複数の表記方法の比較・照合を行うことにある．第 2 の目的は，構成要素中の接頭辞や基本単位をエンティティ辞書に紐付けることで，組立単位の意味内容を理解することにある．なお，表のセルに対してアノテーションを行う際には，PDF のアノテーションツールである PDFAnno [6] を用いた．

4）https://www.wikidata.org/wiki/Wikidata:Main_Page
5）https://main.spsj.or.jp/c5/koron/koron.php

5　解析手法

本節では 4 節で構築したデータセットを学習データとして用い、セルの意味解釈を行うモデルの訓練と評価を行う．

セル文字列の解析は，(1) 文字レベルの系列ラベリングと，(2) 辞書マッチングによるエンティティ・リンキングの 2 ステップで行う．まず (1) ではセルの文字系列を入力として，系列ラベリングによってエンティティのスパンを同定する．ただし物性と高次構造情報については全エンティティを単一のラベルに集約したものを用いる．また，単位についても，基本単位に属する全エンティティを単一のラベルに集約する．次に (2) では，スパンが同定されたエンティティの内，物性，高次構造情報，または基本単位であると予測されたものを入力として，辞書マッチングによるエンティティ・リンキングを行う．各エンティティを別々のラベルとして扱えば，文字レベルの系列ラベリングのみで解析を行うことも可能だが，物性または高次構造情報のエンティティ総数は約 140 種類，基本単位のエンティティ総数も約 80 種類と多く，データスパースネスの影響が無視できないことから，上述の様にパイプライン型のモデルを採用している．また，系列ラベリングのタグ付けスキームは BIEOS 方式とし，スパンが取りうるラベル集合は，表 1 に示したエンティティ・タイプ一覧に基づいて定めた．ネットワーク構造としては，畳み込みネットワークに

skip-connection [7] を入れたブロックを N 層 [6]，積層したものの上に線形層と CRF 層 [8] を持つアーキテクチャーを採用した．

5.1 辞書マッチング

辞書マッチングでは，系列ラベリングで同定したエンティティ・メンションを，各候補エンティティに紐づく同義語辞書とマッチさせることで、エンティティ・リンキングを行う．ただしメンションと同義語の完全一致を必要条件にすると Recall が低下するため，メンションが同義語を部分文字列として含むならば，当該エンティティを候補に加えるものとする．また，辞書のどのエンティティのどの同義語ともマッチしなかった場合はリンキングに失敗したと判定する．

6 実験

構築したデータセットを学習データとテストデータが約 4:1 となるように分割を行った結果，学習データが 3,046 事例，テストデータが 762 事例となった．なお，各表の各セルの（文字列，タグ系列）を 1 事例としている．確率的勾配降下法を用い，学習率を 0.001 として 30 エポック学習を行った後にテストデータでの評価を行った結果を表 2 に示す．まず数量表現については，基本単位に属する主要なエンティティや，指数，基数などについて 90%を超える F1 値が得られた．次に物性や高次構造情報については，エンティティによって F1 値にばらつきが見られた．これらのカテゴリでは基本単位に比べて表記揺れの度合いが大きいため，辞書マッチングが上手く機能しないケースがより多く発生していると考えられる．第 3 に，実験手法や溶媒については F1 値が 50%前後に留まっており，Precision に比べて Recall が低い傾向が見られた．ただしこれらのタイプについては，ドメイン知識を活用することである程度，網羅率の高いエンティティ辞書と同義語辞書を編纂可能であると期待されるため，辞書素性の活用

6） 実装としては N=5 を採用した．

表 2 セルの意味解釈モデルの実験結果．一部のエンティティ・タイプのみを示している．

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

等によって性能を向上させることができると考えられる．

7 おわりに

本研究では材料科学論文の表から物性や単位を検出し，辞書への紐付けを行うタスクに取り組んだ．まずモデル構築の基盤となる言語資源として，高分子論文集の論文約 200 本に含まれる全ての表に対して意味解釈アノテーションを行った網羅的なデータセットを構築した．次に，文字レベルの系列ラベリングと辞書マッチングを組み合わせた解析を行った．今後は表の複数のセルを考慮した大域的なモデルの有効性を検討する予定である．

謝辞

本研究の一部は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の助成による．

参考文献

[1] Daheng Wang, Prashant Shiralkar, Colin Lockard, Binxuan Huang, Xin Luna Dong, and Meng Jiang. Tcn: Table convolutional network for web table interpretation. *Proceedings of the Web Conference 2021*, Apr 2021.

[2] Hiroyuki Oka, Atsushi Yoshizawa, Hiroyuki Shindo, Yuji Matsumoto, and Masashi Ishii. Machine extraction of polymer data from tables using xml versions of scientific articles. *Science and Technology of Advanced Materials: Methods*, 1(1):12–23, 2021.

[3] Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A Smith. Explaining relationships between scientific documents. *arXiv preprint arXiv:2002.00317*, 2020.

[4] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[5] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 564–580. Springer, 2020.

[6] Hiroyuki Shindo, Yohei Munesada, and Yuji Matsumoto. Pdfanno: a web-based linguistic annotation tool for pdf documents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.