新規用途探索を目的とした技術文書からの材料情報抽出

有馬隆広　大熊智子　出羽達也
旭化成株式会社
{arima.tf, okuma.td, izuha.tb}@om.asahi-kasei.co.jp

概要

近年，材料分野において自然言語処理技術に大きな期待が寄せられている．その中でも特に強いニーズとして新規材料の用途探索がある．とりわけ，まだ工業製品で汎用的に採用されていない新規材料には様々な用途が想定されるため，用途探索は材料開発の現場において重要度の高い業務である．本稿では，用途探索が必要な新規材料の例として金属有機構造体（MOF）を対象にした特許文書の要旨を題材にする．特許文書から材料名と発明の用途を固有表現抽出によって取得し，その結果にもとづいて本タスクの課題について議論する．

1　はじめに

材料分野において自然言語処理技術への大きな期待が寄せられている．特許文書や学術論文から材料に関わる情報を抽出して知識ベースに新規化合物を登録することや，材料の合成プロセスを抽出して実験を支援することが望まれている．さらに，この分野特有のニーズとして用途探索がある．特にまだ汎用的に利用されていない次世代の新素材にはその素材の持つポテンシャルにもとづき様々な用途が想定されるため，研究開発の現場において重要度の高い業務である．

このような背景から，技術文書からの材料情報の抽出は国内外で取り組まれてきた．上述のような知識ベースの構築や実験プロセスの自動抽出を目的とした辞書やコーパスが公開されてる．しかし，その抽出対象は主に材料，物性，操作に関するものでありここで対象とする用途に関する情報の抽出に取り組んだ事例は殆どない．

本稿では，材料開発における用途探索業務の支援を目指して，英文特許文書からの情報抽出を行う．用途探索が必要な新規材料の例として金属有機構造体（MOF）を対象にした特許文書の要旨を題材にする．図 1 に示すように材料と用途を示す表現にアノテーションを行った．用途と材料の情報抽出には BERT と BERT+CRF を用いた．評価実験の結果にもとづき，本タスクの課題について議論する．

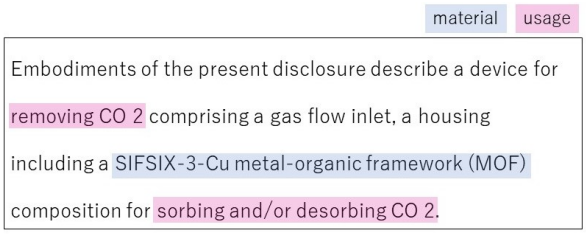

図 1　特許文書における材料（material）と用途（usage）

2　関連研究

特許文書から有用な情報を抽出する試みは以前から取り組まれてきた．[1] では特定の構文パターンを用意することによって技術の応用について書かれた部分を「特徴表現」として抽出している．しかし，その対象にはどのような技術を使ったものなのかなどは含まれておらず，「特徴表現」自身も限定的な範囲に留まっている．このようにあらかじめパターン記述や教師データなどを作成しない情報抽出手法として [2] がある．この手法では請求項間の語の重なりを考慮して発明の新規な部分をキーワードとして抽出するが，本研究では発明中に出現する材料のすべてとその用途を網羅的に抽出することを目的としている．

一方で，材料分野の技術論文を対象にした言語資源がいくつか公開されている．[3] では技術論文から材料の合成プロセスの情報を抽出するためのコーパスを構築している．このコーパスでは材料とそれに対する操作や数量などの個有表現とそれらの関係がアノテーションされている．BioCreativeIV[4] では主に創薬支援を目的とした化学物質と病名などの固有表現を含むコーパスが公開されており，このコーパスを使った固有表現抽出の取り組みがある [5]．

材料分野における言語資源としてコーパスだけで

なく知識ベースやオントロジーも開発されている．日本化学物質辞書（日化辞）は有機化合物のデータベースであるが検索可能な形式で Web 上で公開されている [6]．[7] では日化辞を含む様々なデータベースの情報を対象にした複合名詞の抽出によって化合物情報の可視化を行っている．

しかし，いずれも言語資源に含まれる情報の大半は材料そのものであり用途に対してアノテーションを行ったコーパスは存在しない．本研究ではこれまでにない新たな取り組みとして材料だけではなくそれによってもたらされる機能である用途に対してアノテーションを行ったコーパスを構築し情報抽出を行う．

3 コーパス

3.1 対象データ

本研究では米国で出願された特許のうち “metal organic framework” もしくは “coordination polymer” を含み, さらに ‘application’,‘purpose’,‘use’ のいずれかの語を含む特許の要旨のテキストを取得した．これらの条件はテキスト中に MOF を使った発明の用途が明記されているものを選別するために設定した．さらに，その中から 350 件をサンプリングし，1）MOF を使った発明ではなく MOF の製法に関するもの, 2)MOF に関連しないもの, 3) 用途についての記載がないもの, のいずれかに該当する特許を目視で確認して除外した．その結果，最終的に 263 件の文書が得られた．

3.2 仕様

3.2.1 アノテーションの方針

アノテーションは材料分野に関する知識を有する 1 名が brat[1] 上で行った．

名詞句 今回設定した 2 種類のタグはなるべく体言に対して付与する．また，文頭の冠詞や前置詞などはタグの範囲に含めないようにする．副詞や形容詞が材料を示す名詞句を修飾している場合にはそれらが材料名にとって必須の構成要素になりうるかどうかを個別に判断するよう指示した．これは例えば “inorganic nanoparticles” において ‘inorganic’ が形容詞であっても材料名の中心的な意味を表す場合があるからである．一方で，複合名詞に関してはなるべく広くタグの中に含めるようにした．

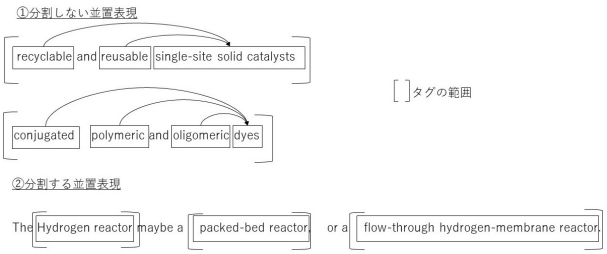

図 2 並置表現を対象にしたアノテーションの範囲

述語表現を含む句 usage には述語や項を含むような表現も存在するため、そのようなフレーズもアノテーションの対象とした．ただし、“sensing platform for detecting the presence or amount of target nucleic acid, particularly HIV-1 ds-DNA” における “particularly HIV-1 ds-DNA” のように補足的な副詞節はその範囲に含めないようにした．

並置表現 材料分野に限らず，特許には並置表現が多く含まれる．今回のアノテーションでは，接続詞で並列になっている名詞句が形容詞句によって均等に修飾されている，あるいは並列する修飾成分が同一の名詞句を修飾している場合は分離せず，そうでないものは別々のタグを付与することにした．この方針の具体例を図 2 に示す．①はエンティティの中心となる名詞句が複数の修飾を受けているため接続詞 ‘and’ を含む名詞句全体に一つのタグを付与している[2]．しかし，②のケースのように名詞句がそれぞれ独立して出現する場合には個別にタグを付与する方針とした．

3.2.2 material

材料を指す表現は分子や化合物から複数の材料によって作られた物質まで粒度の異なる様々なエンティティがある．本稿ではそれらの粒度の区別はせず一律に material タグを付与した．これは，本研究の最終的なゴールがマテリアルズ・インフォマティクスを目的とした実験設定の抽出を目指すものではなく，用途探索を目的とした発想の支援にあるためである．従って下記のような表現が material のアノテーション対象に含まれる．

元素記号／化学式 先行研究では ‘Au’，‘Ag’ などの元素記号や ‘CO2’，‘N(C1-6-alkyl)2’ などの化学式が材料として抽出されている．さらに，‘-NH2’ のように官能基を示す化学式も存在する．

1） https://brat.nlplab.org/

2） discontinuous なタグを付与することも考えられるが今回はアノテーション作業の負荷を考慮して仕様に入れなかった

物質の構造／形態／総称　今回，情報抽出の対象にした MOF のように，特定の物質を指すのではなく物質の構造や形態あるいは物質の総称を指す表現もアノテーションの対象にした．具体的には ‘fluid’，‘mixture’，‘ligand’，‘polymer’, などである．

物質の一部　エンティティの中には物質全体を指すのではなく，その一部を指す表現が含まれる．例えば ‘membrane’，‘layer’ や ‘surface’ などがこれに相当する．

一般名詞　化学記号や化学式ではなく一般名詞として記述されている材料も多い．‘copper’ や ‘silver’ などよく用いられている金属は一般名詞で記載されている．

物質の役割　物質そのものを指すのではなく，その物質が合成プロセスにおいて担う役割を指すエンティティが存在する．‘catalyst’ や ‘reactor‘，‘media’ などがこれにあたる．ただし，実際にアノテーションを行うとこのような物質の役割が用途として書かれていることもあった．今回はその発明が提供する機能かどうかによって material と usage の判断を行ったが，今後はより明確な基準を設定する必要がある．

3.2.3　usage

用途を表す表現は下記に示すように比較的長いフレーズである場合がある．

装置　特許文書であるため，‘battery’ や ‘sensor’ などの装置を指す表現が比較的多い．今回のアノテーションでは MOF が直接関わらないものであってもタグを付与した．

動作　“remove the contaminants from a stream of electronic gas” や “separating hydrogen molecules” など具体的な動作と対象についての記述がある．このように具体的な動作を記述する場合はエンティティというよりも述語と項から構成されるフレーズであることも多い．

3.3　統計情報

表 1 に本実験で作成したコーパスの統計量を示す．usage に比べて material の数が 5 倍以上多いのは，一つの用途に対して複数の材料が用いられることが考えられる．また異なり数の割合が usage の方が高い．これは要旨の中であっても，特許では同じ材料に対する記述が繰り返し現れる傾向があることに起因している．

表 1　コーパスの統計量

	material	usage
Entity 数	3,656	698
異なり数	1,627	508
平均語数	2.0	2.6
平均文字数	15.0	18.8

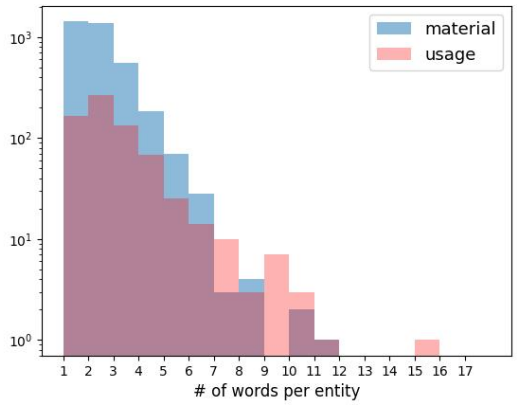

図 3　エンティティの長さ（語数）の分布

図 3 に material と usage のエンティティを構成する単語数の分布を示す．material はエンティティを構成する語の数が多くなるほど頻度が低くなる傾向があるが，usage は語数によらず比較的頻度が一定である．これは物質名が主である material に比べて usage の表現が多様であることを示している．

4　実験

アノテーションでは，特許の要約テキストを対象に，文に含まれる単語に対して固有表現ラベル material, usage を付与した．固有表現抽出は系列ラベリングタスクとして取り組み，アノテーションを付与したデータセットを 6:2:2 の割合で訓練・開発・評価に分割して使用した．具体的には，対象テキストをモデルへ入力して各トークンの分散表現を得，分散表現を全結合層に入力して得られた結果を更に CRF 層へと入力することで各トークンに対する固有表現ラベルを判定する．モデルとしては，BERT に全結合層のみ追加したモデルと，BERT に全結合層と CRF 層を追加したモデルの 2 種類をベースラインとして検討し，どちらのモデルにおいても BERT_{base} を利用して fine-tuning を行った[3)]．

3)　パラメータは以下の通り．最大系列長:512，バッチサイズ:8，エポック数:5，学習率:1e-5，最適化アルゴリズム:Adam

表 2　固有表現抽出の実験結果

	material			usage		
	Pre.	Rec.	F1	Pre.	Rec.	F1
BERT	0.67	0.74	0.70	0.34	0.45	0.39
BERT+CRF	0.73	0.82	0.77	0.58	0.37	0.45

5　実験結果と考察

5.1　実験結果

表 2 に固有表現抽出の実験結果を示す．どちらの固有表現カテゴリにおいても，BERT と比較して BERT+CRF の方がより高い性能を示している．material の F 値は 0.77 という一般的な専門用語抽出とほぼ同等の性能が得られた．しかし，usage は F 値が 0.5 にも満たない性能であり固有表現抽出というタスク設定が適切だったのか再考する必要がある．usage の方が material よりも抽出精度が低い理由として，訓練データに含まれる数が少ないことや，エンティティの長さが比較的長いことが挙げられる．

5.2　エラー分析

抽出エラーについて，スパンとラベルの誤りにどのような傾向があるかを観察した．

スパンの誤り傾向　スパンの誤りに比較的多く見られたのは，図 4 に示すように usage タグの一部に material が含まれているとその部分だけを material と誤って推定してしまうケースである．今回はアノテーション作業の負荷を考えて入れ子構造を仕様に入れなかったがこのようなエラーを回避するためには検討する必要がある．またエンティティがスパンの途中で切れる・スパンの途中から始まる・途中で 2 つに区切られるケースについては，コーパスに含まれる全エンティティ数が十分な量でないことで訓練データに含まれる文字数が比較的長いエンティティの数も少なくなったことによると考えられる．

ラベルの誤り傾向　スパンは一致しているが usage でなく material に誤るケースがある．このエラーは文脈によって材料名にも用途名にもなり得る単語が存在することに起因すると考えられる．

6　おわりに

本稿では材料の新規用途探索を行うための第一歩として，特許文書中に含まれる材料名とその用途に対してアノテーションを行った．その結果，材料名と用途はいずれもその出現形態や内容が多岐にわたり，特に用途は体言では無い表現も多いことが分かった．また，BERT による抽出実験によって材料名については固有表現抽出と同じ方式で実用的な精度に達する見込みがあることを示した．但し，用途については同じ方式で抽出することが難しく，文分類などの別のタスク設定を検討する必要があると思われる．

今後は本研究の最終的な目的である用途探索に向けたテキストマイニングをユーザと共に実施し，今回行った情報抽出の有用性や必要な精度などを検証したい．

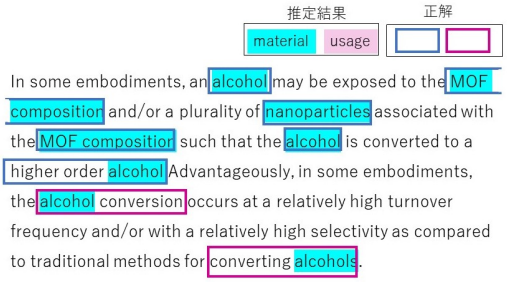

図 4　固有表現抽出結果と正解のギャップ

参考文献

[1] 西山莉紗, 竹内広宣, 渡辺日出雄, 那須川哲哉. 新技術が持つ特徴に注目した技術調査支援ツール. 人工知能学会論文誌, Vol. 24, No. 6, pp. 541–548, 2009.

[2] Shoko Suzuki and Hiromichi Takatsuka. Extraction of keywords of novelties from patent claims. In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics**, 2016.

[3] Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In **Proceedings of the 13th Linguistic Annotation Workshop**, 2019.

[4] Biocreative IV, 2014. https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/.

[5] 渡邊大貴, 田村晃裕, 二宮崇, 牧野拓哉, 岩倉友哉. 化学分野の固有表現抽出のための化合物名を含む文の言い換え学習を用いたマルチタスク学習手法. 自然言語処理, Vol. 29, No. 2, pp. 294–313, 2022.

[6] NBDC 版日化辞 rdf. http://dbarchive.biosciencedbc.jp/jp/nikkaji/desc.html.

[7] Kazunari Tanaka, Tomoya Iwakura, Yusuke Koyanagi, Noriko Ikeda, Hiroyuki Shindo, and Yuji Matsumoto. Chemical compounds knowledge visualization with natural language processing and linked data. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.