情報抽出による質の高い新規用途アイデアの獲得

谷口友紀 高橋拓誠 大熊智子
旭化成株式会社
{taniguchi.tcr,takahashi.tkr,okuma.td}@om.asahi-kasei.co.jp

概要

近年，材料の用途探索に自然言語処理を活用する取り組みが広がっている．材料固有の特性を発揮させる使用法が用途であることから，用途探索は，材料の特性が発揮できる用途を探し出すタスクとなる．本論文では，材料固有の特性に類似する特性エンティティを含む文書を検索したのち，特性エンティティに関係する用途エンティティを同文書中から取得することにより，材料特性を活かした用途アイデアの獲得手法を提案する．さらに本論文では，提案手法で獲得した用途アイデアを対象に専門家による新規性と実現性の評価を実施し，その有効性を確認する．

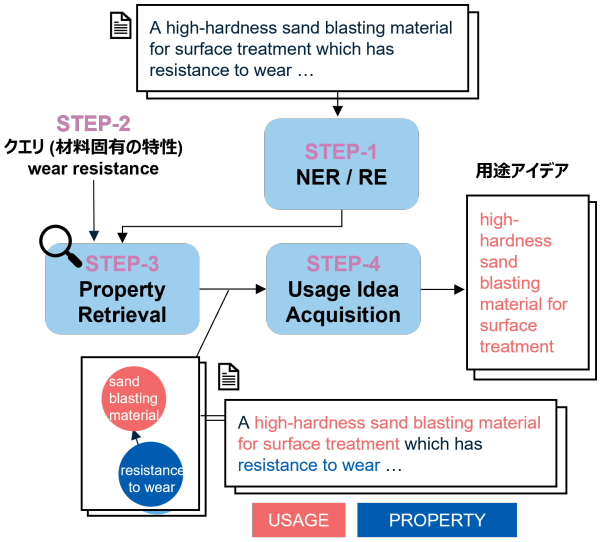

図 1 提案手法の概要

1 はじめに

材料の用途探索に対する自然言語処理への期待が高まっている．産業の上流に位置する基礎材料は応用先が広く，未だ知られていない用途が存在する．しかしながら，用途考案には専門性を持つ従業員による調査・分析が必要であり，具体的な検討につながるアイデアを生み出すのは容易ではない．理想的な用途アイデアには新規性と実現性の両立が求められる．新規性はアイデアの斬新さを，実現性は材料によるアイデアの実行可能性を表す．また，材料が有する特性が発揮できる使用法が用途であることから，用途探索は，材料の特性が発揮できる質の高い用途アイデアを発見するタスクとして定義できる．

用途探索を実現する一般的な方法として，材料固有の特性をクエリとして関連文書を検索し，検索結果である文書集合から用途を抽出する手法が考えられる．しかしながら，クエリと文書の類似度に基づき関連文書を探す文書検索では，検索結果で得られた文書集合が類似してしまう．その結果，そこに含まれる用途アイデアの多様性が失われ，新規性のあるアイデアが埋没する．本論文では，材料固有の特性をクエリとして，その固有特性に類似する特性が記載されている文書を検索したのち，文書中においてその特性と関係している用途を抽出することで，実現性と新規性を兼ね備えた用途アイデアを獲得する手法を提案する．

図 1 に提案手法の概要を示す．提案手法は，固有表現抽出 (**Named Entity Recognition;NER**) により文書から用途と材料の特性に関するエンティティを抽出し，文書内の共起を用いてエンティティを関係 (**Relation Extraction;RE**) づける．その後，探索中の材料固有の特性をクエリとして，同義もしくは類似の特性エンティティを含む文書を検索する (**Property Retrieval**)．そして，検索で得られた文書集合に含まれる用途エンティティを，材料固有の特性と特性エンティティの類似度にしたがってソートし，類似度の高い特性エンティティと共起した用途エンティティを新規用途アイデアとして獲得する (**Usage Idea Acquisition**)．つまり，提案手法は文書中の用途エンティティのアイデア品質を，用途実現への寄与が期待できる特性エンティティとの類似度で見積もる．さらに，提案手法は，材料固有の特性を表すクエリと，文書中に存在する特性エンティ

ティの類似度のみに基づき用途アイデアを検索するため，新規性のある多様なアイデアを獲得することができる．本論文の貢献は以下の 2 点である．

- 用途探索中の材料固有の特性と，文書から抽出した特性エンティティの類似度に基づき，新規性と実現性を両立する用途アイデアを獲得する手法を提案する．
- 提案手法で得た用途アイデアの新規性と実現性を材料分野の専門家により評価する．

2 関連研究

2.1 材料分野における情報抽出

化学や材料分野の情報抽出に関する研究の多くは，文書から材料や特性を抽出し，その知識を活用することを目指している [1, 2, 3]．しかしながら，本論文が対象とする用途情報の活用を対象とした研究はほとんど存在しない．有馬らは用途情報を抽出することを目指して，金属有機構造体を題材としたコーパスを提案した [4]．また，Weston らは検索結果に用途情報を付与することによって，優れた可視化を実現する検索システムを提案した [5]．しかしながら，これらの研究は，テキスト構造化による文書内容の把握に焦点が当てられており，新たな用途アイデアの獲得に踏み込んでいない．

2.2 アイデアの発想支援

近年，生成 AI によるアイデアの発想支援システムが提案されている．Li らは論文の引用関係から技術進展の繋がりを構築して指示文に加えることにより，専門家に匹敵する研究アイデアが生成できることを示した [6]．Wang らは，過去文献と比較しながら，新規性が十分に達成されるまでアイデア生成を繰り返す手法を提案した [7]．これらの取り組みは研究アイデアの発想支援を目指しており，新規用途の開拓を支援するものではない．

3 用途探索コーパス

3.1 対象材料とアノテーション

本論文では，硬化性組成物 (*Curable Composition;CC*)，結晶性樹脂 (*Crystalline Resin;CR*)，非結晶性樹脂 (*Non-Crystalline Resin;NCR*) に関する用途探索を想定する．用途探索コーパスは特許要旨を対象に構築されており，人手によりエンティティと関係ラベルが付与されている．本論文では，その中から用途と材料の特性エンティティを表す *USAGE* と *PROPERTY* ラベルを利用する．*USAGE* は発明の主たる用途を指し，*PROPERTY* は材料が有する物理的な特性を示すラベルである．*USAGE* に関するアノテーション指針は有馬らに従う [4]．本論文では，導電性や熱抵抗などの物理特性に関する表現に *PROPERTY* を付与する．

表 1 エンティティラベルの出現頻度[1)]

領域	文書数	*USAGE*	*PROPERTY*
CC	619	2679 (4.33)	1471 (2.38)
CR	339	986 (2.91)	358 (1.06)
NCR	466	1498 (3.21)	891 (1.91)

3.2 統計量

エンティティラベルの出現頻度を表 1 に示す．用途探索コーパスは文書数が数百と小規模であり，文書には複数の *USAGE* が含まれている．一方で，*PROPERTY* の出現頻度は低い．なお，*USAGE* を構成するエンティティの平均単語数は 2.4，*PROPERTY* は 4.1 である．

4 提案手法

4.1 材料固有の特性に基づく用途獲得

提案手法は以下の 4 ステップから構成される．検索対象の文書集合を D={$d_1, d_2, ..., d_N$} で定義する．

STEP1 NER/RE 文書 d_i から特性エンティティ P={$p_1, p_2, ..., p_M$} と用途エンティティ U={$u_1, u_2, ..., u_L$} を抽出する．そして，文書内で共起関係にある P と U に関係ありのラベルを付与する．

STEP2 固有特性によるクエリ定義 用途探索中の材料固有の特性 T={ $t_1, t_1, ..., t_J$} を定義する．

STEP3 類似度の算出 材料固有の特性 T と文書 d_i に含まれる特性エンティティ P の類似度 s_i を式 1 により算出する．

$$s_i = \sum_{j=1}^{J} \underset{m=1,...,M}{max}[similarity(p_m, t_j)] \quad (1)$$

STEP4 用途アイデアの獲得 STEP1 から STEP3 を繰り返し，文書集合 D について材料固有の特性 T と特性エンティティ P の類似度を算出する．類似度

1) 括弧内の数値は文書あたりの出現頻度を表す

表 2　DyGIE++による $f1$ スコア

領域	*USAGE*	*PROPERTY*
CC	0.785	0.590
CR	0.719	0.760
NCR	0.774	0.632

が高くなった特性エンティティと関係ラベルで紐づく K 個の用途エンティティを取得し，用途アイデアとして採用する．K は定数である．

5　実験設定

5.1　NER モデルの構築と評価

提案手法の NER モデルには，DyGIE++[8] を採用した．DyGIE++は固有表現抽出と関係抽出をマルチタスクで学習する情報抽出技術であり，科学文書において優れた性能を達成できることが報告されている．提案手法は NER 結果を活用するため，その性能が用途アイデアの品質に影響を与える．そこで，提案手法では直接利用しないが，*USAGE* と *PROPERTY* に加えて，用途探索コーパスに含まれる全てのエンティティと関係ラベルを利用してモデルを構築した．DyGIE++による NER の評価結果を表 2 に示す．*PROPERTY* と比較して *USAGE* で良好な $f1$ スコアが得られた．これは，*"The invention discloses a **stainless steel container**"*のように，ほぼ定型の表現で文頭に用途が記載されるためである．

5.2　用途アイデアの評価

CC，*CR*，*NCR* の 3 材料を対象に，提案手法で獲得した用途アイデアを新規性と実現性の観点で評価した．材料分野の専門家 3 名が，得意とする材料についての評価を実施した．新規性と実現性のそれぞれについて，以下の 3 段階の評価基準を採用した．

新規性の評価基準

2: これまで思いつかなかった
1: 広く知られていない
0: 広く知られている

実現性の評価基準

2: 実現の可能性がある
1: やや実現の可能性がある
0: 実現は不可能である

新規性と実現性はトレードオフの関係にあるため，その両立が求められる．そこで，新規性と実現性スコアが共に **2** なら *EXCELLENT*, 新規性と実現性スコアが **2**, **1** または **1**, **2** なら *VERY GOOD*, 新規性と実現性スコアが共に **1** なら *GOOD* として個々のアイデア品質を評価した．さらに *EXCELLENT* を 3 点，*VERY GOOD* を 2 点, *GOOD* を 1 点として，獲得アイデアの総合スコアを評価した．重複した用途アイデアは除去し，得られた 50 の用途アイデアに対して評価を実施した．評価者である専門家は用途アイデアのリストと，用途アイデアの抽出元である特許要旨を参照して評価を行った．また，材料分野の専門家との議論から，*CC*，*CR*, *NCR* が有する材料固有の特性 T として，接着性，摺動性・耐摩耗性，架橋性などを定義した．

5.3　比較手法

提案手法で利用する特性エンティティとの類似度を算出する方法として，以下の 2 種類を評価した．OpenAI 社が提供している text-embedding-3-large[2] による埋め込み表現を使ったコサイン類似度 (PROPERTY-EMB-3) と，Word Mover's Distance[9] から算出した類似度 (PROPERTY-WMD) である．後者の単語埋め込み表現には FastText[10] を利用し，米国の公開特許テキストを使って事前学習を行った．また，ベースライン手法として，特許要旨とクエリである材料の固有特性を text-embedding-3-large で埋め込み表現に変換し，そのコサイン類似度によって関連文書を検索する文書検索手法 (DOC-EMB-3）を準備した．

6　評価結果

6.1　総合スコア

図 2 に *CC*，*CR*，*NCR* における用途アイデアの評価結果を示す．総合スコアから，提案手法に基づく PROPERTY-EMB-3 と PROPERTY-WMD が優れた用途アイデアを獲得することに成功している．アイデア品質の観点では，ベースライン手法である DOC-EMB-3 と比較して，提案手法は *EXCELLENT* もしくは *VERY GOOD* の評価を獲得できており，新規性と実現性を両立することができている．

6.2　新規性・実現性

新規性と実現性の評価軸でみると，提案手法はベースライン手法と比較して，新規性の高いアイデアを獲得している (図 2)．実現性については，*CC* で

2)　https://platform.openai.com/docs/guides/embeddings

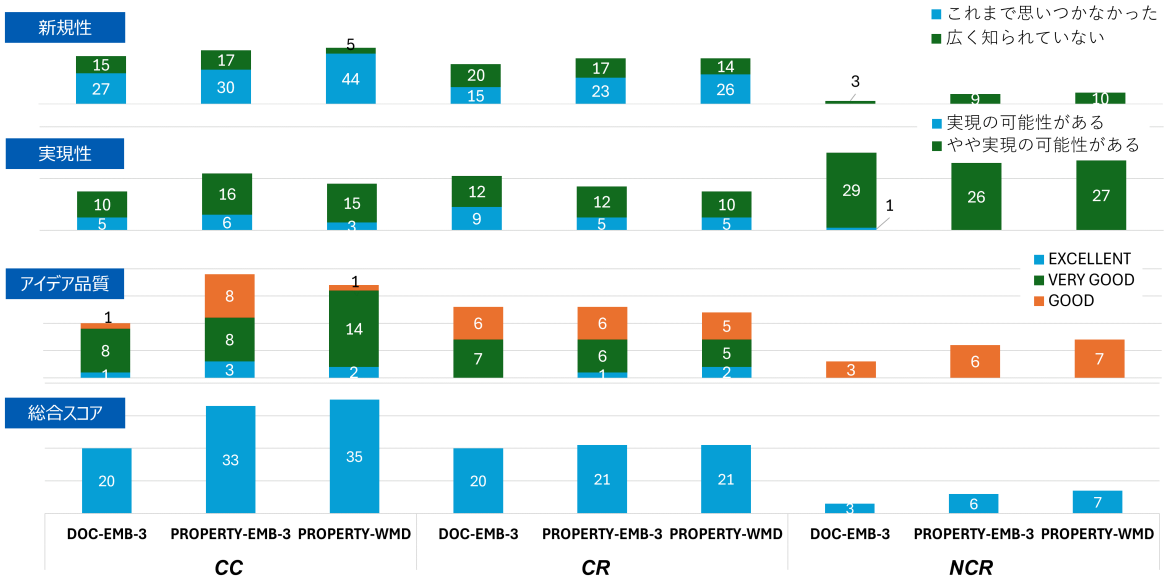

図 2　専門家による用途アイデアの評価結果

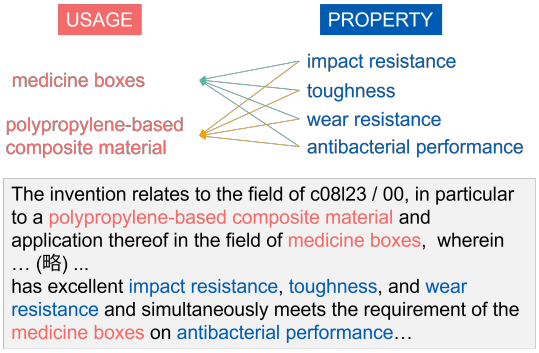

図 3　*CR* で獲得した用途アイデアの具体例

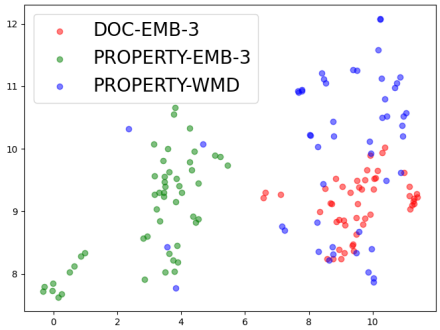

図 4　用途エンティティの埋め込み表現による可視化

は提案手法がやや良好であるものの，*CR*，*NCR* ではやや劣る結果となった．しかしながら，ベースライン技術である DOC-EMB-3 は新規性と実現性を両立できず，提案手法に総合スコアで劣る．

6.3　用途アイデアの具体例

CR において，提案手法である PROPERTY-EMB-3 が獲得した *EXCELLENT* な用途アイデアと抽出元文書を図 3 に示す．提案手法により耐衝撃性，強度，耐摩耗性と抗菌性を有するメディシンボックスに関する用途アイデアが抽出できている．

6.4　考察

獲得した用途アイデアを text-embedding-3-large で埋め込み表現に変換し，UMAP[11] により可視化した (図 4)．提案手法である PROPERTY-EMB-3，PROPERTY-WMD はベースライン手法である DOC-EMB3 と比較して広範囲に用途アイデアが分布している．この結果は，提案手法が多様な用途を抽出しており，新規性の高いアイデアが得られていることを示唆している．

7　結論

本論文では，新規性と実現性を両立する用途アイデアの獲得手法を提案した．材料分野の専門家による評価から，提案手法により新規性と実現性を両立するアイデアが獲得できることを確認した．本論文では，共起によりエンティティ間の関係を定義したが，機械学習に基づく関係抽出モデルの利用が望ましい．文書レベルの関係抽出の難易度は高いが，その性能向上に取り組み，導入を進めたい．また，用途アイデアを生成する発想支援エージェントの構築にも取り組みたい．

参考文献

[1] Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In **ACL**, 2020.

[2] Yunsoo Kim, Hyuk Ko, Jane Lee, Hyun Young Heo, Jinyoung Yang, Sungsoo Lee, and Kyu-hwang Lee. Chemical language understanding benchmark. In **ACL(Industry Track)**, 2023.

[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 **LAW**, 2019.

[4] 有馬隆弘, 大熊智子, 出羽達也. 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. In **NLP**, 2023.

[5] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, L.A.Persson, G.Ceder, and A. Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. In **ACS**, 2019.

[6] Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. Chain of ideas: revolutionizing research via novel idea development with llm agents. In **arXiv preprint**, 2024.

[7] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In **ACL**, 2024.

[8] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contexualized span representations. In **EMNLP**, 2019.

[9] Matt J. Lusner, Yu Sun, Nicholas I. Kolkin, and Killian Q. Weinberger. From word embeddings to document distances. In **ICML**, 2015.

[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Thomas Mikolov. Enriching word vectors with subword information. In **TACL**, 2017.

[11] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. In **arXiv preprint**, 2018.