DEIM2022 E43-2(day4 p11)

特許文書を用いた物質と特徴の関係理解に基づく物質の意外な用途の発見

古屋　昭拓†　　山本　岳洋†　　窪内　将隆††　　大島　裕明†

† 兵庫県立大学 大学院情報科学研究科　〒 650–0047 兵庫県神戸市中央区港島南町 7–1–28
†† 堺化学工業株式会社　〒 590–8502 大阪府堺市堺区戎島町 5–2
E-mail: †ad21m044@gsis.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp, †††kubouchi-m@sakai-chem.co.jp, ††††ohshima@ai.u-hyogo.ac.jp

あらまし　本研究では，物質の意外な用途を推定する手法を提案する．例えば，酸化チタンという化学物質に対して潤滑剤という用途を推定する．酸化チタンは白色の物質であり，紫外線を反射する特徴があるため，顔料や充填剤といった用途で使われている．また，タルクという物質も白色の物質であり，紫外線を反射する特徴があるため，同様の用途で使われている．このように，特徴が類似する物質は共通の用途を持つ傾向がある．タルクは潤滑剤としても使われていることから，タルクと特徴が類似する酸化チタンも潤滑剤として使われる可能性があると考えられる．このような考えに基づき意外な用途の推定を行う．具体的には，特許文書を対象に物質と用途の抽出を行い，対象の物質と特徴が類似する物質の推定とその用途の推定を行うことによって，対象の物質の用途ではない用途を意外な用途として推定する．

キーワード　深層学習，固有表現抽出，特許，HITS

1　は じ め に

物質の新たな用途を発見する際，多くの時間とコストをかけて研究を行うが成果が実らないことも多い．例えば，化学物質である酸化チタンの新たな用途を発見する場合，既に様々な角度から研究が行われているため，新たな用途の発見は困難である．仮に新たな用途を発見できたとしても，かけた時間とコストに見合わないことがあり，費用対効果が悪いという問題がある．

また，新たな用途の意外性を考慮することは困難である．物質の新たな用途として，毎年多くの特許が申請されているが，その内容は似通ったものが多い．進歩性や新規性といった発明としての新しさを有してはいるが，既存の技術を応用した特許が多く，新しい考えに基づいた意外な用途を発見することは困難であると考えられる．そのため，物質の新たな用途の中でも意外性を有した用途を推定する手法が望まれる．

理想的な物質の新たな用途とは，特徴が類似する物質の用途である．例えば，酸化チタンという物質に対して潤滑剤という用途を推定することが望ましい．酸化チタンは顔料や充填剤として頻繁に使われている物質であり，潤滑剤として使われることは少ない．そのため，潤滑剤は意外性のある用途であるといえる．酸化チタンは白色の物質であり，紫外線を反射する特徴があるため，顔料や充填剤として使われている．潤滑剤という用途には，タルクやシリカなどの物質が使われている．酸化チタンはタルクやシリカと共通の特徴を多く持っており，これらの物質は特徴が類似していると考えられる．そのため，酸化チタンは潤滑剤として使用できる可能性が高い．このように，特徴が類似する物質の用途は，意外な用途として新たな用途の発見に利用できる可能性がある．

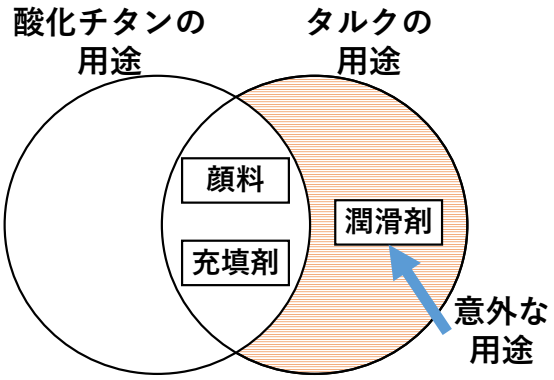

図 1　酸化チタンとタルクの既知の用途の差分を得ることで酸化チタンの意外な用途を推定

このような意外な用途を推定するために，対象の物質と類似物質の既知の用途を推定し，その差分を得ることで意外な用途を推定する手法を提案する．図 1 に例を示す．提案手法は以下の 3 つのステップにて行う．

（1）　対象の物質の用途の推定
（2）　類似物質とその用途の推定
（3）　対象の物質の意外な用途の推定

（1）では，対象の物質の既知の用途を推定する．意外な用途を推定するためには，対象の物質にとって意外な未知の用途を判別する必要があるからである．

（2）では，対象の物質の類似物質を推定し，その既知の用途を推定する．類似物質の既知の用途の中で，対象の物質の既

知の用途ではないものが意外な用途と考えられるためである．

（3）では，(1）と（2）で推定した用途から意外な用途を推定する．対象の物質と類似物質の既知の用途の差分を得ることで，対象の物質の既知の用途にないものが意外な用途であると考えられる．

実験として，10 種類の物質で意外な用途を推定し，その意外性の評価を行った．その結果，大部分の用途は意外性の低い用途だったが，いくつか意外性の高い用途を発見することができた．また，意外な用途の推定に類似物質の中でも類似性の低いものを使うと，より意外性の高い用途を推定することができた．

2 関 連 研 究

2.1 Bidirectional Encoder Representations from Transformers（BERT）

BERT [1] は，近年，自然言語処理におけるさまざまなタスクで利用されている深層学習のモデルであり，特許文書を対象としたタスクにも適用され始めている [4], [9], [10], [11]．BERT は，双方向の Transformer [13] を用いたモデルであり，大規模なデータによって事前学習されている．出力層を一層追加して学習を行うことで，さまざまなタスクに対応することができる．

本研究では，BERT を用いて特許文から物質の抽出を行う．このタスクは固有表現抽出（Named Entity Recognition：NER）といい，BERT をファインチューニングすることによってこのタスクに取り組む．手法の詳細は 5 節で説明する．

2.2 特許文書を対象とした機械学習の研究

近年，特許文書を対象として，機械学習のモデルを用いたさまざまな研究が報告されている．Fall らの研究 [2] では，機械学習のための特許分類タスク用データセットの作成を行っており，そのデータを用いて，さまざまな機械学習アルゴリズムによる特許分類の結果を示している．

自然言語を対象としたタスクにおいて BERT を用いた研究が増えており，特許文書を対象としたタスクにおいても，BERT を用いる研究がいくつも報告されている．Lee らの研究 [10] では，特許文書で追加学習を行った BERT を用いて，特許分類タスクに取り組んでいる．従来の深層学習のモデルである CNN とワードエンベディングを組み合わせたモデルと，追加学習を行った BERT を用いたモデルとで特許分類の精度を比較した結果，BERT が特許分類のタスクにおいて有効であることを示している．加えて，Lu らの研究 [11] では，BERT の最終層 4 層から取り出したベクトルを用いて行列を生成し，その行列を CNN で畳み込むという手法で特許分類タスクに取り組んでいる．

また，Kang らの研究 [4] では，特許検索のタスクに取り組んでおり，こちらの研究では，特許の先行技術調査の際，BERT を用いたモデルを適用することでノイズとなる特許を除去することに成功している．

Lee らの研究 [9] では，自動生成された特許請求項の関連性評価に BERT を用いており，GPT-2 によって生成された特許請求項を BERT で評価することで，特許請求項の自動生成タスクに取り組んでいる．

このように，特許を対象とした機械学習の研究は多数存在している．特に，最新の研究では BERT を用いているものが多く，その有効性が示されている．そのため，本研究でも手法として BERT を用いて特許文から物質の抽出を行う．

2.3 固有表現抽出

文書から特定のラベルが付加された情報を抽出するタスクを固有表現抽出（NER）という．本研究では，BERT を用いて特許文中から物質を表す表現の抽出を行う．

固有表現抽出に深層学習のモデルを用いた研究は数多く存在しており，Lample らの研究 [8] では，深層学習のモデルと Conditional Random Fields（CRF）[7] を組み合わせることの有効性が示されている．深層学習のモデルの中でも，近年，BERT を用いた手法が注目されている [3], [12]．

Souza らの研究 [12] では，BERT を用いたモデルで，ポルトガル語のデータセットから 10 クラスの固有表現抽出を行うタスクにおいて，他のモデルより高いスコアを得ることに成功している．また，Hakala らの研究 [3] も同様に，BERT を用いたモデルで固有表現抽出のタスクに取り組んでおり，その有効性を示している．

2.4 意外な情報検索の研究

意外な情報の発見を行う研究は頻繁に報告されている．佃らの研究 [17] では，Wikipedia から抽出した語句を対象にグラフを作成し、それをもとに意外な情報の検索を行っている．この研究では，共通の上位語を持つような語句を同位語とし，Hyperlink-Induced Topic Search（HITS）アルゴリズム [6] を用いて，任意の語句の同位語の中で，より同位語らしい語句の決定を行うことによって意外な情報の検索を行っている．

中村ら [16] の研究では，意外な検索キーワードを推薦する手法を提案している．意外な検索キーワードを「調べたい事柄と弱い関連があるもの」と定義し，そのようなキーワードの発見をネットワークを用いて行っている．ユーザの検索履歴から，同時に検索されたキーワード間にエッジを持つネットワークを構築し，一つのノードを経由してつながっているノードの意外性の強さを調べることで意外な検索キーワードの発見を行っている．

鈴木ら [14] の研究では，意外性のある検索クエリの推薦を行っている．検索クエリの「思いつきづらさ」と「得られる情報の予想しづらさ」を求め，その積の小さい検索クエリを意外性のある検索クエリとして提示している．学生 14 人が評価した結果，ベースラインと比べて意外性のある検索クエリを推薦した．

また，既存発明の新たな用途を探索する研究も行われている．太田ら [15] の研究では，特許からその発明が達成する効果と解決した課題の抽出を行い，効果は類似していないが，課題が類似している特許の探索を行った．課題が類似していれば既存の技術で解決できる可能性があり，その中でも効果が類似してい

ないものであれば意外性が高い可能性がある．このような考えで既存発明の新たな用途を探索している．本研究は，物質の新たな用途を発見する研究であり，物質の特徴や用途に注目し，類似物質が持つ用途から意外性のある新たな用途を推定する．そのため，太田らの研究とは異なるアプローチで新たな用途の発見を行っている．

3 問題定義

本研究では，物質名を入力するとその物質の意外な用途を出力する手法を提案する．

- **入力：物質名**
- **出力：意外な用途**

入力となる物質名はテキストであり，酸化チタンや酸化亜鉛といった化学物質名などを想定している．出力される意外な用途もテキストであり，充填剤や顔料，潤滑剤などを想定している．

データセットとして特許文書を用いる．特許には，物質と特徴，用途に関する内容が記述されているためである．上記の情報を抽出，分析することで，物質と類似物質の用途を推定し，その差分を得ることで意外な用途の推定を行う．本研究で扱うデータセットの詳細は次節で説明する．

図 1 のように，物質の新たな用途の推定は，物質と類似物質の既知の用途の差分を得ることで実現できると考えられる．このようにして得た用途は，対象の物質の用途とは結ぶつきにくい意外性のある用途と考えられるため，このような用途を意外な用途として推定する．推定された意外な用途の意外性を人手で評価する．本研究では，化学分野の専門家 1 名によって 5 段階で評価を行う．

4 データセット

本研究では，特許庁が提供を行っている特許情報バルクデータ [1]を利用する．提供されている様々なデータの中でも，特許・実用新案公報情報（特許，実用新案登録）の 2019 年の特許データを利用した．

この特許データ 1 件には，書誌情報，明細書，特許請求の範囲などの情報が XML 形式で格納されている [2]．書誌情報には，特許番号，出願番号，特許分類，発明の名称などが記されており，特許分類は，国際特許分類（International Patent Classification：IPC）と日本の特許分類（File Index：FI）に分けて記載されている．それぞれが，特許分類の主分類とそれ以外の分類に分かれて記載されている．主分類は特許に必ず付与されているが，それ以外の分類は付与されていないこともある．また，IPC はセクション，クラス，サブクラス，グループから構成される特許分類であり，FI は，IPC の構成要素に展開記号や分冊識別記号を加えた特許分類である．明細書には，特許文書の本文が段落ごとに記載されており，特許請求の範囲には，請求項が記載されている．

表 1　2019 年の特許データのセクションごとの特許数

セクション記号	セクションタイトル	件数
A	生活必需品	27,907 件
B	処理操作；運輸	29,902 件
C	化学；冶金	20,040 件
D	繊維；紙	1,612 件
E	固定構造物	6,146 件
F	機械工学；照明；加熱；武器；爆破	15,606 件
G	物理学	39,287 件
H	電気	40,053 件
合計		180,553 件

特許分類のセクションは，A から H の 8 つの分野に大別されている．セクションごとの統計量を表 1 に示す．その中でも，セクション C は（化学；冶金）分野となっており，他のセクションと比べ，物質とその特徴に関する記述が頻出すると考えられる．

本研究では，IPC か FI の主分類のセクションが C の特許と，セクション A の特許を扱う．セクションが C の特許は 20,040 件あり，セクション A の特許は 27,907 件ある．このデータを以後，**化学冶金データ**，**生活必需品データ**とそれぞれ呼称する．

5 物質の意外な用途の推定

物質の意外な用途を推定する手法を提案する．対象の物質の意外な用途を推定するためには，対象の物質と類似物質の用途の中で，類似物質のみが持つ用途を推定する必要がある．そのため，特許文書を用いて以下の手順で意外な用途の推定を行う．

（1） 対象の物質の用途の推定
（2） 類似物質とその用途の推定
（3） 対象の物質の意外な用途の推定

各処理に関しては以降で詳細に説明する．

5.1 対象の物質の用途の推定

対象の物質の意外な用途を求める際，その用途を推定する必要がある．そのため，対象の物質名と用途名を含む特許文から物質名と用途名の抽出を行った．

本研究では，物質の用途として化審法 [3]（化学物質の審査及び製造等の規制に関する法律）で用いられている用途を使用する．化審法とは，人の健康及び生態系に影響を及ぼすおそれがある化学物質による環境の汚染を防止することを目的とする法律であり，化学物質の用途ごとに環境への排出量が定められている．そのため，化審法では化学物質の用途が網羅的に定義されている．特許にも化審法で扱われている用途名がよく使われるため，物質の用途として化審法の用途を使用する．

まず，化学冶金データから特許文を抽出する．本研究では，特許文は特許の明細書から抽出する．このとき，特許文は対象の物質名と用途名を含む文を対象とした．例を図 2 に示す．対

1：https://www.jpo.go.jp/system/laws/sesaku/data/

2：https://www.jpo.go.jp/system/laws/koho/shiyo/document/kouhou_siyou_vol4-4/1-02.pdf

3：https://www.meti.go.jp/policy/chemical_management/kasinhou/files/ippantou/yusenyoto_2019fy.pdf

この無機充填剤は、シリカ、カーボンブラック、
用途名
炭酸カルシウム、水酸化カルシウム、マイカ、
水酸化マグネシウム、ケイソウ土、酸化チタン、
物質名
酸化亜鉛、酸化ビスマス、硫酸バリウム、タルク、
炭酸マグネシウム及びアルミナからなる群から選択
される１又は２以上である。

図 2　酸化チタンの意外な用途を求める際には物質名と用途名を含む文を抽出する

この無機充填剤は、シリカ、カーボンブラック、
用途名
炭酸カルシウム、水酸化カルシウム、マイカ、
水酸化マグネシウム、ケイソウ土、酸化チタン、
酸化亜鉛、酸化ビスマス、硫酸バリウム、タルク、
炭酸マグネシウム及びアルミナからなる群から選択
グループ内で物質名が列挙
される１又は２以上である。

図 3　マーカッシュ形式で書かれた特許文には物質が列挙され共通の特徴が書かれることがある

象の物質名と用途名を抽出するため，これらを含む文を特許文書から取り出した．本研究では，このような文を持つ化学冶金データの特許 1000 件を対象として文の抽出を行った．

特許文から物質名を抽出するため，固有表現抽出（NER）タスクでファインチューニングした BERT モデルを使用した．特許文内のマーカッシュ形式で記述された部分から物質名の抽出を行う．特許にはマーカッシュ形式という記述形式が使用されており，この形式は，グループの要素を設定し，グループ内の要素を選択できるように記載する記述形式である．このような記述形式は化学物質を扱う際によく使われているため，化学冶金データ内のマーカッシュ形式で記載された文には物質名が列挙されることが多い．例を図 3 に示す．グループの要素は図 3 のように読点などで区切られて列挙される．列挙された物質は共通の特徴や構造，用途を有している必要がある．そのため，マーカッシュ形式の列挙部分の付近を調べることで物質の用途を推定できると考えられる．また，物質の列挙部分に注目することで類似物質を推定することができると考えられる．類似物質の推定に関しては 5.2 節で述べる．マーカッシュ形式は，著者や列挙する物質によって多様な書かれ方がされており，物質名を抽出するには，ルールベースな手法より深層学習のモデルが適していると考えられる．このモデルを**物質抽出器**と呼称する．

物質抽出器の学習を行うため，学習用のデータセットを作成した．データセットとして，化学冶金データから 500 文を抽出したものを使用した．抽出には化合物名リストを利用した．これは，Wikipedia の化合物カテゴリ [4]の下位カテゴリに存在するページのタイトルを収集したものである．化合物カテゴリから 2 ページ以内のすべての下位カテゴリからタイトルを収集した．亜鉛だけでも酸化亜鉛，塩化亜鉛，臭化亜鉛などが収集されている．合計で 4765 語を化合物名として収集した．そして，化学冶金データを対象に，文中に含まれる化合物名の数ごとに 10 段階でランダムな 50 文を抽出した．化合物名の数が 0，1，2，3，4，5，6，7，8，9 以上の 10 段階で抽出した．化合物名の数で段階的に特許文を収集したのは多用な文を収集するためである．物質名が記載されていない文や物質名は記載されているが列挙されていない文，物質名が列挙されている文など，多用な文を収集した．抽出された全 500 文から文として成立しないものを除去すると，496 文のデータとなった．モデルの学習には，このデータを訓練：検証：テスト=8:1:1 に分割して使用する．このデータを以後**物質抽出データ**と呼称する．

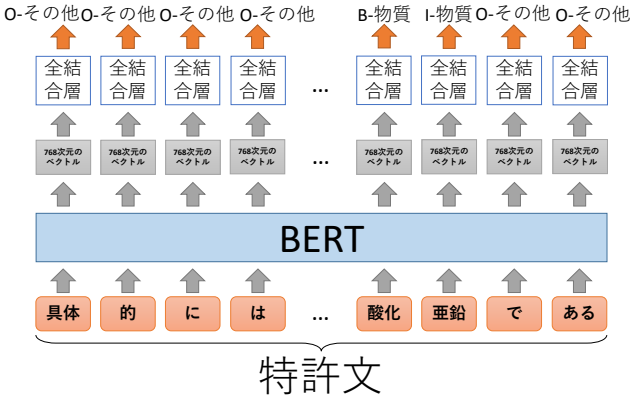

図 4　特許文から物質を抽出するためのモデル

物質抽出データには，文中の語句に物質ラベルを付与した．ラベルは，マーカッシュ形式で列挙されている語句に付与している．図 3 のような場合，シリカ，カーボンブラック，炭酸カルシウム，水酸化カルシウム，ケイソウ土，酸化チタン，酸化亜鉛，酸化ビスマス，硫酸バリウム，タルク，炭酸マグネシウム，アルミナにラベルを付与した．例外として，商品名や会社名にはラベルを付与しないものとする．

特許文から，物質名の抽出を行う．文から物質名を抽出する際，文をサブワードレベルでトークン化し，トークンごとに先頭，それ以降，それ以外のラベルの推定を行う．なお，サブワードに関してはその前のトークンと同様のラベルを付与する．そのため，トークンごとに 3 クラス分類を行うタスクと捉えることができる．

上記のタスクを解くためのモデルを作成した．このモデルの構造としては，BERT に全結合層を 1 層組み合わせたものを用いている．構造の概要を図 4 に示す．特許文を入力とし，サブワードレベルで分割を行った各トークンに対して，BERT と全結合層によってラベルの推定を行う．モデルの学習には物質抽出データを使用した．ハイパーパラメータを以下に示す．

4：https://ja.wikipedia.org/wiki/Category:化合物

- バッチサイズ ：2
- optimizer ：Adam [5]
- 損失関数 ：Cross Entropy Loss
- 学習率 ：2e-7
- 最大入力長 ：128

5epoch 以内に loss が下がらなければ早期終了する．このモデルを用いて特許文から物質名の抽出を行った．

そして，物質の用途は，物質名が特許文から抽出された際，同一文に含まれる用途とする．特許文から物質名が抽出された際，その特許文はマーカッシュ形式で書かれていることが想定され，マーカッシュ形式では，列挙部分の要素は共通の特徴や構造，要素を持っている必要があるため，その特許文に含まれる用途名が物質の用途として考えられる．そのため，物質名が抽出された同一文内の用途名を物質の用途とする．

5.2 類似物質とその用途の推定

対象の物質の類似物質を求め，その用途を推定する．類似物質は HITS アルゴリズムを用いて推定を行い，用途は，生活必需品データから抽出された特許文を用いて推定する．対象の物質と類似物質で用途の推定に用いるデータを変えることにより，異なる分野から意外な用途の推定を行う．そうすることで，(化学；冶金）分野からは見つけにくい意外な用途を推定することができると考えられる．

類似物質は，HITS アルゴリズムを用いて推定することができる．特許にはマーカッシュ形式という記述形式が使用されており，この形式では，語句が列挙された際，それらの語句は共通の特徴や構造，用途を有している必要がある．つまり，同一列挙により多く含まれた物質どうしは，特徴や構造，用途が多く共通しているため類似物質と考えられる．これは，同一列挙に含まれていたかどうかという間接的なつながりによって判断される．このような場合，HITS アルゴリズムを用いることで類似物質を推定することができる．

まず， 5.1 節で抽出された物質名とその物質名が列挙されていたグループの情報を利用して 2 部グラフを作成し，任意の物質に対して特徴が類似する物質を特定する．このようなグループのことを**物質列挙グループ**と呼称する．同一列挙から抽出された物質名を同じ物質列挙グループに属する物質名として扱う．グラフは下記のノード集合から構成される．以下では，任意の物質名を m，特許から抽出した物質名集合を M，物質列挙グループ集合を F，とする．また，枝については，物質名が属している物質列挙グループすべてに対して存在するものとし，枝集合を E とする．

- $F_m = \{f | f \in F, (f, m) \in E\}$
- $M_m = \{m' | m' \in M, (f, m) \in E, (f, m') \in E\}$
- $E_m = \{e | e \in E, e \in F_m * M_m\}$

このとき，グラフ $G' = ((F_m \cup M_m), E_m)$ を作成する．m を酸化チタンとした際の G' の例を図 5 に示す．

この G' に HITS アルゴリズムを適応することで，物質のノードにオーソリティ値，物質列挙グループのノードにハブ値を求めるものとする．このオーソリティ値により，特徴が類似する物質を求めることができる．$f_i \in F_m$，$m_j \in M_m$ のとき，語 f_i のハブ値を x_i，語 m_j のオーソリティ値を y_j とすると x_i と y_j の値は以下の式によって更新される．

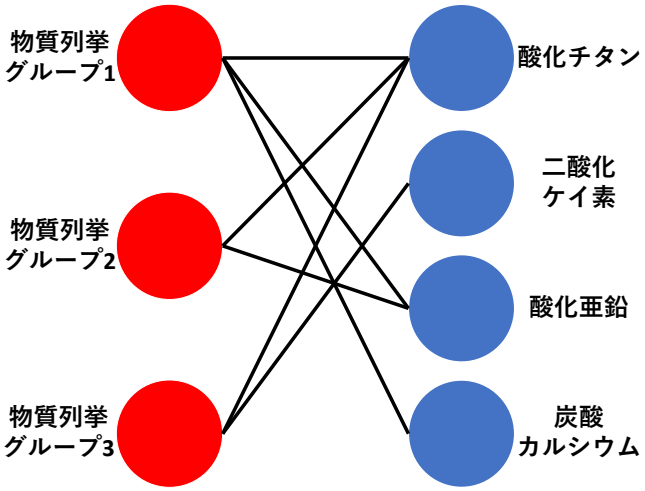

図 5　酸化チタンと特徴が類似する物質を特定する際に作成するグラフの例

$$x_i = \sum_{m_j \in M_m} y_j,\ y_j = \sum_{f_i \in F_m} x_i$$

本研究では，M_m の初期値を 1，それ以外を 0 とし，更新後に正規化を行う．値が収束するまで更新を行う．

例えば，図 5 のような酸化チタンを m としたグラフを得たとする．このようなグラフに HITS アルゴリズムを適用すると，酸化チタンと物質列挙グループをより多く共有し，多くの物質名と枝を持つ物質列挙グループをより多く共有している物質名を求めることができる．本研究では，このような物質を類似物質と仮定する．そのため，図 5 では，酸化亜鉛，炭酸カルシウム，二酸化ケイ素の順に酸化チタンと類似する物質といえる．

そして，類似物質の用途を推定する．本研究では，物質の用途は，物質が特許文から抽出された際，同一文に含まれる用途としている．生活必需品データを用いて 5.1 節と同様に抽出を行った．

5.3 対象の物質の意外な用途の推定

対象の物質の意外な用途を推定するためには，対象の物質と類似物質の既知の用途の差分を得る必要がある．本研究では，物質の意外な用途は，対象の物質と類似物質の既知の用途を比べた際，類似物質の用途のみに存在する用途と考えているため，5.1 節で推定した対象の物質の用途と 5.2 節で推定した類似物質の用途の差分を取ることによって求める．

また，意外な用途は，実際にその用途が役に立つかという実用性と，その用途が知られていないかという意外性の二つの観点から判断されるべきである．本手法では，対象の物質の既知の用途ではない用途を求めているため，後者の観点は考慮しているが，前者の観点は考慮出来ていない．そのため，意外な用途の役に立つ度合いに関するランキングを作成する．意外な用途と類似物質それぞれをノードとし，5.2 節と同様に 2 部グラフの作成と HITS アルゴリズムの適用を行う．枝については，

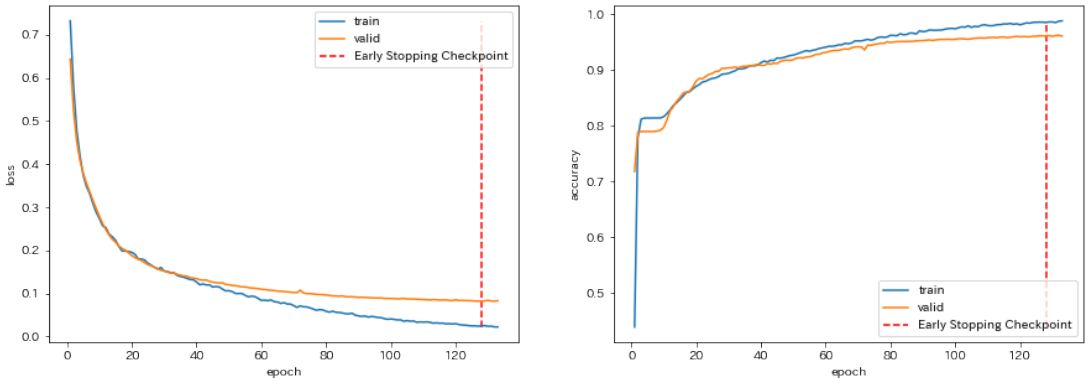

図 6　物質抽出器の学習曲線

類似物質のノードから類似物質が持つ用途のノードすべてに存在するものとする．類似物質のノードにはオーソリティ値，意外な用途のノードにはハブ値を求めるものとする．

上記のような 2 部グラフからハブ値を求めることで，実用性を推定することができると考えられる．ハブ値の高い意外な用途は，2 部グラフ内で多くの類似物質が持っている重要な用途であるため，対象の物質も持ちえる重要な用途となることが想定される．このような考えから，ハブ値の降順で意外な用途をランキングした．

6　実　　験

6.1　実 験 条 件

物質抽出データを用いて物質抽出器の評価を行った．ハイパーパラメータは 5.1 節のとおりである．文中から物質を抽出できているか物質名の完全一致で評価を行った．

対象の物質の類似物質が HITS アルゴリズムによって求められているか確認を行った．酸化チタンの類似物質を推定し，オーソリティ値が上位 6 件の類似物質を確認した．

物質の意外な用途を求め，その意外性を評価した．酸化チタン，炭酸カルシウム，シリカ，酸化亜鉛，硫酸バリウム，酸化ホウ素，酸化インジウムスズ，酸化グラフェン，水酸化バリウム，ステアリン酸アルミニウムの 10 種類の物質を対象として実施した．意外性の評価は，化学物質の専門家 1 名が 5 段階で行った．意外な用途のランキング上位の 10 件に対して評価を行った．

6.2　実 験 結 果

上記の条件で実験を行い，各物質で意外な用途を推定した．意外な用途の推定で使用した物質抽出器の学習曲線を図 6，HITS アルゴリズムを用いた類似物質の用途の推定結果を表 2，意外な用途の評価結果を表 3 に示す．

図 6 の学習曲線から，うまく学習が進んでいることがわかる．物質抽出器でテストデータ内の物質名を抽出した結果，物質名の完全一致の再現率は 0.775，適合率は 0.606 となった．

表 2 は，酸化チタンの類似物質を推定した結果であり，オーソリティ値が上位 6 件の物質を示している．オーソリティ値は，酸化チタンの値が最大となる．炭酸カルシウムやタルクといった酸化チタンと特徴が類似する物質のオーソリティ値が高くなっていることがわかる．炭酸カルシウムは，白色の顔料であるため，化粧品や塗料に使われており，酸化チタンと共通の用途で使われている．タルクも 1 節で述べたように酸化チタンと共通の特徴を多く持っているため，2 部グラフによって特徴が類似する物質をうまく求められていると考えられる．

表 2　酸化チタンを中心に求めたオーソリティ値が上位 5 件の類似物質

物質名	オーソリティ値
酸化チタン	0.06063
炭酸カルシウム	0.03356
タルク	0.03035
シリカ	0.02734
酸化亜鉛	0.02545
硫酸バリウム	0.02439

表 3 より，意外性の高い用途をいくつか推定できていることがわかる．大部分の用途は意外性の低い用途となっているが，意外性の高い用途がいくつか推定されている．例えば，酸化チタンに対して潤滑剤といった用途を推定している．酸化チタンは，一般的に顔料や充填剤として使われることが多い物質であり，潤滑剤は意外性が高い用途である．実際に潤滑剤という用途が抽出された特許文の確認を行った．特許文内に「潤滑剤は，存在する場合には，一般に，ステアリン酸マグネシウム，ステアリン酸またはタルクである．」と記載されていた．酸化チタンの類似物質であるタルクが記載されていることが確認できる．そのため，特許文からうまく意外な用途を推定することができていると考えられる．

7　考　　察

実験結果の分析を行った．表 3 の用途の意外性を見ると，ランキングが低くなるほど意外性が高まる傾向がわずかにみられた．そのため，ランキングごとに意外性の平均値を求めた．図 7 に散布図で示す．図 7 からもランキングが低くなるほど用

表 3　各物質に対して推定した意外な用途の意外性の評価

	酸化チタン		炭酸カルシウム		シリカ		酸化亜鉛		硫酸バリウム	
ランキング	用途	意外性	用途	意外性	用途	意外性	用途	意外性	用途	意外性
1	香料	3	金属イオン封鎖剤	1	抗菌剤	2	酵素	1	有機酸	1
2	乳化剤	1	浸透剤	2	芳香剤	1	芳香剤	2	乳化剤	1
3	潤滑剤	3	不織布	1	pH 調節剤	1	不織布	1	酵素	4
4	芳香剤	3	凍結防止剤	1	コーティング剤	1	合成ゴム	1	酸化剤	1
5	中和剤	1	消毒剤	1	金属イオン封鎖剤	1	金属イオン封鎖剤	1	モノマー	5
6	不織布	1	油剤	2	殺菌剤	2	乾燥剤	3	pH 調節剤	1
7	金属イオン封鎖剤	2	洗剤	1	浸透剤	1	中和剤	1	消臭剤	4
8	消毒剤	1	燃料	4	消臭剤	1	磁性材料	1	中和剤	1
9	磁性材料	2	バインダー成分	1	中和剤	1	漂白剤	1	芳香剤	4
10	pH 調節剤	1	展着剤	2	展着剤	1	展着剤	4	不織布	2

	酸化ホウ素		酸化インジウムスズ		酸化グラフェン		水酸化バリウム		ステアリン酸アルミニウム	
ランキング	用途	意外性	用途	意外性	用途	意外性	用途	意外性	用途	意外性
1	着色剤	1	顔料	4	顔料	2	無機酸	1	安定剤	1
2	色素	1	着色剤	4	着色剤	2	添加剤	1	充填剤	2
3	界面活性剤	2	染料	4	色素	2	界面活性剤	3	湿潤剤	1
4	染料	1	添加剤	1	添加剤	1	香料	1	可塑剤	1
5	安定剤	1	ワックス	5	染料	2	色素	1	キレート剤	1
6	香料	3	石鹸	5	プラスチック	2	無機アルカリ	1	バインダー	1
7	酸化防止剤	1	酸化防止剤	2	安定剤	1	安定剤	1	分散剤	1
8	有機酸	1	プラスチック	3	界面活性剤	3	着色剤	1	抗菌剤	3
9	プラスチック	4	安定剤	1	増粘剤	4	希釈剤	3	安定化剤	1
10	酸化剤	1	触媒	4	紫外線吸収剤	1	酵素	4	イオン交換樹脂	3

表 4　各物質に対して推定した意外な用途の意外性の平均値

	酸化チタン	炭酸カルシウム	シリカ	酸化亜鉛	硫酸バリウム
オーソリティ値が上位 25%以上の類似物質を用いた場合の意外性の平均値	2	1.8	1.2	1.4	2.2
オーソリティ値が上位 25%から 50%の類似物質を用いた場合の意外性の平均値	2.1	1.8	1.2	1.9	2.7

	酸化ホウ素	酸化インジウムスズ	酸化グラフェン	水酸化バリウム	ステアリン酸アルミニウム
オーソリティ値が上位 25%以上の類似物質を用いた場合の意外性の平均値	1.4	3.3	2	1.5	1.6
オーソリティ値が上位 25%から 50%の類似物質を用いた場合の意外性の平均値	2.2	3.4	2	1.6	1.5

途の意外性が高くなることが確認できる．このことから，実用性と意外性はトレードオフの関係があると推測でき，意外性の高い用途を得たい場合，実用性を下げればよいと考えられる．

実用性が低ければ意外性は高くなるのか，物質の意外な用途を推定する際，オーソリティ値が上位 25%以上の類似物質のみを用いた場合とオーソリティ値が上位 25%から 50%の類似物質のみを用いた場合で比較を行った．オーソリティ値が低い類似物質になるほど特徴が似ない物質であると考えられるため，オーソリティ値が低い類似物質を用いて意外な用途を推定した場合，実用性が低くなり，意外性が高い用途が推定されると予想できる．10 種類の物質を対象に，意外な用途の意外性を評価し，その平均値の比較を行った．結果を表 4 に示す．

表 4 から，オーソリティ値が 25%から 50%の類似物質を用いて意外な用途を推定した場合，10 物質中 6 物質でより意外性の高い用途を推定できていることがわかる．酸化チタン，酸化亜鉛，硫酸バリウム，酸化ホウ素，酸化インジウムスズ，水酸化バリウムの 6 種類の物質で，オーソリティ値が 25%から 50%

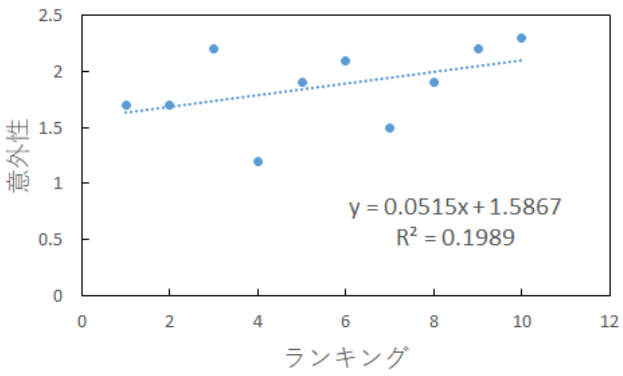

図 7　ランキングごとの意外性の平均値の散布図

の類似物質を用いたほうが意外な用途の意外性が高くなった．

8 まとめ

本研究では，物質の意外な用途を発見する手法の提案を行った．物質と類似物質の既知の用途の差分を得ることで意外な用途を得ることができるという考えに基づき手法を提案した．特許文書を対象として物質，特徴，用途に関する情報を抽出し，用途の推定を行った．

手法の流れとしては，まず，対象の物質の用途を推定した．化学冶金データから特許文を抽出し，文中のマーカッシュ形式の列挙部分から物質名を抽出した．同一文中に書かれた用途名を物質の用途として対象の物質の用途を推定した．次に，類似物質の用途を推定した．物質とその物質名を抽出した列挙部分の情報から 2 部グラフを作成し，HITS を適用することによって類似物質を求めた．生活必需品データの特許文から類似物質の用途の推定を行った．そして，対象の物質と類似物質の用途の差分を得ることで対象の物質の持たない用途を意外な用途として推定した．また，類似物質と意外な用途から 2 部グラフを作成することで，実用性の観点から意外な用途をランキングした．

結果として，物質の意外な用途を求めることができた．大部分の用途は意外性の低いものだったが，いくつか意外性の高いものを求めることができた．また，意外な用途の推定に使用する類似物質の違いによって意外性が変化することがわかった．オーソリティ値が低い類似物質を用いることで，意外性の高い用途を推定できると考えれられる．

今後の予定としては，実用性の評価を検討している．意外性の評価は行ったが実用性については評価されておらず，意外性と実用性の関係を推測するにとどまっているため，定量的な評価を行い，関係性を明らかにしたい．また，今回は用途として，化審法で定められた用途のみを使用しているため，もっと広い範囲の用途を扱った場合どうなるのかも検討したい．

謝　辞

本研究は JSPS 科研費 JP21H03775，JP21H03774，JP21H03554，ならびに，2021 年度国立情報学研究所公募型共同研究（21S1001）の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[2] Caspar J. Fall, Atilla Törcsvári, Karim Benzineb, and Gabor Karetka. Automated categorization in the international patent classification. *Special Interest Group on Information Retrieval Forum*, Vol. 37, No. 1, pp. 10–25, 2003.

[3] Kai Hakala and Sampo Pyysalo. Biomedical named entity recognition with multilingual BERT. In *Proceedings of the 2019 Workshop on BioNLP Open Shared Tasks*, pp. 56–61, 2019.

[4] Dylan Myungchul Kang, Charles Cheolgi Lee, Suan Lee, and Wookey Lee. Patent prior art search using deep learning language model. In *Proceedings of the 2020 Symposium on International Database Engineering & Applications*, pp. 1–5, Article No. 1, 2020.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computing Research Repository*, 2015.

[6] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. Vol. 46, No. 5, pp. 604–632, 1999.

[7] John D. Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 2001 International Conference on Machine Learning*, pp. 282–289, 2001.

[8] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, 2016.

[9] Jieh-Sheng Lee and Jieh Hsiang. Measuring patent claim generation by span relevancy. *Computing Research Repository*, 2019.

[10] Jieh-Sheng Lee and Jieh Hsiang. Patent classification by fine-tuning BERT language model. *World Patent Information*, Vol. 61, Article 101965, pp. 1–4, 2020.

[11] Xiaolei Lu and Bin Ni. BERT-CNN: a hierarchical patent classifier based on a pre-trained language model. *Computing Research Repository*, 2019.

[12] Fábio Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. Portuguese named entity recognition using BERT-CRF. *Computing Research Repository*, 2019.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 2017 International Conference on Neural Information Processing Systems*, pp. 5998–6008, 2017.

[14] 鈴木永史郎, 杉本徹. 意外性のある検索クエリの推薦方法の提案. 情報処理学会 第 78 回全国大会 全国大会講演論文集, pp. 503–504, 2016.

[15] 太田貴久, 南拓也, 山崎祐介, 奥野好成, 田辺千夏, 酒井浩之, 坂地泰紀. 特許文書を対象とした因果関係抽出に基づく発明の新規用途探索. pp. 1–4, 2018.

[16] 中村将人, 木村昌臣. 意外な検索キーワードを推薦する手法の提案. 情報処理学会 第 72 回全国大会 全国大会講演論文集, pp. 895–896, 2010.

[17] 佃洸摂, 大島裕明, 山本光穂, 岩崎弘利, 田中克己. 語の認知度と語間の関係の非典型度に基づく Wikipedia からの意外な情報の発見. 情報処理学会論文誌データベース（TOD）, Vol. 7, No. 1, pp. 1–17, 2014.