# 自然言語処理と Linked Data を用いた 化学物質情報の可視化

田中一成 [1,3] 岩倉友哉 [1,3] 小柳佑介 [1,3] 池田紀子 [1] 進藤裕之 [2,3] 松本裕治 [2,3]
{tanaka.kazunari, iwakura.tomoya, koyanagi.yusuke, nona}@jp.fujitsu.com,
{shindo, matsu}@is.naist.jp
株式会社富士通研究所 [1]
奈良先端科学技術大学院大学 [2]
国立研究開発法人 理化学研究所 革新知能統合研究センター[3]

## 1. 概要

化学物質(化合物)の知識は、新材料や新薬の開発、材料を用いた製品開発に必要不可欠である。これらの開発においては、様々な化合物の中から候補を選択し、試行錯誤を繰り返すことが行われる。そのためには、化合物についての調査が必要であるが、高度なスキルと多大な時間と労力が必要である。

たとえば、化学特許の調査を例にとると、1 件の特許に含まれる膨大な数の化合物名について、それらの構造を理解し、化合物の違いを把握する必要がある。そのために、様々なデータベースを参照するが、この作業はユーザにとって大きな負担になる。

このような調査に関する課題を解決するためには、テキスト中の化合物名の抽出および別称の同定を行い、各種データベースにナビゲートすることが考えられる。しかし、化合物名の抽出や別称の同定においては、既存のデータベースを基に作成される化合物名辞書を用いるだけでは十分ではない。

たとえば、最も大きな化合物 DB の一つである CAS [1]は、1 億以上の化合物の情報を保持している。しかしながら、分単位で増え続ける英語テキスト中の化合物情報に対し、人手で対応し続けており、多くの化合物情報が利活用されていない状況と考えられる。特に、CAS が対象としていない、日本語といった言語においては、活用されていない状況はより顕著である。

本論文では、従来は十分に活用されていなかった日本語テキスト情報を用いた化合物調査の支援を目的に、日本語処理技術で得られた知識と、既存のデータベースを Linked Data (LD)化した知識を統合し、化合物情報の可視化を行うことで化合物情報の活用支援のためのシステムについて紹介する。まず、化合物名抽出および化合物別称の同定について説明する。その後、可視化方法について説明する。

## 2. 化合物情報抽出

テキスト中の化合物名を抽出できれば、たとえば、化合物名の個所をハイライトし、化合物データベースへのナビゲートを行うといった支援が実現できる。しかしながら、全ての化合物名がデータベースに登録されている訳ではないため、未知の化合物名の抽出が必要になる。

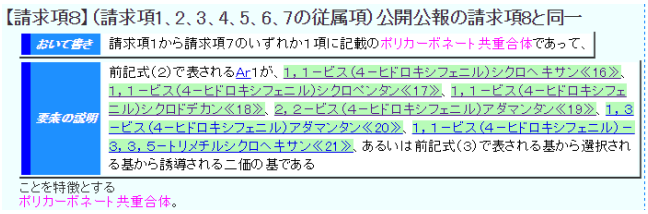
【請求項8】(請求項1、2、3、4、5、6、7の従属項)公開公報の請求項8と同一
おいて書き 請求項1から請求項7のいずれか1項に記載のポリカーボネート共重合体であって、
要素の説明 前記式(2)で表されるArが、1,1-ビス(4-ヒドロキシフェニル)シクロヘキサン≪16≫、1,1-ビス(4-ヒドロキシフェニル)シクロペンタン≪17≫、1,1-ビス(4-ヒドロキシフェニル)シクロドデカン≪18≫、2,2-ビス(4-ヒドロキシフェニル)アダマンタン≪19≫、1,3-ビス(4-ヒドロキシフェニル)アダマンタン≪20≫、1,1-ビス(4-ヒドロキシフェニル)-3,3,5-トリメチルシクロヘキサン≪21≫、あるいは前記式(3)で表される基から選択される基から誘導される二価の基である
ことを特徴とする
ポリカーボネート共重合体。

**図 1. 特許文書にリンクを付けた例**

未知の化合物名の抽出のために、本稿では、固有表現抽出の手法を用いる[2]。固有表現抽出では、前後の文脈や化合物名表記のパターンを利用することで、辞書に登録されていない化合物名の候補を抽出することが可能になる。

この学習には distant supervision の考え方を応用した[3]。 化合物名の辞書として、日化辞を活用した[4]。化学文書のうち、辞書により自動的にラベル付けした箇所を化合物抽出のための正例として利用する。また、化合物名が一般的には、含まれていないと思われるスポーツや政治の記事を負例として学習を行った。

化合物名が抽出することで、図 1 に示すような、特許文章中の化合物名のハイライトや、化合物の個所をクリックすることで関連情報へのナビゲートを行う。また、新規化合物名においては、後述する解析を経て、関連情報を提示するために用いられる。

また、化合物名称に加えて、機能・用途、代替物質の候補といった情報もテキストから抽出し、蓄積する。機能・用途とは、化合物が何に使われるかという情報で、例えば、可塑剤や界面活性剤といったものがある。代替物質の候補としては、例えば、可塑剤という用途において、「ジブチルフタレート」以外に「ジオクチルフタレート」も使われるといった情報である。

## 3. 化合物名別称の同定

抽出された化合物名がデータベースに登録されている場合には、化合物に関する情報を引き出すことができる。しかし、データベースに登録されていない化合物名、または、別称であれば、目的の情報が得られない。そこで、化合物名の別称を生成し、化合物を同定する。

今回は、日化辞のデータを利用して化合物名の言い換えルールを獲得した。獲得のためには、化合物名の命名規則である IUPAC 命名法 [5]に基づき、化合物名を分割し、部分構造単位で言い換えルールの抽出を行った。

獲得方法について、日化辞に登録されている、「アクリル酸 4-tert-ブチルフェニル」と、その別称である「アクリル酸 4-(1,1-ジメチルエチル)フェニル」を例に説明する。まず、「アクリル酸 4-tert-ブチルフェニル」を“アクリル酸”〃、“tert-ブチル”、“フェニル”に分割し、「アクリル酸 4-(1,1-ジメチルエチル)フェニル」を“アクリル酸”、“1,1-ジメチルエチル”、“フェニル“と分割する。続いて、共通部分を除き、残った、「tert-ブチル」と「1,1-ジメチルエチル」を言い換えルールとして獲得する。部分構造

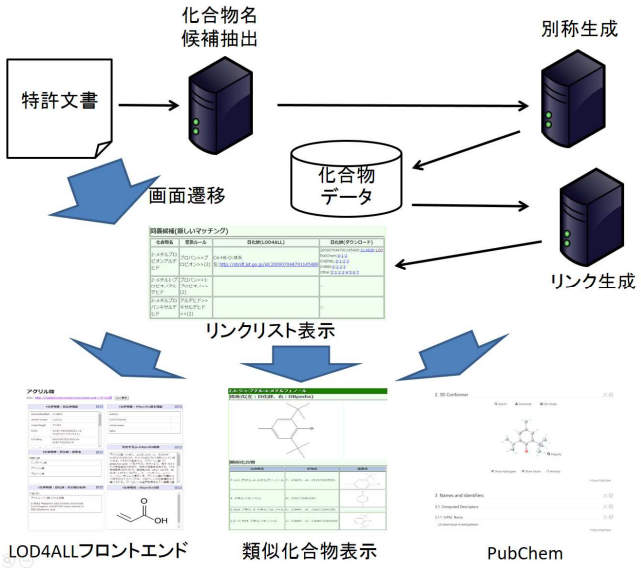


図 2：処理の流れ

への分割は、部分構造の辞書を用いて分割する。

このようにして抽出した言い換えルールを用いて、化合物名を置き換えることで、目的の情報にアクセスできる可能性を高める。

例えば、「2-(p-トリル)エタノール」という化合物名は登録されていないが、「2-(4-メチルフェニル)エタノール」に置き換えるとデータベースでヒットする。

## 4. 情報の統合・可視化

図 2 に文書から化合物の情報へナビゲートするための処理の流れを示す。文書から、化合物名抽出を行い、言い換えルールによって、別称の生成を行う。続いて、抽出された化合物名と生成された別称についてそれぞれ化合物データベースを検索する。データベースとしては、PubChem [6]や ChENBL [7]、日化辞を用いる。

化合物の情報を統合して表示するために、LOD4ALL フロントエンド [8]の仕組みを用いた。LOD4ALL フロントエンドでは、データを LD の表現形式である RDF で蓄積し、SPARQL によって検索し、表やグラフによって結果を表示することができる。テキストから抽出された化合物情報および、各種化合物データベースは、RDF で蓄積する。


図 3. LOD4ALL フロントエンドによる情報表示の例

図 3 に「アクリル酸」についての情報を LOD フロントエンドの仕組みを使って表示した例を示す。

化合物データベースの他にも DBPedia [9]に蓄積されているデータやテキストから獲得した知識も統合して表示する。

## 5. 未知の構造の可視化

言い換えを行うことで、別称が判明すれば、同一である化合物情報をデータベースから獲得可能となる。しかしながら、データベースに登録されていない化合物名は、その限りではない。そこで、化合物の構造の類似性を基に化合物の構造を表示することで、理解の支援を行う。

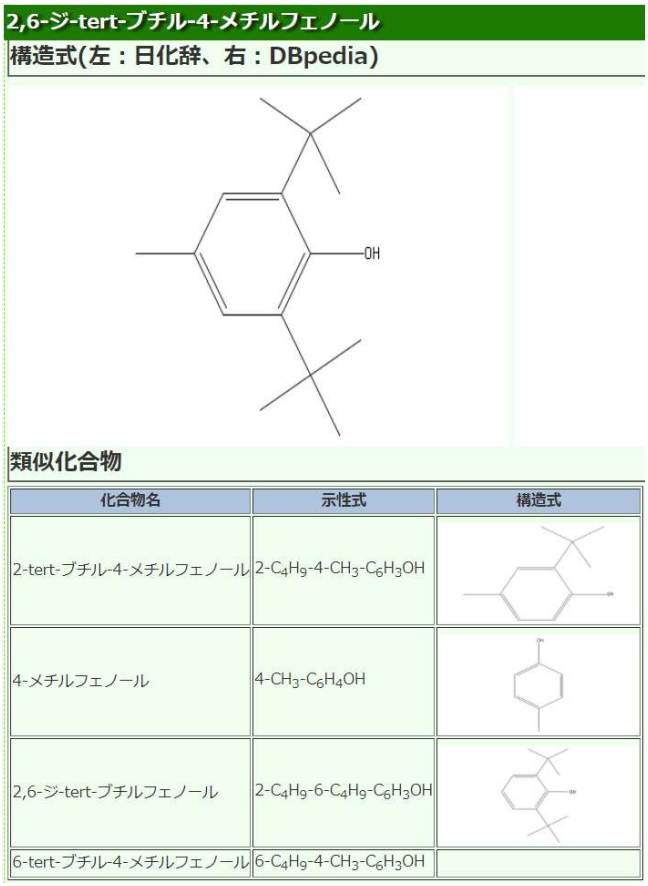


**図 4. 部分的に構造が一致する化合物を表示する例**

図 4 に示す「2,6-ジ-t-ブチル-4-メチルフェノール」の例では、化合物 DB に構造式が登録されているため、全体の構造式が表示されているが、それ以外に、類似化合物として、部分的に構造が一致する化合物の構造式を示している。これにより、部分構造を手がかりにして化合物の全体の把握に役立つと期待される。このような表示は、化合物名を解析して部分構造を明確にすることにより実現する。

化合物構造は置換基という部分構造の組み合わせからできており、化合物名、特に体系名では、置換基名の組み合わせからなっている [10]。そこで化合物名を置換基名毎に分解することで、化合物の構造が明らかになり、一部の置換基を除いて類似化合物の化合物名を生成することができる。置換基の組み合わせを変えることによって化合物データベースに登録されている化合物名と一致する可能性があり、この結果、化合物の構造を知る手がかりを得ることができる。

## 6. 多数の化合物名の可視化

化学特許には、図 5 に示すように多くの化合物名が羅列される場合がある。たとえば、材料の候補として、複数の化合物を挙げる場合がある。特許を理解するために、挙げられている化合物群の共通点や差分を読み解いていくには、高度なスキルと多大な時間が必要であり、実現は難しい。

このような場合にも、化合物名の解析を応用することができる。化合物名を部分構造の連なりを階層的に解析して分解し、化合物の核となる母核と置換基の関係を表形式に整理することで、多数の化合物の類似性を整理することができる。類似する部分構造をそろえて表示することで、共通点を見出し、全体を俯瞰することができる。

図 6 に表形式の表示の例を示す。この例では、11 個の化合物名が書かれているが、置換基の構造は大きく 2 パターンに分類され、それと、母核のバリエーションの掛け合わせでこれだけの化合物名になっていることがわかる。また、置換基の構造には 1 つだけ例外があり、本質的ではないものが混ざっていることも読み取ることができる。

## 7. 評価

特願 2014-263456 の特許のうち段落【0017】に出てくる 36 種類の化合物名について、データベースのナビゲートに関する評価を行った。結果を表 1 に示す。

36 種類の化合物名のうち、日化辞には 12 種類が登録されていた。登録されていなかった化合物名のうち、言い換えルールによって、日化辞に登

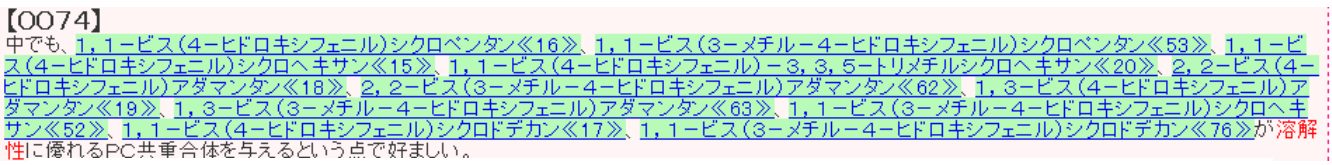
【0074】
中でも、1,1-ビス(4-ヒドロキシフェニル)シクロペンタン≪16≫、1,1-ビス(3-メチル-4-ヒドロキシフェニル)シクロペンタン≪53≫、1,1-ビス(4-ヒドロキシフェニル)シクロヘキサン≪15≫、1,1-ビス(4-ヒドロキシフェニル)-3,3,5-トリメチルシクロヘキサン≪20≫、2,2-ビス(4-ヒドロキシフェニル)アダマンタン≪18≫、2,2-ビス(3-メチル-4-ヒドロキシフェニル)アダマンタン≪62≫、1,3-ビス(4-ヒドロキシフェニル)アダマンタン≪19≫、1,3-ビス(3-メチル-4-ヒドロキシフェニル)アダマンタン≪63≫、1,1-ビス(3-メチル-4-ヒドロキシフェニル)シクロヘキサン≪52≫、1,1-ビス(4-ヒドロキシフェニル)シクロドデカン≪17≫、1,1-ビス(3-メチル-4-ヒドロキシフェニル)シクロドデカン≪76≫が溶解性に優れるPC共重合体を与えるという点で好ましい。

**図 5. 化合物名が羅列されている例**

| 化合物名 | 母核 | 結合位置 | 置換基数 | 母核炭素数 | 第1の置換基 | | | 第2の置換基 | | | 第3の置換基 | 第4の置換基 | 第5の置換基 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 母核 | 置換基1 | 置換基2 | 母核 | 置換基1 | 置換基2 | 母核 | 母核 | 母核 |
| 1,3-ビス(3-メチル-4-ヒドロキシフェニル)アダマンタン | アダマンタン | 1-,-3-, | 2 | 10 | フェニル | 3-メチル | 4-ヒドロキシ | フェニル | 3-メチル | 4-ヒドロキシ | | | |
| 2,2-ビス(3-メチル-4-ヒドロキシフェニル)アダマンタン | アダマンタン | 2-,-2-, | 2 | 10 | フェニル | 3-メチル | 4-ヒドロキシ | フェニル | 3-メチル | 4-ヒドロキシ | | | |
| 1,1-ビス(3-メチル-4-ヒドロキシフェニル)シクロドデカン | シクロドデカン | 1-,-1-, | 2 | 12 | フェニル | 3-メチル | 4-ヒドロキシ | フェニル | 3-メチル | 4-ヒドロキシ | | | |
| 1,1-ビス(3-メチル-4-ヒドロキシフェニル)シクロヘキサン | シクロヘキサン | 1-,-1-, | 2 | 6 | フェニル | 3-メチル | 4-ヒドロキシ | フェニル | 3-メチル | 4-ヒドロキシ | | | |
| 1,1-ビス(3-メチル-4-ヒドロキシフェニル)シクロペンタン | シクロペンタン | 1-,-1-, | 2 | 5 | フェニル | 3-メチル | 4-ヒドロキシ | フェニル | 3-メチル | 4-ヒドロキシ | | | |
| 1,3-ビス(4-ヒドロキシフェニル)アダマンタン | アダマンタン | 1-,-3-, | 2 | 10 | フェニル | 4-ヒドロキシ | | フェニル | 4-ヒドロキシ | | | | |
| 2,2-ビス(4-ヒドロキシフェニル)アダマンタン | アダマンタン | 2-,-2-, | 2 | 10 | フェニル | 4-ヒドロキシ | | フェニル | 4-ヒドロキシ | | | | |
| 1,1-ビス(4-ヒドロキシフェニル)シクロドデカン | シクロドデカン | 1-,-1-, | 2 | 12 | フェニル | 4-ヒドロキシ | | フェニル | 4-ヒドロキシ | | | | |
| 1,1-ビス(4-ヒドロキシフェニル)シクロヘキサン | シクロヘキサン | 1-,-1-, | 2 | 6 | フェニル | 4-ヒドロキシ | | フェニル | 4-ヒドロキシ | | | | |
| 1,1-ビス(4-ヒドロキシフェニル)シクロペンタン | シクロペンタン | 1-,-1-, | 2 | 5 | フェニル | 4-ヒドロキシ | | フェニル | 4-ヒドロキシ | | | | |
| 1,1-ビス(4-ヒドロキシフェニル)-3,3,5-トリメチルシクロヘキサン | シクロヘキサン | 1-,-1-,-3-,-3-,-5-, | 5 | 6 | フェニル | 4-ヒドロキシ | | フェニル | 4-ヒドロキシ | | -3-メチル | -3-メチル | -5-メチル |

**図 6. 多数の化合物名を表形式に整理した例**

録されている化合物名に変換できたものは 4 種類であった。言い換えルールを用いても構造情報が得られなかったもののうち 2 つは類似化合物の構造を得ることができた。

**表 1 ：評価結果**

| | |
|---|---|
| 出現した化合物名 | 36 |
| 直接日化辞でヒットしたもの | 12 |
| 言い換えルールによりヒットするようになったもの | 4 |
| ヒットしなかったもののうち類似化合物の構造が得られたもの | 2 |

この結果から、化合物データベースで全ての化合物を網羅することの難しさを確認できた。今回カバーできなかった「４，４’－ジヒドロキシ－３，３’－ジメチルフェニルエーテル」や「４，４’－ジヒドロキシ－３，３’－ジメチルジフェニルスルフィド」などを見ると、「ヒドロキシ」「メチル」「ジフェニルエーテル」「ジフェニルスルフィド」といった部分構造はデータベースに登録されていた。今回求めた類似化合物は部分構造を 1 つ除くというものだったため、これらは漏れてしまったが、部分構造から全体構造を推定するというアプローチでより多くの化合物をカバーできる可能性がある。一方で、元の化合物から離れるほど全体構造を推定しづらくなるため、図 4 の示性式のように化学式を生成して全体構造を推定しやすくするということも必要であると考えられる。

## 8.おわりに

化合物名を自動抽出して、別称の候補を生成するとともに、化合物名を解析することにより、類似化合物の情報を表示するシステムを構築した。これにより、化合物データベースに登録されていない化合物についてもユーザに情報を提供できる。また、異なる化合物情報を RDF で表現し、LOD フロントエンドの仕組みを利用することによって、統合し、表示を行う。言い換えルールや類似化合物名の生成により、化合物データベースをさらに効果的に活用できることが期待される。

今後は、テキストから、データベースには登録されていない化合物の物性情報など、より多くの情報を提示できる仕組みの構築を目指したい。

## 参考文献


[1]. CAS <http://www.cas-japan.jp/>

[2]. Tomoya Iwakura. A Named Entity Recognition Method using Rules Acquired from Unlabeled Data. Proc. of RANLP’11. Pp. 170—177.

[3]. Mintz, Mike and Bills, Steven and Snow, Rion and Jurafsky, Dan. Distant Supervision for Relation Extraction Without Labeled Data. Proc. Of ACL’09. pp. 1003—1011. 2009.

[4]. 日本化学物質辞書(日化辞) <http://dbarchive.biosciencedbc.jp/jp/nikkaji/desc.html>

[5]. IUPAC命名法<https://iupac.org/what-we-do/nomenclature/>

[6]. The PubChem Project <https://pubchem.ncbi.nlm.nih.gov/>

[7]. ChEMBLhttps://www.ebi.ac.uk/chembl/

[8]. LOD4ALLフロントエンド <http://lod4all.net/frontend/>

[9] DBpedia < http://wiki.dbpedia.org/>

[10] 池田紀子，田中一成: 特許文書からの化学物質情報の抽出，Japio YEAR BOOK 2015, p.274-281（2015）