特許検索における化学物質名の異表記同定に向けた考察

東京工業大学大学院情報理工学研究科准教授　藤井　敦

PROFILE
1998年東京工業大学大学院博士課程修了。博士（工学）。筑波大学大学院准教授等を経て、2009年より現職。自然言語処理、情報検索、Webマイニングの研究に従事。2009年度より特許版産業日本語委員会委員。

谷川国際特許事務所　田中　るみ子

PROFILE
1977年大阪大学理学部高分子学科卒業。2010年筑波大学図書館情報メディア研究科博士前期課程修了。三菱化成株式会社、東海大学等を経て、2006年より現職。特許事務、情報検索に従事。日本特許トランス株式会社兼務。

1 はじめに

化学物質には「構造式」、「結合表」、「名称」など多様な表現法がある。「構造式」は分子構造の図解であり、「結合表」は分子の結合に関する表形式の表現である。ここで、分子構造とは化学物質を構成する元素のつながり方である。例えば、メタンは「1個の炭素原子に4個の水素原子が単結合で結合した様式」という分子構造を持つ。図1と図2にメタンの「構造式」と「結合表」をそれぞれ示す。結合表は図2の一番左の図で示すように各原子に番号をつけ、右側の原子リストに番号と原子の種類を記載し、一番右側の結合リストに各原子に関する結合の種類を示す表である。メタンの名称は、「メタン」のほかに「メタンガス」や「R-50」がある。一般に名称には「体系名」、「慣用名」、「商品名」、「略称」など多様な表記がある。「体系名」は、化学物質の構造を示す表記であり、その指針として「国際純正および応用化学連合」（International Union of Pure and Applied Chemistry：略称IUPAC）が定めたIUPAC命名法がある。「慣用名」は、物質の出所や特性などを表わすラテン語や学名からつけられる。慣用名は、化学物質の構造とは関係がなく、体系名が現れる以前より用いられ、広く浸透している。そのためIUPAC命名法でも一部の慣用名に対しては使用を容認している。

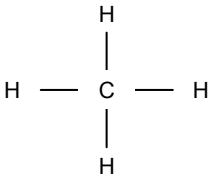
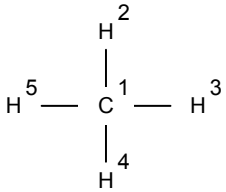


図1



原子リスト		結合リスト		
		1番目の原子	2番目の原子	結合の種類
1	C	1	2	単結合
2	H	1	3	単結合
3	H	1	4	単結合
4	H	1	5	単結合
5	H			

図2

現状では、化学物質の表記法は、以下の中から書き手が適宜選択している。

・慣用名 （例）ベンゼン
・体系名 （例）プロパン -1- オール
・体系名と慣用名の組合せ （例）体系名「メチル」と慣用名「安息香酸」の組合せによる「4 －メチル安息香酸」
・商品名 （例）体系名「2-(アセチルオキシ) ベンゼンカルボン酸」に対する「アスピリン」
・略称 （例）体系名「ジメチルスルホキシド（Dimethyl sulfoxide）」に対する「DMSO」
・番号 （例）CAS 登録番号「110-86-1」
・英語名 （例）caffeine
・分子式 （例）CH_3-COO-CH_3

化学物質が様々な表現法を持つことと書き手による恣意性によって、同一の物質に対し複数の名称が存在する。その結果、情報共有の阻害や情報検索における漏れの要因となっている。

この問題は化合物に関する辞書を利用することで部分的に解決することができる。化合物辞書の例として米国 NLM（米国立医学図書館）の PubChem や日本 JST 日本化学物質辞書 Web（日化辞）がある。表 1 に示すように、「酢酸エチル」に対して日化辞には「ビネガーナフタ」や「エタン酸エチル」など 15 通りの名称が定義され、PubChem では 113 通りの名称が定義されている。しかし、辞書による対応には以下に示すような限界がある。

・日々、新しい物質が作られる。
・表記に関する基準や方針が時代とともに変わる。
・後発医薬品によって商品名が増える。
・書き手が勝手に作る。

「書き手が勝手に作る」ことを示す例として、特許公報において「酢酸エチル」が「酢エチ」（特開 2003-212861）と表記されている。「酢エチ」という表記は日化辞には定義されていない。そこで、化学物質名を対象とした異表記問題を解決する情報処理技術が必要となる。

異表記問題を解決する一つの方法として、化学物質を構造で表現し、同定する方法がある。化学物質を正しい構造式に変換して比較することができれば、異表記問題は解決する。しかし、現状ではあらゆる名称を構造式に変換することは技術的に困難である。そこで、本研究は異表記の現象に着目し、なぜ異表記が発生するかという原因を化学的背景から分析し、異表記問題の本質を探ることを目的とする。

日化辞	PubChem
酢酸エチル	ethyl acetate
エチル=アセタート	Acetoxyethane
Ethyl acetate	Acetic ether
Acetic acid ethyl	Ethylacetate
ビネガーナフタ	Essigester
RCRA waste number U-112	acet-et-ester
エチルアセテート	Acetic acid ethyl ester

表 1

2 本研究の概要

本研究は、化学物質名の異表記問題を検討するため、まず、実際の文書に記載された化学物質名を抽出し、その記載が従来からある化学物質データベースの記載とどのように違っているかを分析する。ここで、「表層的な特徴」だけでなく「化学的な特徴」から分析し、この結果をもとに同定手法の開発に向けた指針について考察する。

分析対象の文書として、特許電子図書館から公開特許公報を検索して利用した。本稿では公開特許公報を「特許公報」と呼ぶ。特許公報中の【特許請求の範囲】には新規性や進歩性の判断対象となる化学物質発明について記載され、【実施例】には新規性・進歩性の判断対象ではないものの、化学物質に関する実施可能な例が示される。特許公報には、表記や翻訳に関する法的拘束がない。さらに、研究者、技術者、弁理士など書き手の属性が多様であるという特徴から物質名も多様になる傾向

がある。国際特許分類「C07D」を検索条件として指定した。「C07D」は複素環式化合物に関する物質名が多く記載されており、慣用名と体系名の組合せによって、名称が漸進的に増えるという特徴がある。

化学物質名の抽出および化学物質データベースの記載と比べる手がかりとして、物質に固有に付与されるCAS登録番号を用いた。CAS登録番号とは、米国化学会が策定し、2009年9月現在約5,000万件の物質を特定する番号である。特許公報ではCAS登録番号と化学物質名が併記される場合があることを利用して、物質名抽出の検索条件として「CAS登録」を指定した。

特許電子図書館における公報テキストでは、公報本文に「CAS登録」の記載があり、国際特許分類IPCが「C07D」であり、公報発行日が「20090917」（2009年9月17日）と指定して検索した。

この条件で検索された公報明細書の一部分を図3に示す。明細書中の「ジアミノジフェニルメタン（DDM；CAS登録番号101－77－9）」、「スルファニルアミド（SAA；CAS登録番号63－72－1）」、「パラフェニレンジアミン（PDA；CAS登録番号106－50－3）」、「ヘキサメチレンジアミン（HMDA；CAS登録番号124－09－4）」の記載を抽出した。

次に、特許公報から抽出した化学物質名とそれと同一物質である化学物質が既存のデータベースではどのように記載されているかを調査した。既存のデータベースとして日化辞を選択した。日化辞は無料で利用できる化学物質データベースで、2010年1月現在の収録件数は約279万件である。最新の文献から毎月約1万件の物質が追加されている。主なデータ源はJSTが作成・提供する文献データベースに収録されている原著論文などの文献で主題となっている有機低分子化学物質と化審法や労働衛生法など法律上の公示物質も収録している。文字列や化学構造から検索でき、分子式、分子量、CAS登録番号、法規制番号、体系名、慣用名、用途語などの化学物質情報が記載されている。「CAS登録番号」を検索項目として検索した日化辞の物質情報を図4に抜粋する。図4では、「体系名」や「慣用名」に記載されている名称が記載されている。

特許公報と日化辞データベースを用いて、同一物質でありながら記載が違う異表記対の集合（以下、「異表記コーパス」）を作成した（3章）。さらに、異表記が生じた原因を目視で分析し、異表記コーパス中の事例を類型化した（4章）。

体系名	4,4'-メチレンジ(アニリン) メチレンビス(p-フェニレン)ジアミン [メチレンビス(p-フェニレン)]ジアミン ジ(4-アミノフェニル)メタン ビス(4-アミノフェニル)メタン 4,4'-メチレンビス(アニリン) 4,4'-メチレンジアニリン [4,4'-メチレンビスアニリン] 4,4'-メチレンビス[ベンゼンアミン] 4,4'-メチレンビスベンゼンアミン 4,4'-メチレンビス(ベンゼンアミン) 4,4'-メチレンビスアニリン
慣用名	トノックス Tonox 4,4'-Methylenebisaniline 4,4'-Methylenebis(benzenamine) 4,4'-ジアミノジフェニルメタン 4,4'-Diaminodiphenylmethane 4,4'-Methylenedianiline Bis(4-aminophenyl)methane 4,4'-Methylenebis(aniline)

図4

【００３９】実施例３ツイン８ｅのジアミン化合物による硬化実施例１で得られた液晶ツインエポキシモノマー８ｅを、架橋剤としてジアミノジフェニルメタン（ＤＤＭ；ＣＡＳ登録番号１０１－７７－９）、スルファニルアミド（ＳＡＡ；ＣＡＳ登録番号６３－７２－１）、パラフェニレンジアミン（ＰＤＡ；ＣＡＳ登録番号１０６－５０－３）、ヘキサメチレンジアミン（ＨＭＤＡ；ＣＡＳ登録番号１２４－０９－４）を使用して硬化させ、その硬化物を製造するとともにそれらの性質を検討した。化学量論量のジエポキシモノマーと架橋剤を乳針で粉砕して、反応混合物を形成した。すなわち、ジアミン化合物は四官能であるので、反応混合比がジエポキシモノマー２モルに対してジアミン化合物１モルとなるよう混合した。反応混合物それぞれについてＤＳＣ、ＰＯＭを用いて予備硬化実験を行い、最良の硬化条件を定めた。

図3

3 異表記コーパスの作成

異表記コーパス作成にあたり、同一物質に対する異表記を特定することが非専門家には難しいという問題がある。表 1 に示した「酢酸エチル」と「ビネガーナフタ」が同一物質であることや、「酢酸エチル」と「酢酸メチル」が別の物質であることは非専門家にはわかりにくい。そのため、同一物質に共通に付与されている物質番号に着目し、物質名と物質番号が同時に記載されている特許公報を利用した。

本研究では CAS 登録番号を化学物質抽出の手がかりとして用いた。また、日化辞では検索項目に「CAS 登録番号」があるため、CAS 登録番号を手がかりに検索し、検索結果の名称一覧より、特許公報と異表記対になる名称を探すことができる。

例えば、図 3 に示した「・・・架橋剤としてジアミノジフェニルメタン（CAS 登録番号 101-77-9）、スルファニルアミド（CAS 登録番号 63-72-1）、・・・」のように CAS 登録番号「101-77-9」の記載があると、日化辞の検索項目で「101-77-9」と入力して日化辞の物質名を検索することができる。その結果、日化辞に「4,4'- メチレンビス [ベンゼンアミン]」、「4,4'- メチレンビスアニリン」、「4,4'- ジアミノジフェニルメタン」などの名称一覧を取得することができる。

そこで特許公報中の名称「ジアミノジフェニルメタン」の記載を調べ、あれば異表記コーパスの対象からはずし、なければ類似名称の「4,4'- ジアミノジフェニルメタン」を選んで異表記対とした。この作業を CAS 登録番号ごとに繰り返し行って異表記コーパスを作成した。

特許電子図書館において公報発行日が 1993 年 1 月から 2009 年 9 月まで、国際特許分類が「C07D」有機化学 複素環式化合物、「CAS 登録」が本文に含まれるという条件で検索を行い、公報 313 件を得た。この公報 313 件に記載されていた CAS 登録番号の異なり数は 978 であり、これをもとに異表記対 201 ペアを抽出した。

図 5 に作成した異表記コーパスの抜粋を示す。左の列が特許公報中の名称で右側が日化辞の名称である。

特許公報中の名称	日化辞の名称
ジアミノジフェニルメタン	4,4'-ジアミノジフェニルメタン
過ホウ酸ナトリウム・4 水和物	過ほう酸ナトリウム・4 水和物
Bay-u-3405	BAY-u-3405
2,3 ジヒドロキシブタン二酸塩	L-酒石酸塩
ドネペジル	塩酸ドネペジル
ロラカルベフ	L-ロラカルベフ
N-クロロこはく酸イミド	こはく酸 N-クロロイミド
α-グルコシルルチン	α-D-グルコシルルチン

図 5

4 異表記の類型化

作成した異表記コーパスについて以下のような手順で類型化した。図５を見ると、「過ホウ酸ナトリウム・４水和物」と「過ほう酸ナトリウム・４水和物」のように見てすぐ分かる表層的な違いによる異表記がある一方、「2,3 ジヒドロキシブタン二酸塩」と「L- 酒石酸塩」のように根本的に違う表記が混在している。まず表層的な特徴に着目し、「ギ酸」と「ぎ酸」のような表記的な違いを持つ対を選り分け、表層的には説明がつかない異表記対は化学的特徴に起因すると考え、試行錯誤しながら見直し、整理統廃合を行った。表記的特徴に類型化した中で、化学的背景を持つ場合は化学的特徴にも分類した。

次に（1）表層的特徴と（2）化学的特徴による類型化の例を説明する。スラッシュ「／」は異表記ペアの区切りを示す。

（1）表層的特徴

・異表記：かたかな、ひらがな、漢字、大小文字による違い。
（例）リン酸／りん酸、蟻酸／ぎ酸

・字訳：字訳基準摘要の違い。以下の例では、字訳基準に従えば「アセタート」が正しい。
（例）アセタート／アセテート

・翻訳：原語読みと翻訳語の違い。以下の例では chloride に対して「クロリド」が原語読みで「塩化」が翻訳である。
（例）クロリド／塩化

・略語：頭字語で略記。
（例）Diaminodiphenylmethane ／ DDM

・誤り：字訳の誤り。
（例）オルニバル／オルバニル

（2）化学的特徴

・命名方針：置換命名法や基官能命名法などによる違い。以下の例では、「1 －ブタノール」が置換命名法で「ブチルアルコール」が基官能命名法である。
（例）1 －ブタノール／ブチルアルコール

・位置番号：置換基や二重結合などの位置番号による違い。以下の例では、二重結合の位置を語の最初で示す「2- ブタジエン」と二重結合の前で示す「ブタ -2- エン」がある。
（例）2- ブタジエン／ブタ -2- エン

・立体表記：幾何異性、光学異性など立体表記を表わす記号の違い。以下の例は光学異性の表し方に関する違いで、絶対配置で示す「S-」と旋光性で示す「(+)-」がある。
（例）S- ロイシン酸／ (+)- ロイシン酸

・記号の使い方：記号の種類や有無の違い。以下の例では、「ブロモメチル」に対しする括弧の有無が違う。
（例）4- ブロモメチル - 安息香酸メチル／ 4-(ブロモメチル) 安息香酸メチル

・異表記：化学的に意味は同じであって表記が違う。以下の例では、「ベンゼン」に対する２置換基の位置が反対側であることを示す表記法が異なる。
（例）パラ／ p- ／ para

・慣用名：体系名と慣用名の違い。以下の例では、「2,3- ジヒドロキシブタン二酸」が体系名で「酒石酸」が慣用名である。
（例）2,3- ジヒドロキシブタン二酸／酒石酸

・説明語：化合物の種類を説明する言葉がない。以下の例では、「エステル」が該当する。
（例）フェニル - 酢酸フェネチルエステル／フェニル酢酸フェネチル

・記載順：部分名称の記載順が違う。以下の例では、「エトキシ」と「テトラフルオロ」の順序が違う。
（例）1- エトキシ -1,1,2,2- テトラフルオロエタン／ 1,2,2- テトラフルオロ -1- エトキシエタン

化学的特徴の類型化は化学命名法を参考にした。化学命名法の命名方針と名称のつけ方は命名法に関する解説書 1,2) を参考にし、字訳は日本化学会字訳基準 3) を参考にした。

5 おわりに

本研究は、主に特許検索の高度化を目的として、化学物質名の異表記対を集めてコーパスを作成し、異表記対の類型化を行った。この分析に基づいて、異表記同定手法を確立することが今後の課題である。

[参考文献]

1) 井藤一良 . 有機化合物命名のてびき . 化学同人, 1990.

2) 廖春栄 . 全有機化合物名称のつけ方 . 三共出版, 1999.

3) 日本化学会化合物命名小委員会 . 化合物命名法 補訂 7 版 . 日本化学会, 2000.