Vol.2010-FI-98 No.9
Vol.2010-DD-75 No.9
2010/3/5

化学物質名の異表記同定手法に関する考察

田中 るみ子† 藤井 敦††

化学物質は構造式, 結合表, 名称など様々な形式で表現することができ, その中で名称は体系名, 慣用名, 商品名, 略名など多様な表記を持つ. どの表記を使うかは書き手次第であり, 情報共有の阻害や情報検索における漏れが生じる. そこで, 化学物質名の異表記を同定する手法が必要である. 化学物質名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, 現状ではあらゆる物質名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. 本研究の方法は, 特許文書に記載された化学物質名の表記と化学物質データベースで定義された表記がどのように異なるかを分析し, 文字表記だけに着目した従来の方法とは異なる, 化学知識を加えた考察を行う. 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検討する.

A Study of Identification for Chemical Substance Names

Rumiko Tanaka† and Atsushi Fujii††

Chemical substances can be represented by various forms, such as structural formulas, connection tables, and names. Chemical substance names can be divided into various representation forms such as systematic names, common names, trade names, and abbreviations. Variants for a single chemical substance decrease the quality of exchanging and retrieving chemical information. To resolve this problem, an automatic method to identify chemical substance names is needed. If each chemical substance name can be converted into a unique structure formula, we can precisely compare two substance names based on their structures. However, existing tools cannot convert any substance names into their structure formulas with a sufficient accuracy. In this paper, we explore reasons for variants of chemical substance names, from spelling and chemical points of view. We use patent documents and a database for chemical substances to collect variants for different substances and analyze those variants. In addition, we discuss the possibility of realizing automatic methods for chemical substance names.

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それぞれの原子の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*1) 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.

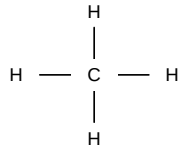
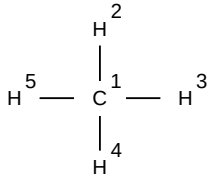


図 1 「メタン」の構造式



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 「メタン」の結合表

† 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba
†† 東京工業大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

*1 IUPAC Nomenclature of Organic Chemistry. http://www.acdlabs.com/iupac/nomenclature (2010.1.31 参照)

ⓒ 2010 Information Processing Society of Japan

Vol.2010-FI-98 No.9
Vol.2010-DD-75 No.9
2010/3/5

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ーメチル安息香酸」
・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*2や日本 JST 日本化学物質辞書 Web*3（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．

*2 PubChem. http://pubchem.ncbi.nlm.nih.gov (2010.1.31 参照)
*3 日化辞（日本化学物質辞書）Web. http://nikkajiweb.jst.go.jp (2010.1.31 参照)

表 1 日化辞と PubChem における「酢酸エチル」の名称例

日化辞	PubChem
酢酸エチル	ethyl acetate
エタン酸エチル	Ethyl ethanoate
エチル=アセタート	Acetoxyethane
Acetic ether	Vinegar naphtha
Ethyl acetate	Acetic ether
Vinegar naphtha	Ethyl acetic ester
Acetic acid ethyl	Ethylacetate
アセチックエーテル	Acetidin
ビネガーナフタ	Essigester
アセチジン	1-acetoxyethane
RCRA waste number U-112	acet-et-ester
Acetidin	Acetic acid ethyl ester
エチルアセテート	CH3-CO-O-CH3
Ethyl=acetate	141-78-6
Acetic acid ethyl ester	・・・
（15 通り）	(113 通り)

本研究の方法は，特許文書に記載された化学物質名の表記と化学物質データベースで定義された表記がどのように異なるかを分析し，文字表記だけに着目した従来の方法とは異なる，化学知識を加えた考察を行う．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検討する．

2 章では，化学物質名を対象とした情報処理技術について「化学物質名の抽出」と「化学物質名の同定」の観点から検討し，本研究の位置付けを明確にする．

2. 先行研究の検討と本研究の位置づけ

2.1 化学物質名の抽出

一ノ瀬ら [5]は，生化学関連の特許文書から化学物質名の抽出をする手法を提案した．化学物質名の部分名称（フラグメント）辞書，接頭語辞書，接尾語辞書，記号辞書，付加語辞書，単位辞書を作成し，辞書に登録される用語の組み合わせによって化学物質名を抽出した．

Klinger ら [2]は，IUPAC 名に準拠した化学物質名を抽出する手法を提案した．条件付き確率場（conditional random field: CRF）に基づく機械学習アプローチを用いてテキス

ⓒ 2010 Information Processing Society of Japan

トをトークン化し，ラベル付けを行うモデルを作成した．慣用名や商品名は辞書で対応するとして，この手法では扱わなかった．

Kemp ら [1]は，化学物質名を結合表などに変換する方法が注目されてきた反面，実際のテキストにある物質名を識別する有効性があまり問題にされてこなかったことを指摘した．化学物質名における文字列のつながりに着目し，化学名称のフラグメント（意味のある断片）辞書を作成して化学物質名の抽出に利用した．国際特許分類 IPC クラス「C07D」に関する 70 件の特許文書を対象に実験した結果，化学物質名 14,855 を抽出し，正解率は 97.4%であった．

本研究で対象とする異表記同定の前処理として，文書から化学物質名を抽出することが必要である．しかし，本研究では既存の抽出手法を利用することを想定し，異表記同定そのものに焦点を当てる．

2.2 化学物質名の異表記同定

同定手法に関する研究には文字列間の類似度を計算し，類似度が高い文字列どうしを同一物質の異表記と特定する研究がある．

Rhodes ら [3]は，化学物質の構造に関する類似度を調べる手法を提案した．具体的には，物質名を線形表記の InChI に変換する．さらに 2 つの InChI 表記どうしの類似度をベクトル空間モデルを使って計算する．ここで，線形表記とは構造を文字列で表した表記である．InChI とは，立体構造を含めた化合物の構造を一意的に記述することを目的に IUPAC が定めた線形表記である．

伊東ら [7]は，化学データベース検索において，化学物質名の異表記による検索を柔軟に行う方法について考察した．まず文字列の 1 文字以内の違いを容認する照合を行った．次に物質名称の性質から導き出した判別規則，例えば 1 文字違いでもその前 1 文字と後ろ 2 文字が一致するなどの規則を併用することによって精度を高めた．

Rhodes らの手法は物質名を構造に変換できない場合は適用することができない．吉川らは物質名をそのまま用いて，表層的な異表記だけを同定するため，表層的にあまり類似しない異表記には適用できない．本研究では構造変換ができない場合を想定し，さらに表層的な異表記に加えて化学的特徴を考慮して異表記同定問題を考察する．

3. 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*4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．特許公報は【明細書】，【特許請求の範囲】，【図面】，【要約】等から成り，【明細書】はさらに【技術分野】，【背景技術】，【発明の概要】，【発明が解決しようとする課題】，【課題を解決するための手段】，【発明の効果】，【発明を実施するための形態】，【実施例】等から成る．日本では，昭和 50 年法により物質特許が認められ，この場合【特許請求の範囲】には新規性・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．化学物質名の多様な記載が見られるため国際特許分類「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み合わせによって，名称が漸進的に増えるという特徴がある．

化学物質名の抽出及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検索の画面を図 3 に示す．ここで用いた検索項目は[公報本文]，[IPC]，[公報発行日]である．[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 4 に示す．明細書の中の「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は日化辞を選択した．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 万件の物質を追加している．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等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，

*4 特許電子図書館　http://www.ipdl.inpit.go.jp/homepg.ipdl (2010.1.31 参照)

ⓒ 2010 Information Processing Society of Japan

Vol.2010-FI-98 No.9
Vol.2010-DD-75 No.9
2010/3/5

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の一部分を図 5 に示す.「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.



図 3　特許電子図書館における公報テキスト検索の画面

【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 4　特許電子図書館における公報テキスト検索結果の例（特開平 09-118673）



図 5　日化辞における CAS 登録番号を用いた化学物質検索の例

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」）を作成する（4 章）. 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する（5 章）. 最後に，異表記同定手法に向けた考察を行う. 具体的には異表記コーパス中の事例を構造に変換し，構造変換できないもしくは構造が一致しない異表記対はどのように同定すべきかについて検討する（6 章）.

4. 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 に示す「酢酸エチル」と「ビネガーナフタ」と「エタン酸エチル」が同一物質であることや,「酢酸エチル」と「酢酸メチル」が別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 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.

ⓒ 2010 Information Processing Society of Japan

Vol.2010-FI-98 No.9
Vol.2010-DD-75 No.9
2010/3/5

例えば，図 4 の公報に「・・・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），・・・」の中に CAS 登録番号「101-77-9」の記載があると，日化辞の CAS 登録番号検索項目で「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」などの名称の一覧を取得できる．

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，なければ類似名称を一つ「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．

特許公報データの収集から異表記対の作成までのデータの流れを表 2 と表 3 に示す．表 2 に示すように，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（IPC）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．表 3 に示すように，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．

表 2　異表記コーパス作成用特許公報の検索条件

公報発行日	1993 年 1 月～2009 年 9 月
国際特許分類	C07D 有機化学 複素環式化合物
公報全文	「CAS 登録」が本文に含まれる

表 3　異表記対作成に関するデータの件数

公報件数	313
CAS 登録番号の異なり	978
異表記対	201

図 6 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．

5. 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 6 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」のように見てすぐわあ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ものを選り分け，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．

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
DDM	4,4'-Diaminodiphenylmethane
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
パラフェニレンジアミン	p-フェニレンジアミン
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 6　作成した異表記コーパスの抜粋

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

ⓒ 2010 Information Processing Society of Japan

・命名方針: 置換命名法や基官能命名法などによる違い．以下の例では表記が「1ーブタノール」が置換命名法，「ブチルアルコール」が基官能命名法である．
（例）1ーブタノール／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．以下の例では光学異性の表し方の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，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，「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

なお，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，名称のつけ方については命名法に関する解説書 [6,12]を参考にした．字訳については日本化学会字訳基準 [10]を参考にした．

6. 異表記同定手法に向けた考察

化学物質は，各成分元素がそれぞれ定められた元素間の結合，結合間隔，結合角によって三次元空間に配列された立体構造であり，構造に沿って名称がつけられたものが体系名である．化合物の表現については解説書 [11]を参考にした．そのため，まず異表記コーパス中の事例に対して名称を構造変換し，構造で比較した．次に構造変換できない，もしくは構造が一致しない異表記対はどのように同定すべきか検討した．本稿では名称から構造への変換を「名称構造変換」と呼ぶ．

本研究の対象は日本語物質名であるため，日本語物質名から構造への変換を検討した．荒木 [4]は日本語物質名を構造式へ変換する研究を行った．しかし，その手法は一般に利用される形でツール化されていないため使えない．そこで日本語物質名を図 7 に示す「化合物名の和英翻訳*5」を用いた．例えば，図 7 の和名欄に「ジアミノジフェニルメタン」と入力すると，「diaminodiphenyl methan(e)」と示される．「化合物名の和英翻訳」名称の翻訳は、「日本化学会の字訳基準」[10]に沿った逐語翻訳を行うため，次の後処理が必要であった．

・複数の接辞語候補からの選択
（例）メタン「methan(e)」→「methane」
・不要な空白を削除
（例）ジフェニルメタン「diphenyl methane」→「diphenylmethane」
・語順の入れ替えおよび必要な修正
（例）酢酸エチル「acetic acid ethyl」 → 「ethyl acetate」
・翻訳失敗への対応

後処理を行った後，名称構造変換を行った．

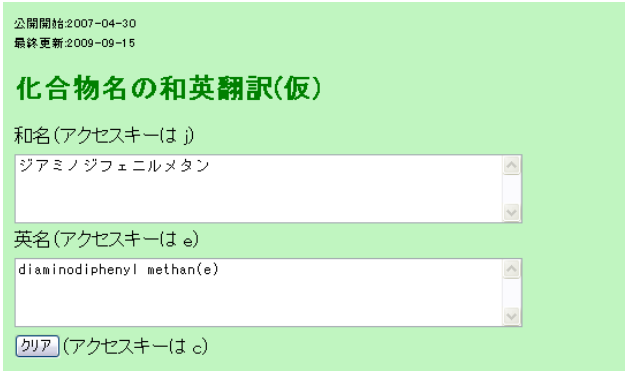


図 7 「化合物名の和英翻訳」の画面

*5 http://homepage1.nifty.com/nomenclator/chemjtra/chemjtra.htm (2010.1.31 参照)

ⓒ 2010 Information Processing Society of Japan

Vol.2010-FI-98 No.9
Vol.2010-DD-75 No.9
2010/3/5

名称構造変換に用いたツールは以下の通りである．

・Reaxys（リアクシス）*6
・OPSIN（オプシン）*7

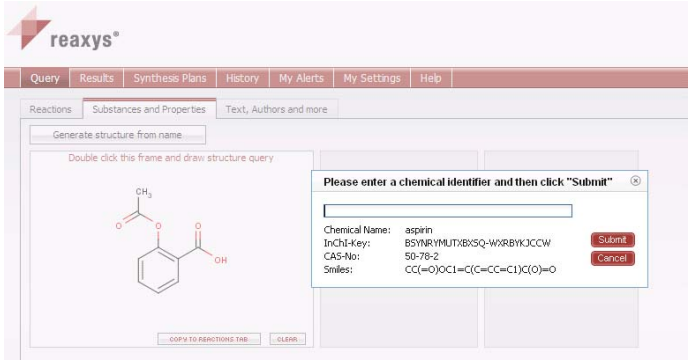


図 8　Reaxys で「aspirin」を名称構造変換した例

図 8 に Reaxys における名称構造変換「aspirin」の例を示す．右半分中央にある「Please enter a chemical identifier and then click "submit"」と書かれた入力ボックスに名称を入力すると左半分の画面に構造式が示される．異表記対 201 件の名称をそれぞれ人手で入力し，示された構造式を目視で比較した．OPSIN は，名称構造変換プログラム [8]に参考に異表記対 201 件の構造変換を行った．

異表記対 201 件のうち複数の分類に属する対は重複して計上したため異表記対の合計は 226 件となった．

表 4 に Reaxys，OPSIN の名称構造変換の実行結果を示す．「構造一致」は異表記対の両方が構造に変換でき，かつ両者の構造が一致した件数を表す．「構造一致」の例は「1,2,3,4-テトラヒドロ-キノリン-7-オール」と「1,2,3,4-テトラヒドロ-7-キノリノール」がある．「構造不一致」は構造変換ができたものの両者の構造が一致しなかった件数を表す．「構造不一致」の例は「2-エトキシ-1,1,1-トリフルオロエタン」と「2,2,2-トリフルオロエチルエチルエーテル」がある．「構造変換できない」は対のうち少なくとも片方が構造変換できなかった件数を表す．「構造変換できない」の例は「トリオルトトリルホスフィン」と「トリ-o-トリルホスフィン」である．

表 4 の合計の欄に示すように Reaxys において 2 つとも構造が一致した割合は 24%，変換できても一致しなかった割合は 11%，変換できなかった割合は 65%であった．

*6 https://www.reaxys.com (2009.11.30 参照)
*7 http://wwmm.ch.cam.ac.uk/wikis/wwmm/index.php/Oscar3 (2010.1.31 参照)

OPSIN はそれぞれ 4%，0%，96%であった．ただし，OPSIN の「構造不一致」は 0 件であったため，表に記載していない．この結果，名称構造変換できる割合が低く，名称構造変換に依存した物質名の同定手法は実現性に乏しいことがわかった．

表 4　名称同定に構造変換を適用した例

分類		Reaxys			OPSIN		合計
		構造一致	構造不一致	構造変換できない	構造一致	構造変換できない	
表記的特徴	異表記	2	1	1		4	4
	字訳	21	1	11	1	32	33
	翻訳	2	1	4		7	7
	略語	1	3	3		7	7
	誤り	1		4		5	5
化学的特徴	命名方針	7	8	49	2	62	64
	位置番号	3	1	5	1	8	9
	立体表記			10		10	10
	記号の使い方	12	2	30	3	41	44
	異表記	1		4		5	5
	慣用名		2	8		10	10
	説明語		5	4		9	9
	記載順	4	1	14	2	17	19
合計 (%)		54 (24)	25 (11)	147 (65)	9 (4)	217 (96)	226 (100)

仮に構造変換と類型の間に強い相関があれば，類型の特徴に対応した同定手法，例えば構造変換や他の手法の提案ができる．しかし，表 4 の結果からはそのような相関は得られなかったため，分類別ではなく全体として考察する．

表層的特徴か化学的特徴かにかかわらず文字列の類似があることを考慮すると，文字の類似度比較に柔軟に対応できる構成文字単位の N グラムが考えられる．N グラムとは連続した N 文字の単位で照合する手法である．例えば，「イルガフォス」と「イルガホス」を 2 文字単位で比較すると「イル」，「ルガ」，「ガフ」，「フォ」，「ォス」と「イル」，「ルガ」，「ガホ」，「ホス」で「イル」と「ルガ」が一致する．

また，化合物名称は化学的に意味のある部分名称に分けることができる．この特徴を利用することで部分名称単位で比較することができる．千原 [9]は，化合物に関する

ⓒ 2010 Information Processing Society of Japan

Vol.2010-FI-98 No.9
Vol.2010-DD-75 No.9
2010/3/5

情報検索の効率化のために名称の分割を行った．

(1) 英数字以外のかっこ，ハイフン，カンマをデリミタとして切り離す．
（例）２，４‐ジクロロ‐１‐ニトロベンゼン → ジクロロニトロベンゼン

(2) 化学的に意味のある単位毎に切り離す．
（例）ジクロロニトロベンゼン →「ジ」,「クロロ」,「ニトロ」,「ベンゼン」

(3) 分割して得られた「ジ」「クロロ」「ニトロ」「ベンゼン」のようなフラグメントでの並べ替えを行う．

(1)から(3)の手順で，名称フラグメントの並べ替えを行うことによって，名称の一致度を見ることができる．

また，名称フラグメントをNグラムの構成単位と考えたNグラムマッチングが考えられる．しかし，本研究では計算機上で異表記を同定する手法を具現化することはできなかった．

7. おわりに

本研究の成果は，異表記問題を検討するため，実際の文書から異表記対を集めコーパスを作成し，異表記対の類型化を行った点にある．残された課題は，コーパス作成用の収集データに偏りがないか検討することと，異表記同定手法を確立することである．

参考文献

1) Nick Kemp and Michael Lynch. Extraction of Information from the Text of Chemical Patents. 1. Identification of Specific Chemical Names. Journal of Chemical Information and Computer Sciences, Vol. 38, No. 4, pp. 544–551, 1998.
2) Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. Detection of IUPAC and IUPAC-like Chemical Names. Bioinformatics, 24(13), pp. i268-i276, 2008.
3) James Rhodes, Stephen Boyer, Jeffrey Kreulen, Ying Chen, and Patricia Ordonez. Mining patents using molecular similarity search. Proceedings of the Pacific Symposium on Biocomputing, Vol. 12, pp. 304–315, 2007.
4) 荒木啓介．化合物の名称解析と立体構造処理システムの研究．筑波大学博士論文, 1992.
5) 一ノ瀬桂子，廣田勇二，千原秀昭．特許公開公報から英文キーワードの自動作成．情報科学技術研究集会発表論文集, Vol. 34, pp. 109-115, 1998.
6) 井藤一良．有機化合物命名のてびき．化学同人, 1990.
7) 伊東靖史，吉川雅修，片谷教孝．化学データベースにおける名称検索の適合率の向上 (4)．全国大会講演論文集 情報処理学会第 52 回平成 8 年前期(4), pp. 263-264, 1996.
8) オープンバイオ研究会．オープンソースで学ぶバイオインフォマティクス．東京電機大学出版局, 2008.
9) 千原秀昭．化学構造検索の実用化に関する研究．昭和 55・56 年度文部省科学研究費補助金(試験研究 1) , 1982.
10) 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.
11) 船津公人監訳．ケモインフォマティックス．丸善, 2005.
12) 廖春栄．全有機化合物名称のつけ方．三共出版, 1999.

ⓒ 2010 Information Processing Society of Japan