研究論文

文章からの化学物質名を含む単語の認識法の確立と化学物質名の選択法の検討－特許公開公報を用いて

A Study for Recognition·Selection of Chemical Substance Name in Documents – Using Patent Documents

田中るみ子[1*]，中山伸一[2]
Rumiko TANAKA[1*], Shin-ichi NAKAYAMA[2]

1 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
〒305-8550　茨城県つくば市春日1-2
E-mail: s1630517@u.tsukuba.ac.jp
2 筑波大学図書館情報メディア系
Faculty of Library Information and Media Science, University of Tsukuba
〒305-8550　茨城県つくば市春日1-2
E-mail: nakayama@slis.tsukuba.ac.jp
*連絡先著者 Corresponding Author

化学物質名は，多様であり，記載法も書き手に委ねられているため自動抽出が難しく化学知識の共有化を妨げている．化学物質名を自動抽出できれば，共有化に役立つ．日本語の化学物質名を抽出するために，化学物質名をタグ付けした特許公開公報のコーパスの作成を行い，文章から形態素の切り出し，切り出した形態素の連結をすることによる化学物質名の認識法を確立した．あわせて連結した単語群から化学物質名を選択する方法の基礎的な検討，及び化学物質名と間違えやすい置換基名との選択比較も行った．

The chemical substance names described have various descriptions and the description of the name depends on the author. Such variation causes hindering information sharing of chemical knowledge. Auto-extraction of chemical substance names is useful for information sharing. In order to find a method for extracting the names of chemical substances in Japanese documents, we created a corpus of patent documents tagged with chemical substance names. We studied cutting out words from sentences and recognized chemical substance names by concatenating cut-out words using the part of speech information. We also studied selecting chemical substance names from concatenated cut-out words and made a selection comparison between chemical substance names and functional group names that are similar to chemical substance names.

キーワード: 化学物質名，化合物名，置換基名，特許公開公報，情報抽出
Keywords: chemical substance name, compound name, functional group, patent publication, information extraction

1 はじめに

私たちの生活の中で，多くの工業製品，医薬品，化粧品などが化学知識を使って作られている．これらの製品は，さまざまな有用な用途に用いられ，生活の向上に多大な寄与をもたらしている．また，化学知識を用いて解決すべき環境問題やエネルギー問題などがあり，化学知識は，問題解決のために欠かせない知識である．

ニュース，新聞記事，科学技術文献などには，さまざまな化学に関する文章があり，そこには化学知識が包含されている．それらは日々膨大な数が生産され，そこに記述される化学知識も増え続けている．化学が扱う問題を解決するには，化学知識を将来にわたって活用できるように効率よく抽出，整理，蓄積することが必要であるが，それには多大な時間と労力が必要である．そこで膨大な文章から化学知識の自動抽出が求められている[1].

化学知識は化学物質とその属性や構造，機能，製造方法，化学反応，用途などから成り立ち，その中核となるのが化学物質である．化学知識の抽出にはまず膨大な文章から化学物質を表す語句を抽出することが必要である[2].

化学物質の名称には，新しい名称が作られる，開発の時期により名称が変わる（医薬品など），命名法による表記のゆれがある，書き手により任意の表記が用いられる，カタカナ，記号，数字が多い，複合語が多いなどの特徴があるのと同時に，表記の仕方も名称，構造式，結合表，化学式，記号など多様である．例えば，「酢酸エチル」の表記には訳語 ethyl acetate（英訳），エチルアセテート（翻字），141-78-6（CAS 登録番号），$C_4H_8O_2$（分子式），$CH_3COOCH_2CH_3$（示性式），などがあり，単に酢エチと記載される場合もある．このように，同一物質でも表記が異なったり，同じ表記でも違う物質を表したりすることから，化学物質名の認識・選択は難しい問題となっている．

文章から化学物質名を抽出する研究は，英文に対する研究が先行している．その理由は，英文は単語をスペースや記号によって分けることができることから，単語の認識が容易であるためと考えられる．英文から化学物質名を抽出する方法は，ルールベースや機械学習，この 2 つを組み合わせたものなどがある．福田らは，物質名を構成する文字の特徴と周辺に現れる語句を手がかりにタンパク質名を抽出した[3]．Kemp らは，化学物質名フラグメントを備えた辞書を用いて，特許文書中の化学物質名を人手で作ったルールに基づいて抽出した[4]．Kemp らによれば，国際特許分類「C07D」の 70 件の特許明細書における特定の化学物質名 14 855 の 97.4%を抽出することができた．英文からの化学物質名抽出については，他にも医薬品開発に向けて，化学的性質や化合物の部分構造を含めて用語の抽出を行った研究[5] [6]や，化学物質名特有の特徴を見出し，機械学習モデルを使って物質名を抽出した研究[7]などがある．

海外では，化学物質名抽出のワークショップが活発に行われている．BioCreative (Critical Assessment of Information Extraction in Biology) IV ワークショップ（2014）[8] では PubMed のタイトルと抄録から，さらに，BioCreative V ワークショップ（2015）[9]では特許文書から化合物，遺伝子，タンパク質などの名称を，それぞれ抽出するタスク（CHEMDNER（Chemical Compound and Drug Name Recognition））が設定され，多く

の研究グループが参加した[10]. BioCreative V の CHEMDNER タスクには，21 チームが参加し, 最高点は再現率 0.91, 適合率 0.87, F 値 0.89 であった．BioCreative V で特許文書を対象としたのは，ノイズの多い特許文書から化学的及び生物学的データを抽出する方法を見出すことによって，ほかの種類の文書においてもその方法が役立つと考えられたためである．

一方，英文における研究と比べて，日本語の文章から化学物質名を抽出する研究はまだ少ない．日本語の場合は，化学物質名を選択する前に文章から単語を認識する必要がある．先行研究では，化学物質名を構成する文字種に着目し，カタカナ，英数，「酸」などの漢字，括弧などが連続したものを候補として認識した後，機械学習を用いて物質名としてふさわしいかどうか選択する抽出法が報告されている[11] [12] [13]. これらは主に命名法に基づく記載に適用されるため，慣用名を抽出するのは難しい．

日本語の文章から化学物質名を認識する方法として，形態素解析を用いる石川らの事例がある[14]．形態素解析は文を形態素という意味の最小単位へ分割し，各形態素へ品詞を付与し，各形態素を原型に復元するという 3 つの機能を持つ．しかし，一般的な形態素解析では，化学物質名は途中で分断され，化学物質名として正しく認識できない場合が多い．例えば，「1－（6－ブロモ－ピリジン－3－イルメチル）－4－エチル－ピペラジン」や「酸化第一銅」は「1/－/（/6/－/ブロモ/－/ピリジン/－/3/－/イルメチル/）/－/4/－/エチル/－/ピペラジン」や「酸化/第/一/銅」のように細かく形態素に切り出されるため，これらから化学物質名としてひとかたまりの単語にまとめることが必要である．石川らは，形態素解析ツール茶筌（ChaSen）を用いて形態素解析を行った後，出力された形態素の各品詞情報をもとに，形態素の品詞が「名詞，未知語，記号，接頭辞」で構成される形態素のまとまりを用語としている．このように化学物質名の認識には，品詞情報をもとに形態素をひとかたまりにする方法が考えられる．ただ，石川らの研究は化学物質名だけでなく，手段・効果の用語も含めた関連語を認識することを目的としており，認識の精度については不明確であり記述されていない．

形態素をひとかたまりの単語に認識した後，化学物質名の選択を検討しなければならない．その手掛かりとして，化学物質名を構成する文字の表記，化学物質名とその周辺に現れる単語，化学物質名を修飾する単語，化学物質名を含む文の意味など多様な方法が考えられる．

このように，形態素解析を介して日本語の化学物質名を抽出するには，文章からの形態素の切り出しと連結による化学物質名を含む単語の認識という段階と，得られた単語群から化学物質名を選択するという段階が必要である．本論では，これらの段階についてそれぞれ検討したので結果を報告する．

2 コーパスの作成

日本語文章中に出現する化学物質名の特徴分析や抽出結果の妥当性を検証するためには，化学物質名をタグ付けしたコーパスが必要である．その作成には，特許公開公報の化学分野の電子データを利用した．特許公開公報は進歩性, 新規性を明らかにし, 発明の知的財産権を表すのに有効な文書であり，他の科学技術文献には掲載されてい

ないデータを持つ重要な情報源である．先行研究でも BioCreative V などのように特許文書を対象にしている例が多く，また特許公開公報における化学物質名の記載は書き手に委ねられている要素が強く，表記が多様であるため，材料として適切であると考えた．さらに，特許公開公報の電子データはオープンソースであり，インターネット経由で容易にダウンロードが可能である[15].

Kemp らは 70 件の特許明細書を対象にしている[4]ことを参考に，本研究では 2016 年 7 月に公開された特許公開公報から，化学物質名が多く記載されていると考えられる国際特許分類「C 化学；冶金」に該当する公報 50 件（総文字数：1120210）を取り出し，タイトル，抄録，本文に記載されている化学物質名に対してタグ付けを行った. タグ付けを行った化学物質名は構造が明確な単一物質, 複合物質, ポリマー, 混合物, 及び部分的に構造が明確な物質群であり，さらに商品名も含む．形式的には分子式, 示性式のタグ付けは行ったが，記号の連なりである CAS 登録番号などの記号番号には行っていない．また特許公報に頻出の RNH_2 のようなマーカッシュ形式は後述の置換基としてタグ付けを行った.

化学物質名の記載は IUPAC（International Union of Pure and Applied Chemistry）が定める，化合物の体系名の命名法への準拠が一般に推奨されている．命名法は中心となる母体（環を含むと母核と呼ばれることもある）化合物の水素を置換基で置き換えた誘導体として命名される．このため，命名法に基づいた化学物質名は，置換基名を含んで成り立っている．例えば塩化ビニルは化学物質名であるが，ビニル基（CH_2=CH-）は置換基名である．化学物質名の名称が命名法への準拠が推奨されることを考慮すると，化学物質名の抽出の際に置換基名が混在する影響が考えられる.

そこで化学物質名を選択する段階で置換基名の影響を検討できるように，コーパスの作成においては，化学物質名にタグ付け<chem>すると同時に，置換基にもタグ付け<group>を行った. なお複数の名称をまとめて記載した「カルシウムアルミネート及び／又はアルカリ金属アルミン酸塩」のような場合，「及び」，「又は」，「及び／又は」の前後で区切ってタグ付けした．タグ付けの例を図 **1**に示す.

代表的なものとしては, 例えば<chem>ビニルシラン</chem>等の（Ｃ１）<group>アルケニル基</group>を有する<chem>珪素含有化合物</chem>と，例えば<chem>ヒドロシラン</chem>等の（Ｃ２）<group>ヒドロシリル基</group>を含有する<chem>珪素化合物</chem>とを総<group>ヒドロシリル基</group>量が０．５倍以上，２．０倍以下となる量比で混合し，（Ｃ３）<chem>Ｐｔ</chem>触媒などの付加縮合触媒の存在下反応させて得られる<group>Ｓｉ－Ｃ－Ｃ－Ｓｉ結合</group>を架橋点に有する化合物等を挙げることができる.

図 1　タグ付け

各特許公開公報に何個の化学物質名と置換基名があったかは後述の表 3に示す．今回タグ付けした公報は国際特許分類「C 化学；冶金」に属しているが，化学物質名がほとんど記載されない装置，製造，物流などの分野もあり，化学物質名タグが付与されない公報もあった．全体として，<chem>タグは 15834 個，<group>タグは 2991 個付与された.

3 化学物質名を含む単語の認識

3.1 方法

日本語の文章から化学物質名を抽出するためには，まず文章を単語単位に認識する必要がある．この単語単位の認識は形態素解析ツールを用いて行う．形態素解析ツールには JUMAN，茶筌(ChaSen)，MeCab などがあるが，より改良され高速になった MeCab を用いることとする[16].

MeCab を用いると，「ジメチルアミノエチルメタアクリレート塩化メチル 4 級塩の重合物が広く使用されている.」は表 **1**のように形態素に認識され，入力文字（表層形）に続き，品詞，品詞細分類 1，品詞細分類 2，品詞細分類 3，活用形，活用型，原形，読み，発音が表示される．

表 1　形態素解析の例

ジメチルアミノエチルメタアクリレート　名詞,一般,*,*,*,*,*
塩化　名詞,サ変接続,*,*,*,*,塩化,エンカ,エンカ
メチル　名詞,一般,*,*,*,*,*
4　名詞,数,*,*,*,*,*
級　名詞,接尾,助数詞,*,*,*,級,キュウ,キュー
塩　名詞,一般,*,*,*,*,塩,シオ,シオ
の　助詞,連体化,*,*,*,*,の,ノ,ノ
重合　名詞,サ変接続,*,*,*,*,重合,ジュウゴウ,ジューゴー
物　名詞,接尾,一般,*,*,*,物,ブツ,ブツ
が　助詞,格助詞,一般,*,*,*,が,ガ,ガ
広く　形容詞,自立,*,*,形容詞・アウオ段,連用テ接続,広い,ヒロク,ヒロク
使用　名詞,サ変接続,*,*,*,*,使用,シヨウ,シヨー
さ　動詞,自立,*,*,サ変・スル,未然レル接続,する,サ,サ
れ　動詞,接尾,*,*,一段,連用形,れる,レ,レ
て　助詞,接続助詞,*,*,*,*,て,テ,テ
いる　動詞,非自立,*,*,一段,基本形,いる,イル,イル
.　記号,句点,*,*,*,*,. ,. ,.

形態素から化学物質名を単語として認識する方法を検討するため，作成したコーパスの中から<chem>でタグ付けした化学物質名だけを取り出し，そこに含まれる各形態素の品詞を調べた．化学物質名に出現していた品詞ののべ語数と事例，出現箇所を表 **2**に示す.

表 2　品詞の割合と例

品詞	語数	事例	出現箇所
感動詞	7	Km ウン	ウンデカン
記号	13062	－β LCH＝ XQNR・ZOFBI （～） ［p’］…	
形容詞	0		
助詞	28	デ ノ へ べ ん	ヘキサデシル
助動詞	0		
接続詞	0		
接頭詞	306	第 不 過 重…	
動詞	0		
副詞	26	フッ ジーン	フッ化白金
名詞	42940	亜砒酸 フルオロウラシル ホルムアミド …	
連体詞	0		
総合計	56369		

化学物質名に付与された品詞は感動詞，記号，助詞，接頭詞，副詞，名詞の 6 種類であった．化学物質名を構成する要素として記号，接頭詞，名詞は妥当であるが，表 2 の事例に見られるように「ウンデカン」の「ウン」が感動詞として，「フッ化白金」の「フッ」が副詞として解析されていることなどから，感動詞，助詞，副詞は誤って付与されたと考えられた．その原因は形態素解析ツールのシステム辞書にウンデカン，フッ化などの形態素がないためであると考えられる．

そこで，基本的な化学物質名の形態素を含むユーザー辞書を作成することとした．日本化学物質辞書 Web（日化辞 Web）[17] から RDF データ NBDC_NikkajiRDF_main.tar.gz をダウンロードし，ndl:transcription のタグのついた例えば，"4-(トリ#イソ#プロピル#シリル#オキシ#)フラン#-2-カルボ#アルデヒド#"のデータを#で分割すること

により，トリ，イソ，プロピル，シリル，オキシなどを形態素とし，最終的に 15654 の形態素からなるユーザー辞書を作成した．このユーザー辞書を用いて，化学物質名だけを形態素解析すると，記号 59 種類，接頭詞 20 種類, 名詞 3106 種類という結果が得られた．誤って付与された品詞がないことが確認されたことから化学物質名を構成する形態素の品詞は記号，接頭詞，名詞であると考え，記号，接頭詞，名詞が連続していた場合，それらを連結して一つの単語とすることとした．

なお記号は「ー」や「()」のように，化学物質名の一部として使われるものもあれば，「。」や「、」のように化学物質名の一部にはならない句読点もある．間に「、」が入ることもある．そのため，連結した後に，単語の先頭や後尾に「、酸化第一銅」のように句読点がある場合，これらの句読点を連結後に削除することとした．

3.2 結果及び考察

上記の方法により認識された単語群を「Z」とし，単語群には予めタグ付けした化学物質名群を「C」，<group>でタグ付けした置換基名群を「G」とすると，「C」と「G」はそれぞれ「Z」にすべて含まれていることを確認した．このように,ユーザー辞書を用いた形態素解析結果を用いて，特定の品詞を連結する後処理を行うことによりすべての化学物質名を単語として認識することができた．表 **3**に公報ごとの全単語群(Z)の数, chem (C)の数, group (G)の数, C/Z, G/Z を示す. 今回対象とした 50 公報の単語のうち，全体として一割弱の語が化学物質名であった．

表 3　公報毎の単語数及びタグの付与頻度

公報番号	全単語群(Z)の数	chem (C)の数	group (G)の数	C の数/Z の数	G の数/Z の数
2016129515	11932	139		0.012	0.000
2016129862	1328	49		0.037	0.000
2016129882	3363	490	29	0.146	0.009
2016129977	736	8		0.011	0.000
2016130183	1887	287		0.152	0.000
2016130193	6672	788		0.118	0.000
2016130203	1084	242		0.223	0.000
2016130213	1513	272		0.180	0.000
2016130234	1854	96		0.052	0.000
2016130249	11266	910	1082	0.081	0.096
2016130271	4712	1041	164	0.221	0.035
2016130281	3761	533	55	0.142	0.015
2016130291	7473	1691	150	0.226	0.020
2016130301	1505	345		0.229	0.000
2016130311	2188	209	1	0.096	0.000
2016130321	5126	441	94	0.086	0.018
2016130331	1880	0		0.000	0.000
2016130341	2256	294		0.130	0.000
2016130351	776	34		0.044	0.000
2016130361	1832	179		0.098	0.000
2016130372	1960	221	3	0.113	0.002
2016130783	1184	198	34	0.167	0.029
2016130861	3260	74	13	0.023	0.004
2016131193	1145	8		0.007	0.000
2016131244	4058	383	152	0.094	0.037
2014020939	3556	411		0.116	0.000
2014021084	2831	247		0.087	0.000
2014021205	2916	147		0.050	0.000
2014021257	1909	294		0.154	0.000
2014021316	4116	88		0.021	0.000
2014021351	6185	813	189	0.131	0.031
2014021388	5968	684	10	0.115	0.002
2014021419	2501	332	276	0.133	0.110
2014021459	2747	170	2	0.062	0.001
2016521114	11677	188	33	0.016	0.003
2016521125	2126	115		0.054	0.000
2016521195	4535	45		0.010	0.000
2016521222	3583	264	16	0.074	0.004
2016521241	1312	268		0.204	0.000
2016521251	4718	545	494	0.116	0.105
2016521262	5107	420	2	0.082	0.000
2016521295	9508	835	50	0.088	0.005
2016521305	1998	319	34	0.160	0.017
2016521316	3548	36		0.010	0.000
2016521374	2986	331	91	0.111	0.030
2016131495	1203	75	3	0.062	0.002
2016131516	1582	2		0.001	0.000
2016131541	2420	24		0.010	0.000
2016131902	3418	200	14	0.059	0.004
2016131932	2403	49		0.020	0.000
合計	179604	15834	2991	0.088	0.017

4 文字に注目した化学物質名の選択方法及び化学物質名と間違えやすい置換基名との選択比較

3.で取り出した単語群には化学物質名とそうでないものが混在している．そこから

化学物質名のみを取り出す方法として，化学物質名は命名法に基づいた記載が多いことから，カタカナ，記号，数字，限られた漢字から構成されることに着目した．そこで化学物質名を構成する文字の出現頻度が認識する手がかりにならないかと考え，試行的に化学物質名を構成する 1 文字(1-gram)の出現頻度による方法を検討した．

4.1 方法

タグ付けした化学物質名群「C」に含まれる一文字ののべ出現頻度（n）とこの方法で得られた全単語群「Z」に含まれる一文字ののべ出現頻度（N）をそれぞれ数え，その比率（n/N）を求める．n/N が 1 以上，0.9 以上，0.8 以上，0.7 以上，0.6 以上，0.5 以上に該当する文字のリストを作成する．n/N が高い文字が化学物質名に特有な文字と考え，その文字が含まれる語句をそれぞれ「Z」から取り出し， n/N により<chem>でタグ付けした化学物質名の選出がどのように変化するかを調べることとした．

また誤って置換基名が化学物質名として選択されてしまう影響を検討するため，<group>でタグ付けした置換基名が各取り出し条件においてどの程度含まれるかを調べることとした．

全単語群 Z と n/N で取り出した単語群 R との関係を図 2に示す．

Z: 全単語群
C: 全 chem 群
G: 全 group 群
O: それ以外
R: 取り出された単語群
c: 取り出された chem 群
g: 取り出された group 群
o: 取り出されたそれ以外

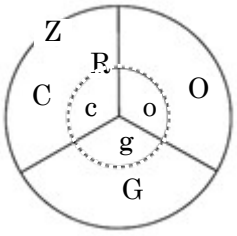

図 2 タグ付けと抽出との関係

全単語群「Z」には，化学物質名群「C」と置換基名群「G」，それ以外の「O」が含まれる．一文字の n/N 別に取り出された単語群「R」は「Z」の部分集合であり，そこには化学物質名 c と置換基名 g とそれ以外の o が含まれる．n/N により，化学物質名と置換基名の適合率，再現率がどのように変動するかを調べることにより，1-gram による方法の化学物質名選出の可能性の検討を行うとともに，置換基名の化学物質名選出への影響を検討することとした．

4.2 結果及び考察

タグ付けした化学物質名群「C」と全単語群「Z」に含まれる一文字の頻度の n/N 別の文字を表 **4**に示す．n/N 別に該当する文字が含まれる語句を抽出し，抽出された単語の数（R_n）と抽出された単語群中の化学物質名の数（c_n）を数え上げた．タグ付けされた全化学物質名数 C_n (15834 個)をもとに，再現率 recall(c_n/C_n)，適合率 precision(c_n/R_n)，F 値を求めた結果を表 5に，再現率と適合率の関係のグラフを図 3にそれぞれ示す．

表 4 化学物質名に特有な文字

n/N	個数	文字
1	12	ω|苛|吉|錫|酒|藻|弗|没|酪|砒|硼|蔗
0.9	16	ω|苛|吉|錫|酒|藻|弗|没|酪|砒|硼|蔗|酢|´|灰|硝
0.8	21	ω|苛|吉|錫|酒|藻|弗|没|酪|砒|硼|蔗|酢|´|灰|硝|-|ゾ|硫|六|窒
0.7	29	ω|苛|吉|錫|酒|藻|弗|没|酪|砒|硼|蔗|酢|´|灰|硝|-|ゾ|硫|六|窒|亜|ホ|錯|珪|黄|－|O|α
0.6	49	ω|苛|吉|錫|酒|藻|弗|没|酪|砒|硼|蔗|酢|´|灰|硝|-|ゾ|硫|六|窒|亜|ホ|錯|珪|黄|－|O|α|メ|芳|エ|ビ|息|ピ|，|燐|土|'|ェ|｛|チ|ボ|炭|素|｝|ニ|ジ|石
0.5	80	ω|苛|吉|錫|酒|藻|弗|没|酪|砒|硼|蔗|酢|´|灰|硝|-|ゾ|硫|六|窒|亜|ホ|錯|珪|黄|－|O|α|メ|芳|エ|ビ|息|ピ|，|燐|土|'|ェ|｛|チ|ボ|炭|素|｝|ニ|ジ|石|酸|鉛|ポ|リ|肪|網|雲|系|鉄|ル|塩|キ|白|ミ|ネ|オ|ノ|シ|ブ|ヒ|サ|ア|‘|'|γ|蟻|五|黒|族|陶|―

n/N；抽出された単語群中の化学物質名文字の頻度(n) /抽出された単語群の文字の頻度(N)

表 5　1-gram による化学物質名の抽出

n/N	抽出された単語群の数(R_n)	抽出された化学物質の数(c_n)	再現率 recall	適合率 precision	F 値
1	54	54	0.003	1.000	0.007
0.9	351	333	0.021	0.949	0.041
0.8	1575	1335	0.084	0.848	0.153
0.7	7106	4276	0.270	0.602	0.373
0.6	20643	10584	0.668	0.513	0.580
0.5	36782	13742	0.868	0.374	0.522

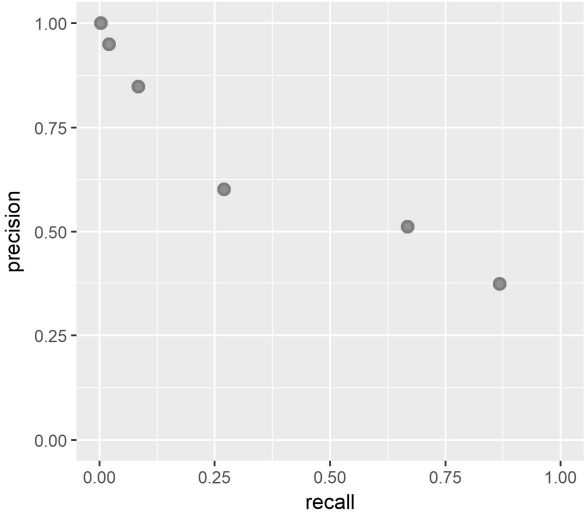

図 3　適合率– 再現率の関係

高い n/N の文字では適合率は高いことから一定の 1-gram の有効性が認められた．しかし，n/N を下げた文字を加えていくと，再現率は高められるが適合率は急激に下がった．この結果から 1-gram による方法では一部の化学物質名は選択できるが，すべての化学物質名を高い適合率で選択することは難しいことがわかった．

表 **6**には，各抽出条件において抽出された置換基名の数(g_n)とその他の数(o_n)を示し，置換基名とその他のそれぞれの再現率，適合率を示し，図 4に，化学物質名,置換基名とその他との適合率・再現率の関係をグラフで示した．置換基名は高い n/N では抽出されず再現率・適合率が低いため化学物質名の抽出に影響を与えないが，n/N が 0.6 以下と選出のしきい値を下げるとその他の語に比べて再現率が上昇し，影響が出ることがわかった．

表 6　置換基名及びその他の抽出状況

n/N	抽出された置換基名の数(g_n)	置換基名再現率	置換基名適合率	抽出されたその他の数(o_n)	その他再現率	その他適合率
1	0	0.000	0.000	0	0.000	0.000
0.9	0	0.000	0.000	18	0.000	0.051
0.8	39	0.013	0.025	201	0.001	0.128
0.7	282	0.094	0.040	2548	0.016	0.359
0.6	1171	0.392	0.057	8888	0.055	0.431
0.5	2357	0.788	0.064	20683	0.129	0.562

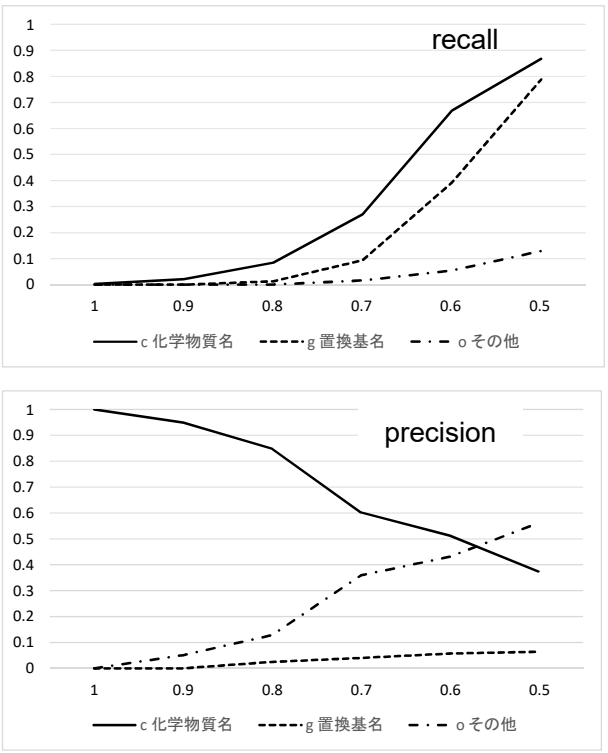

図 4　化合物名，置換基名，その他の各 n/N における適合率・再現率

5 まとめ

日本語の文章から化学物質名を自動抽出する方法を検討した．まず化学物質名の特徴を見出すために，特許公開公報に手動で化学物質名<chem>及び置換基名<group>のタグ付けを行った．日本語の文章から単語を認識す

るために形態素解析ツールを用いた．化学物質名は細かく認識されたため，品詞の種類や形態素解析のユーザー辞書を利用して化学物質名としてひとかたまりの単語となる認識方法を提案し，その妥当性を明らかにした．得られた単語が化学物質名か否かを選択する方法として1-gramの適用を検討し，一定の適合率での抽出が可能であるが，再現率を上げようとすると急激に適合率が低下するため，1-gram単独の方法では選出法として妥当ではないことを明らかにした．また1-gramでは，化学物質名の選出に置換基名が影響を与えることも合わせて示した．本研究では，化学物質名を正確に単語として認識することには成功したが，認識した単語群からの選択の方法は十分達成できていない．選択の方法には1-gramだけでなく2-gramや3-gram，共起語，構文情報や辞書との併用などによる自動抽出方法が考えられる．また，ビッグデータの充実から機械学習モデルを用いた認識・選択も考慮に入れている．この点については今後の検討課題としたい．

注記および参考文献

[1] 田中一成ほか: 「自然言語処理と Linked Data を用いた化学物質情報の可視化」, 言語処理学会 第 24 回年次大会 発表論文集, pp.1243-1246, 2018
[2] 本論文では文章から化学物質名を含む単語を切り出すことを認識，その中から化学物質名を取り出すことを選択といい，それらを合わせた操作を抽出という．
[3] 福田賢一郎ほか: 「医学生物学文献からの専門用語の抽出に向けて:タンパク質名の自動抽出」, 情報処理学会論文誌, Vol.39, No.8, pp.2421-24330, 1998
[4] Kemp, Nick.; Lynch, Michael.: "Extraction of Information from the Text of Chemical Patents. 1. Identification of Specific Chemical Names", J. Chem. Inf. Comput. Sci., Vol. 38, No. 4, pp.544-551, 1998
[5] Zhang, Yaoyun, et al.: "Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning", Database, pp.1-10, 2016
[6]Eltyeb, Safaa; Salim, Naomie: "Chemical named entities recognition: a review on approaches and applications", Journal of Cheminformatics, pp.6-17, 2014
[7] Krallinger, Martin et al., "Information Retrieval and Text Mining Technologies for Chemistry", Chem. Rev., Vol.117, No.12, pp.7673-7761, 2017
[8] Arighi, Cecilia N. et al.: "BioCreative-IV virtual issue", Database, https://doi.org/10.1093/database/bau039 Published: 22 May 2014
[9] Krallinger, Martin et al.: "Overview of the CHEMDNER patents task", Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, pp.633-75, 2015
[10] http://www.biocreative.org/tasks/biocreative-v/track-2-chemdner (2018 年 9 月 17 日参照)
[11] 田中一成; 池田紀子: 「オープンデータを用いた化学特許情報活用へのアプローチ」, Japio YEAR BOOK, pp.206-211, 2017
[12] 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, pp.280-287, 2015
[13] 池田紀子, 田中一成: 「特許文書から抽出した化学物質情報の知識化」, Japio YEAR BOOK, pp.204-208, 2016
[14] 石川大介ほか: 「特許文献における因果関係の抽出と統合」, 情報知識学会誌, Vol.14, No.4, pp.105-118, 2004
[15] https://www.publication.jpo.go.jp/index.action (2018 年 9 月 17 日参照)
[16] http://taku910.github.io/mecab/ (2018 年 9 月 17 日参照)
[17]https://dbarchive.biosciencedbc.jp/jp/nikkaji/download.html (2018 年 9 月 17 日参照)

（2018年12月16日　受付）
（2019年 8月31日　採択）