日本語文章からの化学物質名の抽出に関する研究：特許公開公報を対象にした検討

２０２２年３月

田中るみ子

日本語文章からの化学物質名の抽出に関する研究：特許公開公報を対象にした検討

筑波大学
図書館情報メディア研究科
２０２２年３月

田中るみ子

日本語文章からの化学物質名の抽出に関する研究：特許公開公報を対象にした検討

概要

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．

化学知識を有効活用するために，化学領域においてはこれまで，膨大に蓄積された論文（多くは非構造化データである文章）から人手によりファクトデータベースが整備されてきたが，多様な属性をもつ対象について大規模かつ網羅的なファクトデータベースを構築することは困難な作業であった．しかしながら現在は，多くの論文等が電子化されて提供されるようになったことから，膨大に存在するそれらの文章からファクトデータベースを自動構築することができれば，論文等の増加に対応が可能ではないかと考えられている．

化学領域のファクトデータベースの多くは化学物質名とその構造，機能，製造方法，化学反応，用途などの多様な属性から成り立つ．それらを文章から抽出するにはさまざまな側面からの検討が必要である．その中核となるのが化学物質名であることから，まず膨大な文章から化学物質名を表す語句を識別・同定することが必要である．

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，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．このように，化学物質名を識別し，同定

することは難しい問題となっている.

文章から特定の属性を持つ語を抽出する問題は, 自然言語処理の分野においては固有表現抽出（認識）(NER(named entity recognition)）として検討されている．もっとも基本的で正確な方法は，辞書を使った方法である．化学物質名の NER であれば，あらかじめ化学物質名のリストを辞書として用意しておき，文章中の語と辞書の文字列を確認し，完全一致もしくは近似一致されるものを固有表現として決定するという方法である．この方法は，辞書に含まれている化学物質名以外の語は抽出できないという欠点がある．そこで，辞書にはない化学物質名をも抽出できるよう, ルールを組み合わせる方法も考えられている．その一つはパターンベースのもので，化学物質名に特徴的に出現する文字や文字列を用いて抽出する方法であり, もう一つはコンテキストベースのもので, 文章上の配置を用いて化学物質名を抽出する方法である. これらにより，辞書ベースの方法に比べて多様な化学物質名への対応が可能になる.

辞書ベースの方法では，化学物質名フラグメントを備えた辞書を用いて，英文特許文書中の化学物質名を人手で作ったパターンに基づいて抽出する例がある．その結果，国際特許分類「C07D：複素環式化合物」を付与されている 70 の特許明細書において出現する化学物質名 14,855 のうち 97.4%を抽出することができたとしている. 辞書ベースの方法の抽出精度は使用する辞書の品質とマッチングアルゴリズムによって高い値を実現できる．ただ，新たな化学物質名に対応していくためには辞書を常にアップデートしていく必要があるため，辞書の品質の維持のためのコストが問題になる．またルールを組み合わせた方法についても, ルールに例外事象等が発生した場合はそれを修正する必要があり, 文章の増加に伴ってルールを維持していくためのコストが問題になる.

これに対して，辞書やルールを人手で作るのではなく，機械学習を用いて NER をブラックボックス化する方法が近年注目されている．BioCreative (Critical Assessment of Information Extraction in Biology) IV ワークショップ（2014）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）では特許文書から化学物質名を抽出する CHEMDNER（Chemical Compound and Drug Name Recognition）タスクが設けられた. BioCreative IV の CHEMDNER タスクは, 主要な化学分野から取り出された 10,000 の PubMed のタイトルと抄録について化学物質名を手作業でアノテー

ションしたコーパスを作成し, そのうち 3,000 のサブセットに対する化学物質名の抽出を競うものである．26 チームが参加し，多様な方法で検討が行われた結果，最高点は再現率 0.86，適合率 0.89，F 値 0.87 であった．これは人間のアノテーター間の合意比率が 91%であったことと比較すると大変有望な結果であると評価された. 一方 BioCreative V の CHEMDNER タスクは, 21,000 の医薬化学特許の発明の名称と要約に化学物質名を手作業でアノテーションしたコーパスを作成し, そのうち 7,000 のサブセットに対する化学物質名の抽出を競うものである．ここで用いられた国際特許分類は「A61P：生活必需品医学または獣医学；衛生学化合物または医薬製剤の特殊な治療活性」,「A61K 31/00：有機活性成分を含有する医薬品製剤」である．このタスクはノイズの多い特許文書から化学的およ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, 多様な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設定されたものである．こちらは 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.

英語文章を対象とした化学物質名の NER 研究に比べて，日本語文章を対象とした研究はそれほど多くない．その理由として，日本語文章の場合，単語が英語文章のようにスペースによって分けられておらずその切り出しが困難であること, アノテーションされた大規模なコーパスが無いことなどが考えられる．

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切り出す必要がある．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．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．また，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事例がある. 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．基本的に辞書を参照しながら形態素を切り出すため，辞書に登録されていない形態素は切り出すことはできない．さらに，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．著者は，形態素解析をベースにして，1）化学物質名を含む単語群の切り出し，2）得られた単語群からの化学物質名とそうでない単語の識別というプロセスを提案し，その詳細を検討し，そうすることにより全ての化学物質名が分断されずにひとかたまりになることを示した．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．本論文では，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．

Extraction of chemical substance names from Japanese texts: A case study of Japanese Published Unexamined Patent Applications

Abstract

Many chemically synthesized products are used in our daily lives, such as industrial products, medicines, and cosmetics, which are based on chemical knowledge. These products have various applications and considerably contribute to our quality of life. However, there are problems associated with the environmental impact and energy use of such products that need to be addressed; for solving these issues, chemical knowledge is indispensable.

Media, newspaper articles, and scientific and technical literature all contain a lot of chemical texts, which are based on chemical knowledge. The number of such texts and the chemical knowledge therein are both increasing daily. To solve problems in chemistry and for easier future reference, it is necessary to extract, organize, and accumulate chemical knowledge efficiently. However, this process requires a considerable amount of time and effort.

In the past, factual chemical databases were manually developed in the field of chemistry using numerous accumulated papers, mostly as unstructured data in the form of text. However, this made it difficult to construct a comprehensive database for a subject with so many attributes. In contrast, today, as papers are available in electronic forms, it is possible to keep up with the increasing number of papers by creating a factual database that can be automatically constructed using present technology.

Most factual databases in the field of chemistry comprise the names of chemical substances and attributes, such as structures, functions, manufacturing methods, chemical reactions, and applications. To extract this information from texts, it is necessary to consider few aspects. The

first step in extracting chemical information is identifying the name of the chemical substance, as it is the most important attribute, and the core of the vast database.

The names of chemical substances can be characterized by one or more of the following features: the creation of new names, changes in names according to the time of development (e.g., pharmaceuticals), variations in the notation due to the nomenclature, usage of arbitrary notations by authors. Additionally, there are often various notations, such as names, symbols, abbreviations, structural and chemical formulas, and bonding tables; thus, it is a difficult problem to recognize and identify chemical substance names.

The problem of extracting words with specific attributes from texts has been studied in the field of natural language processing, as named entity recognition (NER). The most basic yet accurate method to do so is to use a dictionary. For the NER of chemical substance names, a list of chemical substance names is first prepared as a dictionary. Thereafter, the words extracted from the text are checked against the strings in the dictionary, with the ones matching perfectly or approximately being determined as unique expressions. This method has the disadvantage that chemical substance names other than those contained in the dictionary cannot be extracted. Another commonly used method is to combine the rules to extract chemical substance names that are not in the dictionary. For example, a pattern-based method can extract chemical substance names using characters or strings that appear in the chemical substance names in a characteristic manner, whereas a context-based method can extract chemical substance names using their arrangement in a sentence. These methods allow the handling of a wider variety of chemical substance names compared to that possible with the dictionary-based method.

To elucidate the dictionary-based method, an example can be discussed, wherein the chemical substance names in English-language patent documents were extracted based on manually created pattern using a

dictionary with fragments of chemical substance names. As a result, of the 14,855 chemical substance names present in 70 patent specifications with the international patent classification "C07D: Heterocyclic Compounds," 97.4% were effectively extracted. Thus, the extraction accuracy of the dictionary-based method can be high, depending on the quality of the dictionary and the matching algorithm used. However, the cost of maintaining the quality of such a dictionary can be a problem, as it needs to be constantly updated to accommodate new chemical names. For the combining rules method, it is necessary to modify the rules when exceptions occur, and the cost of maintaining these rules can also become a problem as the number of sentences increases.

To address this issue, instead of manually creating dictionaries and rules, employing a method of black-boxing NER using machine learning has proven to be effective, and it has also been attracting attention in recent years. In the BioCreative (Critical Assessment of Information Extraction in Biology) IV workshop in 2014, a CHEMDNER (Chemical Compound and Drug Name Recognition) task was created to extract the chemical substance names from PubMed titles and abstracts. Furthermore, in the BioCreative V workshop (2015), a CHEMDNER task was created to extract the chemical substance names from patent documents. The CHEMDNER task in BioCreative IV involved creating a manually annotated corpus of chemical substance names for over 10,000 PubMed titles and abstracts from major chemical disciplines and hosting a competition to extract chemical substance names from a subset of 3,000 entries. Twenty-six teams participated, and various methods were examined. The highest scores were 0.86, 0.89, and 0.87 for recall, precision, and the F-value, respectively. These were very promising results, compared to the agreement ratio of 91% among human annotators. The CHEMDNER task in BioCreative V involved creating a corpus of 21,000 medicinal chemical patents, including the names of the chemical substances manually annotated on the titles and abstracts of inventions. Twenty-one teams

competed to extract chemical substance names from a subset of 7,000 entries. The international patent classifications used here were "A61P: Special therapeutic activity of essential medical or veterinary; hygienic compounds or pharmaceutical preparations" and "A61K 31/00: Pharmaceutical preparations containing organic active ingredients." This task was created to find a method for extracting the chemical and biological data from noisy patent documents. Such methods can ultimately be useful for a wide variety of document types. The highest scores achieved were 0.91, 0.87, and 0.89 for recall, precision, and the F-value, respectively.

Compared to the number of NER studies on the chemical substance names in English, there are significantly fewer studies on the chemical substance names in Japanese. One possible reason for this is that Japanese sentences are not divided into words using spaces (like English sentences), which makes their separation difficult. Additionally, the large number of character types makes it difficult to apply character-based extraction methods, and there exists no large-scale annotated corpus of chemical names.

For extracting chemical substance names from Japanese text, it is necessary to distinguish a word from a sentence before selecting a chemical name. In previous studies, we reported on extraction methods that focus on characters forming up chemical substance names and can recognize katakana, alphanumeric and kanji characters (such as "酸"), and a series of parentheses as candidates. Furthermore, these methods then use machine learning to select the appropriate substance name. Since these methods are mainly applied to the descriptions based on nomenclature, it is difficult to extract customary chemical names. In addition, morphological analysis has been conducted to discern chemical substance names from Japanese sentences. Morphological analysis has three functions: dividing sentences into the smallest units of meaning, called morphemes; assigning parts of speech to each morpheme; and restoring each morpheme to its original form. In other words, specific words are extracted by referring to a

dictionary, and words that are not registered in the dictionary cannot be extracted. However, a potential problem in general morphological analysis is that chemical substance names are often split in the middle and cannot be correctly recognized as chemical substance names. For example, "1－（6－ブロモ－ピリジン－3－イルメチル）－4－エチル－ピペラジン" and "酸化第一銅" are separated into smaller morphemes such as "1/－/（/6/－/ブロモ/－/ピリジン/－/3/－/イルメチル/）/－/4/－/エチル/－/ピペラジン" and "酸化/第/一/銅", so it is necessary to combine them into a single word as a chemical substance name. Based on morphological analysis, the author proposed a process of 1) cutting out words containing chemical substance names, and 2) discriminating between chemical substance names and non-chemical substance names from the obtained word groups. The details of this process were discussed, and it was shown that all chemical substance names can be formed as words by doing so. Once the morphemes are recognized as single words, we then have to consider the selection of chemical names. We can use various methods as aid, such as the notation of letters that form the chemical name, words that appear in and around the chemical name, words that modify the chemical name, and the meaning of the sentences containing the chemical name. Overall, the extraction of Japanese chemical substance names using morphological analysis consists of two stages: the recognition of words containing chemical substance names by separating and concatenating the morphemes from sentences, and the selection of the chemical substance names from the obtained groups of words. In this thesis, we report the results of our studies done for each of these stages.

目次

図目次

表目次

第 1 章　序論

1.1　導入

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，環境問題やエネルギー問題など人類が解決すべき課題の多くが化学知識が必要なものである．このように，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．

化学領域においてはこれまで，膨大に蓄積された論文（多くは非構造化データである文章）から人手によりファクトデータベースが整備されてきた[1]が，多様な属性をもつ対象について大規模かつ網羅的なファクトデータベースを構築することは困難な作業であった．しかしながら現在は，多くの論文等が電子化されて提供されるようになったことから，膨大に存在するそれらの文章からファクトデータベースを自動構築することができれば，論文等の増加に対応が可能ではないかと考えられている[2][3]．

化学領域のファクトデータベースは，構造化された化学知識であると考えられ，大規模なファクトデータベースが構築できれば，新規物質の開発，医薬品の探求，翻訳などさまざまな分野で応用が期待される．新規物質の開発には，物質の属性，化学反応における他の物質の相互作用，反応条件などのデータが欠かせない．過去の成功，失敗の事例とともに大規模なそれらのデータベースがあれば新規物質の開発に有効に活用できる．

近年の計算機能力の劇的な向上に伴い，事前に仮説を立てることをせずにデータを収集し，得られたデータを分析した上で研究を進めるデータ駆動型手法も現れ，大量のデータの処理，分析によっては，最初に予測した結果以外の思わぬ成果も持たらす．この新しい科学的手法の促進が社会貢献につながること

も期待される．このように，大量のデータを適切に蓄積し，社会に有用なコンテンツを見つけ出すことは今後ますます期待される. 大量のさまざまなデータにどう接し，データに潜む有用な知識発見を学ぶかは，社会を支える情報基盤技術として必要不可欠な技術となっている. そしてその基盤となるのがファクトデータベースであると考えられる.

わが国では，実用的なデータベース構築はほぼ海外に依存してきた．近年，ビッグデータの増加に伴い，AI 技術を活用するための自国のデータベースの確保が最重要と認識されるようになった. 日本語の論文や特許公報等を始め日本語のみで書かれた多くの化学知識を含む文章が多数存在する化学分野においても，既存データを利用して新たな知見・技術が求められるため，自国のデータベースの充実が望まれる．このような点から，経済産業省では，化学産業界からの要請を受けて，学術論文，特許，社内文書などから材料開発に必要となるデータを自動抽出するプラットフォーム構築を支援するプロジェクトを 2019 年度からスタートさせている[4].

化学領域で作られているファクトデータベースの多くは化学物質名とその構造，機能，製造方法，化学反応，用途などの多様な属性から成り立つ．それらを文章から抽出するにはさまざまな側面からの検討が必要である[5]．その中核となるのが化学物質名であることから, まず膨大な文章から化学物質名を表す語句を識別・同定することが必要である.

化学物質名には, 名称, 構造式, 結合表, 化学式, 記号などの表現法がある．さらに名称には体系名，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 [6]がある．最新の IUPAC 指針は 2013 年 12 月に発行された．慣用名は，物質の出所や特性などを表すラテン語や学名からつけられ，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.

化学物質名称の多様な表記について, 国内外の大規模化学物質データベースを用いて例証する．表 1 は日本化学物質辞書 Web（日化辞 Web） [7]と PubChem [8]における「酢酸エチル（ethyl acetate）」の名称である．「酢酸エ

チル」に対して日化辞で 17 通り，PubChem で 34 通りの名称がつけられている．

表 1　「酢酸エチル」の名称

日化辞	PubChem
エタン酸エチル	Acetate d'ethyle
エチル=アセタート	Acetato de etilo
酢酸エチル	Acetic acid ethyl ester
Acetic ether	Acetic acid, ethyl ester
Ethyl acetate	Acetic ether
Vinegar naphtha	Acetidin
Acetic acid ethyl	Acetoxyethane
アセチックエーテル	Aethylacetat
ビネガーナフタ	AI3-00404
アセチジン	Caswell No. 429
RCRA waste number U-112	CCRIS 6036
Acetidin	EC 205-500-4
エチルアセテート	EINECS 205-500-4
Ethyl=acetate	EPA Pesticide Chemical Code
Acetic acid ethyl ester	044003
Ethanoic acid ethyl	Essigester
酢エチ	Ethyl acetate
（17 通り）	Ethyl acetate (natural)
	Ethyl acetic ester
	Ethyl ester
	Ethyl ethanoate
	Ethylacetaat
	Ethylacetate
	Ethyle (acetate d')
	Ethylester kyseliny octove
	Etile (acetato di)
	FEMA No. 2414
	HSDB 83
	NSC 70930
	Octan etylu
	RCRA waste no. U112
	RCRA waste number U112
	UN1173
	UNII-76845O8NMZ
	Vinegar naphtha
	(34 通り)

科学技術文書における化学物質名の記載はIUPACに準拠することが推奨されているものの，どの名称を使うかは書き手に委ねられており，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

このように化学物質の名称には，多様な表記がある，慣用名や商品名など特定の規則性がない表記がある，書き手により任意の表記を用いる場合があるなどの特性があるため，それを識別し，同定することは難しい問題となっている．

本論文では，ファクトデータベースの自動構築に向けて日本語文章から化学物質名を抽出する方法を検討することを目的とする．次節では，関連した内外の先行研究を探る．

1.2　先行研究

1.2.1　NER の方法

文章から特定の属性を持つ語を抽出する問題は，自然言語処理の分野においては固有表現抽出（認識）（NER(named entity recognition)）として検討されている．

NER とは，テキストから人名，地名，日付などの固有表現を抽出するタスクで，「Named Entity」という用語は，「the sixth Message Understanding Conference (MUC-6)」で定義された[9]．固有表現の種類については，アメリカ合衆国の Defense Advanced Research Projects Agency（DARPA）が組織した評価型プロジェクトである MUC では，「組織名（ORGANIZATION），人名（PERSON），地名（LOCATION），日付表現（DATE），時間表現（TIME），

金額表現（MONEY），　割合表現（PERCENT)」の 7 種類と規定されている．固有表現の種類は分野により異なり，分野ごとの専門知識が必要なため，コーパスの作成に時間や労力を要する．

NER システムの種類を図 1 に示す．

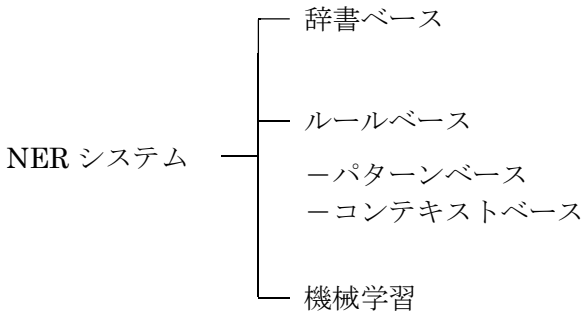

図 1　NER システムの種類

辞書ベースのシステムは，辞書内の用語のリストを使用して，テキスト内のエンティティの出現を識別する．システムは，テキストから選択された単語または単語群が辞書の用語と一致するかどうかを同定する．このアルゴリズムは，次の 2 つのタイプに分けることができる．

1. 完全一致：テキストに対して指定された用語リストから一致した語を同定する．

2. 部分一致：同定時，一部の文字の挿入，削除，または置換を可能にする．

部分一致はあいまい一致を実行し，ほとんどの NER アプローチで使用されている[10]．

ルールベースのシステム[11]では，通常，次の 2 つのルールが使用される．

1.パターンベースのルール：特徴的に出現する文字や文字列を用いて抽出する．

2.コンテキストベースのルール：文章上の配置を用いて抽出する．

コンテキストベースのルールの例としては，「人の肩書きの後に固有名詞が続く場合，その固有名詞は人の名前である」がある[12]．

機械学習（ML）アプローチ[13][14][15]に基づく NER システムは，アノテーションされた文書に依存する観測データの特徴ベースの表現を利用して，特

定のエンティティ名を認識するための統計モデルを使用する． ML ベースのシステムを開発するには，次の 2 つの基本ステップが必要である．

1. トレーニング：機械学習モデルは，アノテーションされた文書の中に存在するアノテーションを手がかりに正解と不正解を見分けるようにトレーニングする必要がある．

2. アノテーション：文書中のデータに対して関連する情報（メタデータ）を注釈として付与すること．

ML アルゴリズムは，求められる結果に基づいて分類される． NER で使用される一般的な ML アルゴリズムには教師あり学習，教師なし学習，半教師あり学習，ハイブリッド学習などがある．

教師あり学習アルゴリズムは, 学習インスタンスに正解データをラベル付けすることで学習し，学習過程をフィードバックする（教師あり学習）．たとえば，通常は教師あり学習によって解決される分類問題では，コンピュータは作成されたタグ付きデータを学習し，それに応じて出力を生成する[16]．

教師なし学習アルゴリズムは, 方法を説明せずにコンピュータに何かをする方法を教えることを目的としており，学習時にラベルがわからないため，難しい．したがって，教師なし学習におけるプログラムの目的は，データから特徴を見つけ出すことである．クラスタリングは，教師なし学習の一例で，学習の中から類似性を見つけることを目的としている． [17][18]．

半教師ありアルゴリズムは, ラベル付きデータとラベルなしデータの両方を使用する．これらのタイプのシステムには，学習プロセスを開始するために手動で定義された信頼できる小さな種のセット，すなわち，わずかな正解データのセットが含まれる．たとえば，「病名」を抽出するシステムには，関連する例として少数の病名が用意されている. これらの例文を含む文章をシステムを使って検索し，例文に共通する文脈上の手がかりを特定することを目指す．そして，似たような文脈の中で現れた他の例を再び検索する．この学習プロセスは，新たに発見された例に対して継続的に再利用され，新たな関連する文脈を発見する．このようにして，このプロセスを繰り返すことで，多数の病名を認識することができる[19]．

1.2.2　化学物質名の NER 研究

化学物質名に関わるコンピュータ処理については，1960 年代に Eugene Garfield が先駆的な研究を行っており，体系名をアルゴリズム的に分子式と線表記に変換する方法を開発した [20]．Zamora らは，一般的な化学物質，化学式, 化学用語辞書を含む自然言語処理技術や化学用語の形態素識別を行うことによって, テキストから反応情報を自動的に抽出することを試みた[21] [22]. Hodge らは, テキストフィールド内の化学物質名を認識して CAS 登録番号を割り当てる技術について論じた[23]．また，Blower と Ledwith は，辞書検索技術と単語の形態論を利用して, 有機化学雑誌の実験欄に記載されている物質を識別し，化学反応に関する反応物，生成物，反応条件，収率などの具体的な項目をルールベースのアプローチで抽出した[24].

Kemp と Lynch 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, 文書中から物質名を自動識別する手法を提案した. 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する研究を行なった．抽出リストより KLIC Index を作成し，この KLIC Index と substrings，stopwords，stopstrings をもとに，タグをつける語の選択を行った．タグ付けは SGML（Standard Generalized Markup Language）で行い，化学物質名には<chem.>タグを付与した．その結果，国際特許分類「C07D：複素環式化合物」を付与されている 70 の特許明細書において出現する化学物質名 14,855 のうち 97.4%を抽出することができたとしている[25].

Klinger らは, 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）に基づいた機械学習アプローチを用いてテキストをトークン化し，ラベル付けを行うモデルを作成した. 化学名すべてを対象とする代わりに IUPAC らしい用語に認識を制限すること（IUPAC 名のフラグメントおよび部分を認識する）がパフォーマンスを増加させることを示した. 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[26].

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[27]などがある．

Hettne ら[28]および Rebholz-Schuhmannetal[29]は，文字列一致法を使って薬物名と分子名を抽出した．一般に，辞書ベースの方法は，使用される辞書およびマッチングアルゴリズムの品質が高ければ，高精度を提供するが，テキストのスペルミスの場合の再現率は低くなる．この方法は，辞書ベースのシステムで古い辞書が変更された場合，辞書の保守にコストと時間がかかる．

辞書にはない化学物質名をも抽出できるよう，ルールを組み合わせる方法も考えられている．これらにより，辞書ベースの方法に比べて多様な化学物質名への対応が可能になる．

Narayanaswamy らは，化学名（例：インドメタシン，N-メチルホルムアミド）や化学名の一部（例：メチル，メタ）など，さまざまなカテゴリを対象としたルールベースの化学に関連する用語を引き出すためのタガーを考案した．このタガーは，個々の単語に対して，化学的な中核用語（化学 c ターム）と化学的な機能用語（化学 f ターム）に分類することを基本としている．化学コア用語の認識には，形態素，大文字，数字，特殊記号などの表面的な特徴のルール，化学語源の形や接辞の検出ルール，IUPAC の化学物質の命名規則に基づくルールなどを用いた [30]．ルールベース NER は，必要なリソース（専門家が作成したルールのセットなど）が利用できる場合には良い性能を発揮するが，システムの移植性に欠ける．ルールに例外事象等が発生した場合はそれを修正する必要があり，文章の増加に伴ってルールを維持していくためのコストが問題になる．

辞書は，通常，データベースやシソーラスなどの公開ソースから手動または自動で作成できる．化学および生物医学分野の辞書とコーパスの例をそれぞれ表 2，表 3 に示す．

表 2　化学および生物医学分野の辞書の例

データベース名	アドレス
ChEBI	http://www.ebi.ac.uk/chebi/
ChEMBL	https://www.ebi.ac.uk/chembl/
ChemIDplus	https://chem.nlm.nih.gov/chemidplus/
ChemSpider	http://www.chemspider.com/
DrugBank	http://www.drugbank.ca/
Jochem	http://www.biosemantics.org/index.php?page=jochem
KEGG COMPOUND	http://www.genome.jp/kegg/compound/
MedlinePlus	https://medlineplus.gov/druginformation.html
MeSH	https://www.nlm.nih.gov/mesh/
NCI Drug Dictionary	https://www.cancer.gov/publications/dictionaries/cancer-drug
NIAID ChemDB	https://chemdb.niaid.nih.gov/
PubChem	https://pubchem.ncbi.nlm.nih.gov/
日本化学物質辞書 Web	https://jglobal.jst.go.jp/info/nikkaji

表 3　化学および生物医学分野のコーパスの例

コーパス名	アドレス
IUPAC training corpus	http://www.scai.fraunhofer.de/chem-corpora.html
SCAI	http://www.scai.fraunhofer.de/chem-corpora.html
NLM-Chem corpus	https://www.ncbi.nlm.nih.gov/research/bionlp/Data/
GENIA corpus	http://www.geniaproject.org/genia-corpus
European Patent Office and the ChEB	http://chebi.cvs.sourceforge.net/viewvc/chebi/chapati/patentsGoldStandard
CHEMDNER Corpus	http://www.biocreative.org/tasks/biocreative-iv/chemdner/

辞書ベースの方法の抽出精度は使用する辞書の品質とマッチングアルゴリズムによって高い値を実現できる．ただ，新たな化学物質名に対応していくためには辞書を常にアップデートしていく必要があるため，辞書の品質の維持のためのコストが問題になる．またルールを組み合わせた方法についても，ルールに例外事象等が発生した場合はそれを修正する必要があり，文章の増加に伴ってルールを維持していくためのコストが問題になる．これに対して，辞書や

ルールを人手で作るのではなく，機械学習を用いて NER をブラックボックス化する方法が近年注目されている．

ほとんどの NER システムは，形態素，言語，文脈，語彙などの領域に依存しない特徴量を使用しているが，これらの特徴量が性能の有効性に与える影響を検証した研究は少ない．ニュース領域の研究[31]やバイオメディカル領域の研究[32]などのいくつかの研究では，NER システムにおける異なる特徴量の使用やその組み合わせの有効性が調査されている．しかし，化学物質のエンティティは，特に形状の特徴の点で，ニュースのエンティティとは異なる．化学名には記号，ローマ数字，ダッシュ，大文字，小文字が含まれているため，大文字，記号，単語の形状パターンなどの正字・形態素の特徴は，パターンベースのルールや教師あり学習の NER アプローチにおいて非常に重要である．

最近の多くのシステムでは，領域固有のリソース（例えば，医薬品 FDA や ATC の命名リスト）や他の NER システムの出力を使用するなど，領域固有の特徴量を使用している．領域固有の特徴量の使用が全体のパフォーマンスに貢献していることが示されている報告がある[33]．これらの特徴量の使用に加えて，空白と句読点などの記号で分割するトークン化は NER システムの重要な問題である．化学の NER システムでは，化学的実体の形状を考慮した特別なタイプのトークン化装置が必要となる．例えば，「(R)-acetoin」という単語から括弧が削除されることはない．しかし，一般的なトークン化装置は，括弧があるところはどこでもトークン化する[34]．ほとんどの研究では，辞書や辞典，トークンの接頭辞（“di”, “tri”, “tetra”など）や接尾辞（“-yl”, “-oyl”, “-one”, “-ate”など）の情報を使用することで，すべてのタイプの NER システムのパフォーマンスが向上することが確認されている．

ハイブリッド NER システムは，各アプローチの優れた特性を活用するために，複数の NER アプローチを実装している．化学物質名の NER では，辞書アプローチは通常，パフォーマンスを向上させるためにルールベースまたは機械学習アプローチと組み合わされる．たとえば，ChemSpot [35]は，テキスト内の化学物質（慣用名，薬，略語，分子式，および IUPAC）を識別するための NER ツールである．これは，CRF モデルと辞書を組み合わせたハイブリッドアプローチを実装する．組み合わせの主な目的は，これらのクラスのさまざまな命名特性をカバーすることであると述べている．IUPAC エンティティ

は, 他のエンティティよりも形態学的に複雑である. IUPAC エンティティは, ルールベースに従うのが難しく, 辞書で見つけるのが最適である. ChemSpot は, CRF モデルと辞書を個別に使用して, テキストにアノテーションを付ける. 最後に, 両方のアプローチのアノテーションがマージされる. 辞書または CRF によって抽出されたエンティティは重複する可能性があるが, ChemSpot は抽出されたすべてのエンティティの和集合を維持し, CRF モデルから一致したものを選択することによってこの重複を解決する. ChemSpot が用意した辞書コンポーネントは, 抽出されたエンティティを CAS レジストリ ID に紐づけるためにも使用される.

化学物質の NER アプリケーションでは, 条件付確率場（CRF）[36][37]や隠れマルコフモデル（HMM), 最大エントロピーマルコフモデル（MEMM) [38]などの教師あり学習モデルが, 広く研究されている.

1.2.3 化学物質名の NER のワークショップ

化学物質名の NER 研究を進展させるため, 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 systems in Biology) IV ワークショップ（2014）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)では特許文書の発明の名称と要約から, 化学物質名, 遺伝子, タンパク質などを抽出するタスクが設けられた. そこでは, 専門家チームによって手作業でアノテーションおよび分類された CHEMDNER （Chemical Compound and Drug Name Recognition）コーパスが用意された [39][40]. BioCreative IV の CHEMDNER タスクは, 10,000 のアノテーションしたコーパスのうち 3,000 のサブセットに対する化学物質名の抽出を競うものである. 26 チームが参加し, 多様な方法で検討が行われた結果, 最高点は再現率 0.86, 適合率 0.89, F 値 0.87 であった. 一方 BioCreative V の CHEMDNER タスクは, 21,000 の医薬化学特許の発明の名称と要約に化学物質名を手作業でアノテーションしたコーパスを作成し, そのうち 7,000 のサブセットに対する化学物質名の抽出を競うものである. ここで用いられた国際特許分類は「A61P：生活必需品医学または獣医学；衛生学化合物または医薬製剤の特殊な治療活性」, 「A61K 31/00：有機活性成分を含有する医薬品製剤」である. このタスクは

ノイズの多い特許文書から化学的およ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, 多様な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設定されたものである．こちらは 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.

英文からの化学物質名抽出については，BioCreative V までの研究のレビューが紹介されている[41].

CHEMDNER タスクのコーパスを用いた研究はその後も継続して多くの研究者によって行われている．ディープニューラルネットワーク BiLSTM-CRF モデルなどの新しい手法を適用するなどしてそれぞれの指標で 0.01-2 程の値の向上が認められている. ディープニューラルは過去の手法を組み込むことで改善が示された[42]. Awan らは BiLSTM-CRF に 4 つのコーパスを使用してネットワーク埋め込み言語モデル（ELMo）を組み込むと F1 値が有意に改善されることを示した [43]. Saad はラベルのない特許コーパスから自動的に生成された機能に基づいて，BiLSTM モデルを改良した[44].

1.2.4　日本語文章に対する化学物質名の NER 研究

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. 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．ただ，化学物質名はスペースや記号を含むため，化学物質名と認識するには，分断されたものをさらに連結する必要がある．さらに化学物質名の特徴として，構造の一部を表す置換基名が使われていることで，置換基名を単位に考えると，ある一定の文字種から成り立っていると言える．日本語は，明確な区切りは句読点のようなものしかない．英文はスペースで区切られた部分は一種類の品詞であるが，日本語の場合は句読点で句切った場合，「10（名詞）m（名詞）M（記号）の（助詞）安息香酸（名詞）が（助詞）遊離（名詞）する（動詞）と（助詞)、(記号)」のように多種類の品詞で構成される．そのため，化学物質名を選択する前に文章から品詞ごとの形態素を認識する形態素解析が必要である. また, 日本語のコーパスが海外に比べて充実していないことも日本語文章の研究が少ない要因である.

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[45]．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．

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[46]．

一ノ瀬らは，生化学関連の特許文書から化学物質名を抽出する手法を提案した．物質名の表記の特徴をとらえ，フラグメント辞書（フェニル，メチル），特性基接頭語，接尾語辞書（ニトロ，イミド），接頭語，接尾語辞書（ジ，トリ），記号辞書（－，’），付加語辞書（誘導体，組成物），単位辞書（重量%，モル），の 6 種類の辞書を作成し，辞書に登録される用語の組み合わせにより化学物質名を抽出した．また化学文献タイトルに現れる体系的化合物名にはハイフン，カンマ，ピリオド，数字などが連続していて，機械翻訳の構文解析を妨げるため，化合物名をひとつの単位として認識する方法，母体化合物と基（グループ）名を登録し，フラグメントに分けて翻訳する方法を開発した．ここでフラグメントとは，化学物質名の構成要素で化学的に意味があるメチル基，エチル基などの置換基，置換基の位置を示す位置番号などのことである[47]．

池田らは，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を報告している[48]

[49][50]．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．

邊土名らは Wikipedia の化合物記事から原材料と製造方法を抽出するために，外部知識源から収集・構築した化合物名辞書を用いて化合物名を置換する手法を提案した．化合物属性の構造化情報は，薬剤や材料の開発，化合物特許の解析において重要であるとして，「関根の拡張固有表現階層[51]」の定義による化合物の属性である種類・原材料・製造方法・別名・用途・特性の中でもパターンベースで抽出しにくい原材料と製造方法の構造化データを作成することを目的とした．しかし，化合物名の抽象化のために 1 種類の文字列で置換してしまうと，記事タイトルとなっている化合物とその他化合物それぞれの生成に関する文が混在している場合には区別できなくなるという問題が生じてしまうため，化合物名を「タイトル化合物」か「その他化合物」に置換し抽象化することで，抽出精度が向上することを確認した[52]．

先行研究では，英文の文献からの化合物名抽出は，英文が空白や記号で分かれており，大規模なコーパスが複数あり，抽出の手法では機械学習が最もよく利用されている．日本語の文章は形態素解析が必要で，コーパスが少なく，先行研究の数は少なかったが，最近議論が交わされるようになった．

1.3　研究目的の設定

本研究では，化学物質名，属性，構造，機能，化学反応，用途などを統合したファクトデータベースの自動構築に活用できるように，日本語文章から化学物質名を認識する方法を見出すことを目的としている．

日本語の化学物質名を抽出するには，文章からの形態素解析による形態素の切り出し，形態素の連結により化学物質名を含む単語を形成する段階と，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．

具体的には，文章から単語を切り出しただけでは，細かい形態素に分かれてしまうため，連結により意味のある単語にまとめなければならない．一方，そのような連結により，余計な部分が化学物質名についてしまう場合もあるため，化学物質名ではない部分を取り除く必要がある．得られた単語群から化学物質名を選択する手掛かりとして，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．

本論文では，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．

第 2 章　研究方法

日本語文章から化学物質名を認識するためには, 日本語文章から化学物質名を含む単語を切り出すことと, 切り出した単語を化学物質名かそうでないかを識別するという 2 つの段階がある．本章では単語の切り出しと識別についての方法を提案し，以降でその方法を行なった結果をまとめる．

2.1　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評価を行うためには, 化学物質名をアノテーションしたコーパスが必要である．日本語のコーパスがなかったため，その作成にあたっては，CHEMDNER タスクにならって特許公開公報の化学分野の電子データを利用する．

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, 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[53].

2.2　化学物質名を含む単語の切り出し方法の概要

日本語の文章から化学物質名を抽出するためには, まず文章を化学物質名を含む単語単位に切り出す必要がある. 日本語の文章から単語を切り出す方法としては, 形態素解析が一般に用いられている．一般的な形態素解析ツールは文を形態素という意味の最小単位へ分割し, 各形態素に品詞を付与し, さらに各形態素を原型へ復元するという 3 つの機能を持つ．基本的に辞書を参照しながら形態素を切り出すため, 辞書に登録されていないものは切り出すことができない. そのため, 化学物質名のように複合的な単語の場合は細かく分断され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」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」のように細かく切り出されてしまう．

そこで，形態素解析を行った後，得られた形態素の品詞情報をもとに，品詞情報が妥当であるかどうかをチェックし，辞書を修正するとともに，化学物質名の品詞情報を利用して．一定の品詞が連続して出現する場合は連結してひとかたまりの単語にする．

上述のような化学物質名をひとかたまりの単語にする方法だと，化学物質名には化学物質名以外の語句も連結されるため，それらの語句がどのように連結されたかを分析し，置換処理によって化学物質名と切り離す．

2.3　化学物質名とそれ以外の単語の識別方法の概要

化学物質名を含む単語群が切り出されたなら，その中から化学物質名を識別・同定する方法を確立する必要がある．辞書ベース，ルールベース，機械学習の方法の中から，本研究では，ルールベースのパターンベースと機械学習を用いる．

第 3 章　化学物質名認識までの方法の検討結果

3.1　作成したコーパス

Kemp らは 70 件の特許明細書を対象にしていることを参考に，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 507 公報を公報番号の順で 10 公報おきに取り出した．最初の公報 50 件（総文字数：1120210）については，発明の名称，要約，明細書に記載されている化学物質名に対して前後を<chem>と</chem>で囲むことでタグ付けを行った．タグ付けの妥当性を高めるため，一部のタグ付け結果を化学の専門家がチェックし，作業にフィードバックした．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，および部分的に構造が明確な物質群であり，さらに商品名も含めた．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，パターンベースのルールで化学物質名の識別を行う際には，文字列のマッチング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 2 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 2　タグ付けの例

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表 4 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.

後述するような形態素を連結するという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」に連結処理後，すべて含まれていることを確認した．このように,タグ付けしたものを切り出した後，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 4 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 4　公報毎の単語数およびタグの付与頻度

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

3.2 化学物質名の切り出し方法の検討

3.2.1 形態素解析による方法

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切り出す必要がある. この単語単位の切り出し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[54].

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 5 のように形態素に切り出され, 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 5 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 6 に示す.

表 6　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	－ß 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る．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）[55]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[56]．この方法により切り出された単語群の数が，表 4 に示した「Z」である．

なお，表記が異なると別の単語として扱われるため，文字はすべて全角に統一した．

3.2.2 化学物質名の分離処理

前項の方法に従って，コーパスを作成した公報 50 件のもとのテキストから，発明の名称，要約，明細書の部分を取り出し，それを形態素解析ツール MeCab を使って単語単位に分割し，助動詞，動詞を原形にした後，記号，接頭詞，名詞が連続していた場合に連結して一つの単語とすると図 3 のようになる．

（形態素解析前テキスト）

液体アンモニア（５０．０ｋｇ）を、Ｔ：４０℃にて１－（６－ブロモ－ピリジン－３－イルメチル）－４－エチル－ピペラジン（１４．２ｋｇ）、酸化第一銅（２００ｇ）、およびＭｅＯＨ（５７ｋｇ）の脱気した混合物に加える。

（形態素解析後テキスト）

液体 アンモニア（ ５０ ． ０ ｋｇ ） を 、 Ｔ ： ４０ ℃ にて １ － （ ６ － ブロモ － ピリジン － ３ － イルメチル ） － ４ － エチル － ピペラジン（ １４ ． ２ ｋｇ ） 、 酸化 第一 銅（ ２００ ｇ ） 、 および ＭｅＯＨ（ ５７ ｋｇ ） の 脱気 する た 混合 物 に 加える 。

（形態素解析後連結テキスト）

液体アンモニア（５０．０ｋｇ） を 、 Ｔ：４０℃ にて １－（６－ブロモ－ピリジン－３－イルメチル）－４－エチル－ピペラジン（１４．２ｋｇ） 、 酸化第一銅（２００ｇ） 、 および ＭｅＯＨ（５７ｋｇ） の 脱気 する た 混合物 に 加える 。

図 3 テキストの連結

このように化学物質名は分断されずひとかたまりになったが，文章からの場合化学物質名以外の記号や文字列が前後に付加されるものがあった．

付加された文字列の内容，場所等を分析し，化学物質名に付着する余計な文字列の分離のために対応を検討した．例えば，記号については「－」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．

文字列については，例えば，『Ｎ－メチル－２－ピロリドン等』，『０．９質量％塩化ナトリウム水溶液１００ｇ』，『１，４－ジオキサン濃度』，など，化学物質名を含むものの『等』や『０．９質量％』などの文字列が付加された．文字列には『等』や『以外』などの例示や除外を示す文字，『５０％』や『４℃』などの数値や単位，『水溶液』や『融点』などの状態を示すもの，『商品名』や『略称』などの別名を示すものなどがあった．(　)書きで，複数の化学物質名と属性が羅列されているものなど，物質名個々として取り出すのが難しい文字列もあった．これらの文字列は形態素解析後の品詞の種類による処理だけでは，化学物質名と分離できないため，文字列に応じた方法で切り離すことにした．

このような検討を行い，最終的に付加された文字列の内容，場所を分析し，化学物質名に付着する余計な文字列の分離のために，以下のような正規表現による置換を行った．

1. 商品名や「；」で併記された表記を分離する
   ((「(通称|商品名)：?)|；)
2. 物質名のうしろに現れる文字に着目して区切る
   (以外|等|中|水溶液|溶液|濃度)
3. かっこに続く属性の説明の前で区切る
   (（(融点|以下|登録商標))
4. 物質名の最後の文字としては不適切な開き括弧などの記号の前で区切る
   ([％（,．／：[〔＋]　)
5. 物質名の最初の文字としては不適切な閉じ括弧などの記号の後で区切る
   (　[・％),．／：]〕＋])
6. 物質名の前に付加される箇条書き記号の後で区切る
   化学式に頻出のＣＨＯＮＳを除いて(　(?[０-９Ａ-ＢＤ-ＧＩ-ＭＰ-ＲＴ-Ｚａ-ｚ]*))
7. 物質名の直前に付記される特定の文字で区切る
   (前記|名称|各|上記)
8. かっこのあとに単位や数値があった場合にかっこの前で区切る
   (（([０-９.,　当量ミリモル℃ｇｋＫｍＭｌＬμ×]*))
9. 特有の語句の前で区切る
   (含有量|含有ガス)
10. 特有の語句の後ろで区切る

(重合体|樹脂|炭素繊維|架橋剤)

11. 物質名の後に付加される箇条書き記号の前で区切る

(([０-９A-Ｚａ-ｚ]*)　)

12. 数字と特有の語句の組み合わせの前で区切る

([０-９～．]*質量部)

これらの置換により，化学物質名と同定できた同定率の変化を表 7 に示す．

表 7　物質名同定率の変化

置換処理	例	新たに同定できた数
物質名の後ろの出現文字列	(以外|等|中|水溶液|溶液|濃度)	681
併記された表記を分離	(([([「](通称|商品名)：?)| ;)	348
高分子物質名に併記	(重合体|樹脂|炭素繊維|架橋剤)	296
水素，酸素などガスに併記	(含有量|含有ガス)	237
単位，商標などの補足説明	(((融点|以下|登録商標)) ([%（,．／：[〔＋]　) ([０-９～．]*質量部)	223
量，温度などの説明	(([０-９.、　当量ミリモル℃ｇｋＫｍＭｌＬμ×]*))	195
文の箇条書き	(　[・%),．／：]〕＋]) (　(?[０-９A-ＢD-ＧＩ-ＭP-ＲＴ-Ｚａ-ｚ]*)) (前記|名称|各|上記)	148
物質名に着く順序の記号	(([０-９A-Ｚａ-ｚ]*)　)	129

形態素解析と記号，接頭詞，名詞の連結により同定できた物質名は 9391 個であり，タグ付けした化学物質名の総出現数 14486 個に占める割合は 64.8%であった．なお，表 4 には化学物質名の総出現数は 15834 個となっているが，その後以下の処理を検討する過程で見直しを行いタグ付けを変更したところがあるため，14486 個となった．

以上の整形により，最終的に総出現数 14486 個のうち 11648 個，80.4%が同定できた．同定できなかった化学物質名は０．３Ｍシュウ酸浴のシュウ酸，レピドクロサイト型チタン酸塩のチタン酸塩などである．

化学物質名の切り出し処理ではタグ付けした化学物質名のうち約 80%の化学物質名が同定できたが残りの 20%は，学習データでは化学物質名としてラ

ベル化されていない.

なお，化学物質名の前後に現れる文字列と物質名を切り離すにあたり，化学物質名を構成する文字種について，日本化学物質辞書 Web で出現する化学物質名の文字種を調べ，例えば，『等』が物質名を構成する文字種ではないことを確認した.（表 8）

表 8　化学物質名文字の出現頻度

文字	頻度	文字	頻度	文字	頻度	文字	頻度	文字	頻度
ー	1519117	b	128500	N	43627	ビ	23619	ム	7074
e	553936	5	128461	R	41006	テ	23226	I	6561
y	426407	s	125900	ピ	40476	f	22868	コ	6258
l	424098	ロ	125633	8	40326	ソ	21438	Z	6126
,	421854	x	114053	ラ	39818	デ	20266	ゼ	5991
o	421059	シ	110733	0	39506	ヘ	18468	:	5367
a	373315	キ	107313	ミ	38974	B	17966	パ	5118
i	357694	ジ	105424	ヒ	38718	サ	17797	グ	5101
t	342894	u	98077	エ	38197	ボ	17473	ダ	4827
n	332792	メ	90068	9	38014	／	16856	q	4215
2	330519	6	89405	ク	37678	M	16233	G	3721
1	316866	オ	84827	ス	36865	E	15944	ザ	3712
ル	308365	(空白)	82535	O	35102	セ	15881	ィ	3682
h	303699	ニ	80652	D	32761	ペ	15802	バ	3582
（	267164	ト	78926	＜	32756	L	15682	F	2898
）	265083	フ	73128	＞	32756	レ	15364	ポ	2699
r	255309	リ	70603	ベ	30667	ホ	13592	マ	2677
3	233515	ア	65279	α	29931	P	13333	酢	2617
p	214882	H	61432	タ	28895	T	13220	v	2569
4	211798	z	60122	一	28560	’	12880	V	2412
d	192030	ド	59919	β	27011	A	12630	k	1917
ン	179151	7	56426	カ	26884	ナ	11269	安	1871
c	169893	ノ	50782	プ	26318	g	9390	息	1869
［	147686	イ	49514	酸	25904	C	9328	香	1864
］	140412	ェ	47968	ゾ	25579	＊	8610	・	
m	135877	′	46734	.	24106	ウ	8237	・	
チ	129297	S	43720	ブ	23823	モ	7276	・	

3.3　化学物質名とそれ以外の単語の識別方法の検討

前章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，ルールベースとして化学物質名は構成する文字列に特徴があり，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，限られた漢字から成り立つことに着目した．ルールベースとして文字列に着目したのは，主に化学物質名の特徴を文字列が最も反映しているもので，視覚的に識別結果がわかりやすく，一文字，文字列の組み合わせ，を検討できると考えた．この方法だけでの識別の可能性は低いと考えられるが，どのくらい識別できるかを調べるために実験を行った．次に，先行技術文献調査からも機械学習が主流となっている，例えば，BioCreative IV では参加チーム 26 チーム中 20 チームが機械学習の手法を用いている[39]ことより，機械学習の二値分類の方法を検討した．

3.3.1　ルールベース

3.3.1.1　1-gram の方法

化学物質の名称は IUPAC に準拠することが推奨されている．IUPAC に従って命名した場合，名称は体系名，慣用名に大別される．体系名は化合物の既知構造に対して，一意的かつ系統的にその構造を表しうるように組み立てられた名称であり，慣用名は古くから慣用された名称のうち，あいまいさがなく IUPAC が使用を認めた名称である．一方，この慣用名を基本名として系統的命名規則を加えてつくった名称は半慣用名と呼ぶ[57].日本語による化合物命名は，日本化学会により「化合物名の日本語表記の原則」が，化学物質名を統一して機械的に処理できるよう制定されたが，多くの化学者にはあまり有効に利用されなかった[58]．日本語で化学物質名を書くときは IUPAC 命名法による記載をカタカナに翻字した場合が多かった．

このように，化学物質名は IUPAC 命名法に基づいた記載が多いことから，記号，数字，英字と，IUPAC 命名を字訳したときにはカタカナが使われ，さらに日本語で慣用名と半慣用名を記載する場合は漢字が加わることがわかる．IUPAC 命名が推奨されていることもあり，漢字の使用頻度は減少の傾向にある．しかしながら，現時点では漢字だけで表した化学物質名は一定以上あり，漢字を含む文字に着目した識別方法は有効であると考えた. 化学物質名を構成

する文字の種類が限定されること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．なお，この検討で扱う C, Z, G は表 4 で示した値を用いている．

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識別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識別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 4 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

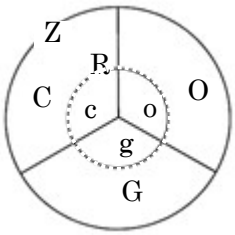

図 4　タグ付けと識別との関係

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, そこには取り出された chem 群化学物質名 c と取り出された group 群置換基名 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識別の可能性の検討を行うとともに，置換基名の化学物質名識別への影響を文字ベースの可能性について検討することとした．

3.3.1.2　1-gram の方法の結果および考察

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 9 に示す. n/N 別に該当する文字が含まれる語句を取り出し，取り出された単語の数（R_n）と取り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 10 に，再現率と適合率の関係のグラフを図 5 にそれぞれ示す.

表 9　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；取り出された単語群中の化学物質名文字の頻度(n) /取り出された単語群の文字の頻度(N)

表 10　1-gram による化学物質名の識別

n/N	取り出された単語群の数(R_n)	取り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

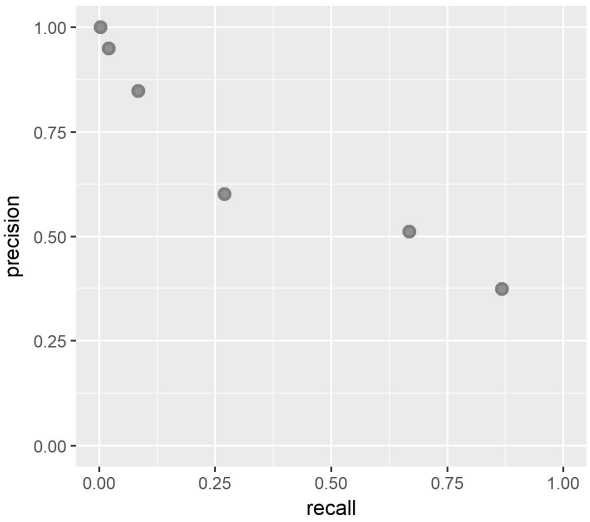

図 5　適合率–再現率の関係

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識別できるが, すべての化学物質名を高い適合率で識別することは難しいことがわかった．

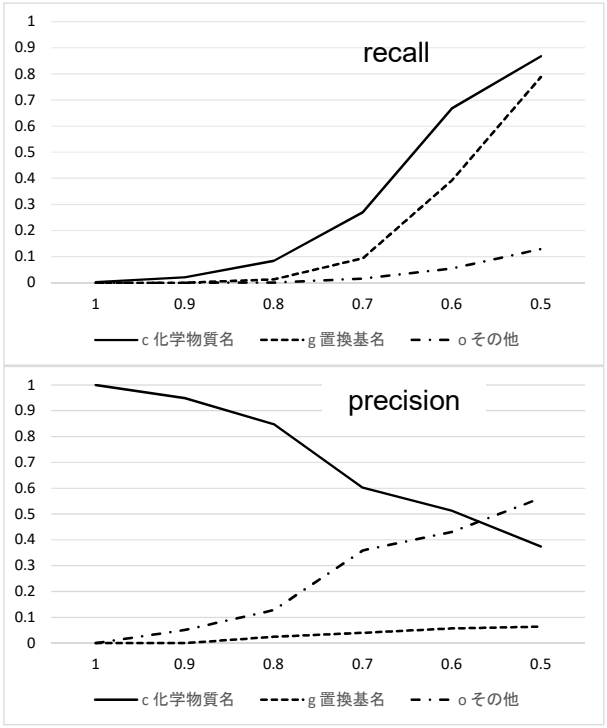

図 6　化合物名，置換基名，その他の各 n/N における適合率・再現率

n/N を 0.5 と下げても，取り出せなかった化学物質名には特有の文字が含まれていない MgCl2，cisplatin などがあった．これらは英数字だけの羅列であり，特徴のある文字種，具体的には主に化学物質名だけに使われる漢字を含んでいなかったため，識別が難しかった．

表 11 には，各条件において取り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 6 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 では取り出されず再現率・適合率が低いため化学物質名の識別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 11　置換基名およびその他の状況

n/N	取り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	取り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

化学物質名を文字単位 1-gram により識別する方法を検討したが，すべての化学物質名に適用することは難しいことがわかった．

3.3.2　機械学習

3.3.2.1　機械学習の方法

化学物質名を識別する方法として，作成したコーパスを基に，化学物質名と予めわかっている情報から，未知のデータが化学物質名であるかどうかを予測する機械学習を用いることとした．機械学習はコンピュータがデータから反復的に学習し，そこに潜むパターンを見つけ出し，学習した結果を新たなデータにあてはめることで，パターンにしたがって将来を予測することができる．機械学習は大量なデータとデータの性質に見合ったさまざまな手法の組み合わせが考えられる．本研究では，文章から単語を取り出すこと，その単語が化学物質名であるかどうかを決める二値分類を図 7 の流れで行う．なお，以降の検討は分離処理を行なった単語群を用いて行なっている．

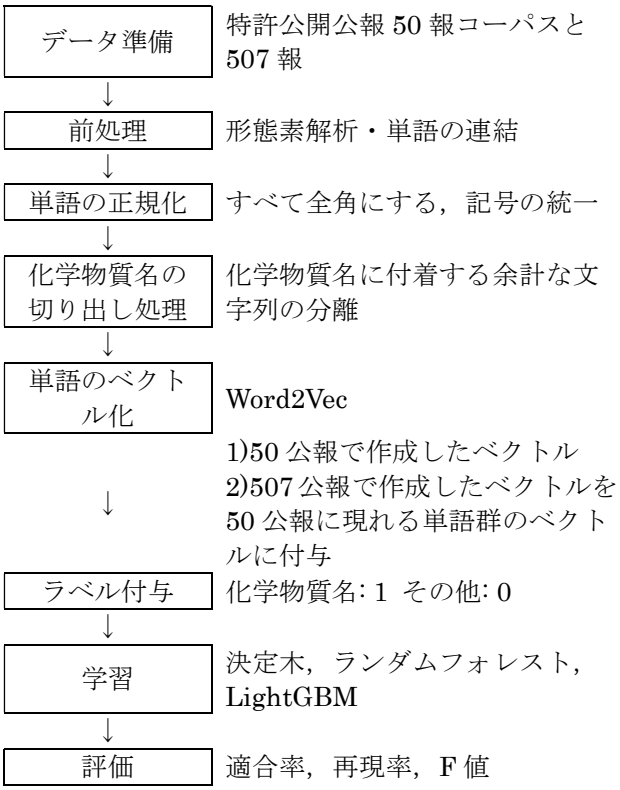

図 7　処理の流れ

切り出した単語に，文章上の配置，前後関係，の情報を加味するため，Word2Vec を用いてベクトル化を行う[59]．Word2Vec とは 2013 年，Google 社より発表された手法で，この手法のベースにあるのは，同じ文脈の中にある単語はお互いに近い意味を持つという考え方に基づいている．CHEMDNER コーパスを用いて高い性能を報告している実験においても特徴量の生成に Word2Vec を用いている[60]．また，化学物質名が頻出する特許公報や化学論文では，実験条件，実験例を示すことも多く，文脈のパターンが類似しているという点でも Word2Vec は適していると考える．

単語をベクトル化し視覚化したものを図 8 に示す．アセトンの周辺にはトルエン，エタノール，メチルエチルケトン，メチル，イソブチルケトンなど意味の似た種類の単語が近くに集まることを示している．

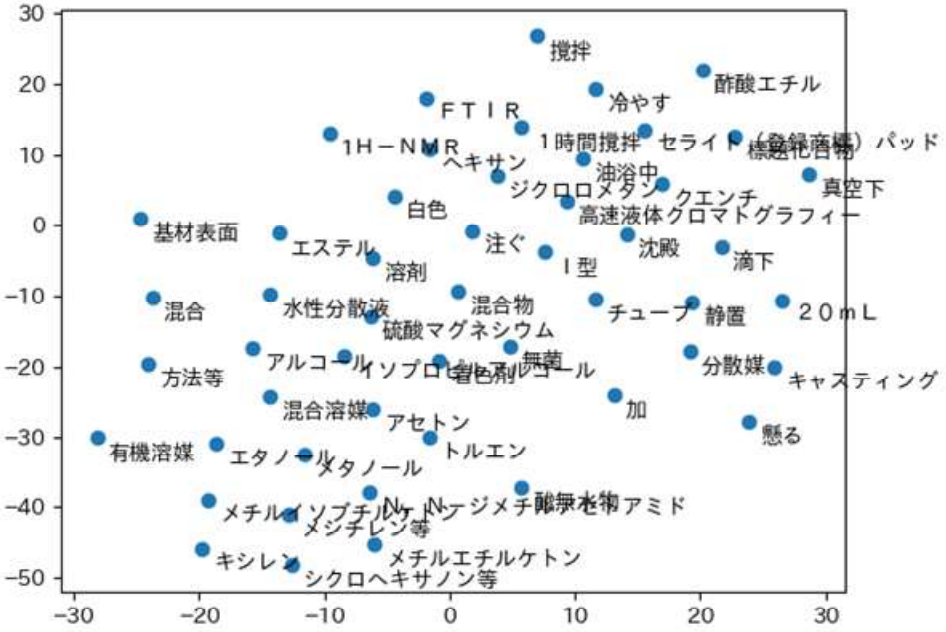

図 8　Word2Vec の例

この単語ベクトルを入力データとし，予め化学物質名とわかっている単語とそうでない単語に 2 種類のラベルをつけ，テストデータからどれくらい化学物質名が認識できるかを機械学習の手法を用いて検討する．

Word2Vec による単語ベクトルの作成は，化学物質名をタグ付けした特許公開公報 50 公報のコーパスに加えて，その 50 公報にタグ付けをしていない公報も加えて全部で 507 公報についても行う．Word2Vec は，単語間の関連性をベクトル表現に反映しているため，文書が大規模になればなるほど，より多くの文脈から精度の高い意味ベクトルが生成されると考えられる．日本語の文章に対応する単語分散表現モデルを SNS データ，Wikipedia，Web ページといった複数カテゴリを媒体とした日本語大規模コーパスから作成したモデルは wikipedia のみを学習コーパスとして用いたモデルと比較して 7 ポイント程度良い性能を得られた例が報告されている[61]．したがって，小規模な文書よりは意味ベクトルの精度が増すと考え，507 公報についても 50 公報と同様の方法で空白で区切られた単語群を作成し単語のベクトル化を行った．なお，507 公報については，化学物質名かそうでないかが分からないため，それが分かっ

ている 50 公報の単語に 507 公報で生成した単語ベクトルを付与し，大規模文書を用いることによる影響を検討する．

形態素解析と単語の連結，余計な文字列の分離を行った後，空白で区切られた単語群を Word2Vec を用いて多次元ベクトルに変換する．Word2Vec のパラメータを表 12 に示す．パラメータにはある単語から周辺の単語を予測する Skip-gram と周辺の単語からある単語を予測する CBOW のどちらの方法を用いるか，次元数，最小単語出現数，ある単語の周囲，何単語までを関連性のある単語としてみなすか，などがある．Word2Vec の周辺単語の予測を行う Skip-gram/CBOW について，Skip-gram は CBOW に比べて，計算に時間がかかるが，条件を変えても比較的よい精度が得られるとの報告があり[62]，Skip-gram で行うこととした．次元数は，大きすぎると訓練時の計算量やモデルの記憶容量が多く必要となる反面，モデルの表現力も向上する．一般的によく用いられている 50，100，200 次元を用いた．最小単語出現数は，この数字未満の出現数の単語は無視する．テキスト中に少なくとも一定回数以上出現していないと，意味のあるベクトル化ができない．また，最小単語出現数を指定することで，ほとんど出てこない単語の処理に無駄に計算時間を費やすことを避けることができる．ここでは，最小の 1 は避け，計算時間を考慮して 2 と 5 を用いた．ウインドウ幅は，テキスト中の単語が与えられた際に，その単語からどのくらい離れた単語までを周辺単語とみなすかを制御する．例えば，ウインドウ幅が 1 の場合には単語と隣接した前後の単語 2 語が周辺単語として用いられる．ウインドウ幅が増えるほど，ベクトル化の計算量が増えるため，単語を中心にその意味に影響を与える範囲を考え，10 を用いることとした．

表 12　Word2Vec のパラメータ

パラメータ[63]	説明	用いた値
sg	モデル（Skip-gram: 単語から周辺の単語を予測する，CBOW: 周辺の単語群からある単語を予測する）	Skip-gram
size	ベクトルの次元数（単語の置かれるベクトル空間の広さ）	50, 100, 200
min_count	最小単語出現数（テキスト中に出現する最小回数値）	2, 5
window	ウインドウ幅（前後の周辺単語とみなす数）	10

次に，507 公報を使って 50 公報と同様のやりかたでまず単語切り出し，連結を行い，その後分離のための処理を行う．それを用いて 50 公報と同じパラメータで Word2Vec にる単語のベクトル化を行い，得られたベクトルを 50 公報の単語群に付与したものを作成する．

生成した単語とその単語ベクトルに，化学物質名か(1)そうでないか(0)のラベルを追加する．実際のやり方を以下に述べる．

タグ付けした化学物質名を用意し，Word2Vec で生成したベクトルの単語名と用意した化学物質名データベースの単語名と一致した場合，新たに設けたラベルという列に「1」を，一致しなかった場合は「0」を入力する．完全一致の場合のみ「1」が入力されるため，タグづけされていても前後に余計な語がついた単語は「1」がつかない．この連結したテーブルにおいて，Word2Vec で生成したベクトルを示す列（50 次元であれば 50 個の列）を説明変数とし，連結して追加した化学物質名のラベルを目的変数とし，機械学習を行う．Word2Vec で生成したベクトルのデータベースと化学物質名データベースの連結を表 13 に，機械学習に用いられる説明変数と目的変数の具体例について表 14 に示す．

表 13　データベースの連結

形態素解析後連結単語

単語名	ベクトル			
例えば	-0.382337	-0.257865	…	-0.280278
ビニルシラン	-0.064069	-0.086978	…	0.004514
等	0.023807	-0.305011	…	-0.168578
の	-0.236233	-0.138371	…	0.013835
（Ｃ１）アルケニル基	-0.090783	-0.105178	…	0.009295
を	-0.064005	-0.061771	…	0.090873
有する	-0.135142	-0.19288	…	-0.080293
珪素含有化合物	-0.101084	-0.126997	…	0.003961

←連結

化学物質名

単語名	ラベル
ビニルキシレン	1
ビニルシラン	1
珪素含有化合物	1
ヒドロシラン	1
珪素化合物	1
Ｐｔ触媒	1
フタル酸	1
尿素	1

表 14　説明変数と目的変数

	説明変数				目的変数
単語	ベクトル				ラベル
例えば	-0.382337	-0.257865	…	-0.280278	0
ビニルシラン	-0.064069	-0.086978	…	0.004514	1
等	0.023807	-0.305011	…	-0.168578	0
の	-0.236233	-0.138371	…	0.013835	0
（Ｃ１）アルケニル基	-0.090783	-0.105178	…	0.009295	0
を	-0.064005	-0.061771	…	0.090873	0
有する	-0.135142	-0.19288	…	-0.080293	0
珪素含有化合物	-0.101084	-0.126997	…	0.003961	1

教師あり二値分類では高速性，正確性，大規模データに対応可能かどうかを考慮して，決定木，ランダムフォレスト，LightGBM の 3 通りの機械学習を検討した[64]．決定木はデータを最もよく分割する条件に基づいて Yes，No の質問を繰り返し，ターゲットのクラスへ分類するもので，ランダムフォレスト，LightGBM は決定木モデルを応用したモデルである．

機械学習を行うにあたり，データを学習用とテスト用に分割する必要がある．BioCreative V の CHEMDNER データセットでは学習用とテスト用に 2：1 に分割しているが，本実験では学習データ数をできるだけ多く取るために学習用とテスト用 4：1 になるように計算時のパラメータで設定する．

評価の方法は，適合率，再現率，F 値を使って行う．

3.3.2.2　機械学習の結果および考察

決定木，ランダムフォレスト，LightGBM の機械学習による学習実行の一例についてプログラムと実行画面を図 9 に示す．

```
# -*- coding: utf-8 -*-
# xml68s50.py
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import lightgbm as lgb #LightGBM

# データの準備
data = pd.read_csv("test2.csv") #utf-8
# 説明変数
X = data.loc[:, ['A1', 'A2', 'A3', 'A4', 'A5', 'A6', 'A7', 'A8', 'A9', 'A10', 'A11',
'A12', 'A13', 'A14', 'A15', 'A16', 'A17', 'A18', 'A19', 'A20', 'A21', 'A22',
'A23', 'A24', 'A25', 'A26', 'A27', 'A28', 'A29', 'A30', 'A31', 'A32', 'A33',
'A34', 'A35', 'A36', 'A37', 'A38', 'A39', 'A40', 'A41', 'A42', 'A43', 'A44',
'A45', 'A46', 'A47', 'A48', 'A49', 'A50']].values #.as_matrix()
# 目的変数
y = data['A103'].values #.as_matrix()
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
# 学習
dt = DecisionTreeClassifier()
rf = RandomForestClassifier()
lr = lgb.LGBMClassifier()
models = [ (dt,'決定木'), (rf,'ランダムフォレスト'),(lr,'LightGBM')]
for clf, name in models:
clf.fit(X_train, y_train)
pred = clf.predict(X_test) # 予測
print(name)
print(classification_report(y_test, pred))  # Precision, Recall,
F1-score # 評価
```

```
C:¥Users¥user¥Desktop>python xml68s50.py
決定木
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     86923
         1.0       0.98      0.99      0.98      1773

    accuracy                           1.00     88696
   macro avg       0.99      0.99      0.99     88696
weighted avg       1.00      1.00      1.00     88696

ランダムフォレスト
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     86923
         1.0       1.00      0.98      0.99      1773

    accuracy                           1.00     88696
   macro avg       1.00      0.99      0.99     88696
weighted avg       1.00      1.00      1.00     88696

LightGBM
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     86923
         1.0       0.97      0.83      0.90      1773

    accuracy                           1.00     88696
   macro avg       0.99      0.91      0.95     88696
weighted avg       1.00      1.00      1.00     88696
```

図 9　機械学習の実行画面

(Python version 3.8.3 / Windows 10)

表 15 ではベクトルの次元数，最小単語出現数，ウインドウ幅を変化させた場合の学習アルゴリズムによる評価結果を示す．表中，テストデータの個数はテストデータにおいて化学物質名としてラベル付けされた単語の数とその横に全体の単語数を括弧で示す．

表 15　学習アルゴリズムによる適合率，再現率，F 値の比較（上段 適合率/中段 再現率/下段 F 値）

	テストデータの個数	決定木	ランダムフォレスト	LightGBM
size 50 min_count 2 全体の数(88696)	1718	0.98 0.98 0.98	1.00 1.00 1.00	0.96 0.83 0.89
size 50 min_count 5 全体の数(84089)	1297	1.00 1.00 1.00	1.00 1.00 1.00	1.00 1.00 1.00
size 100 min_count 2 全体の数(88696)	1824	0.98 0.99 0.99	1.00 0.98 0.99	0.98 0.86 0.92
size 100 min_count 5 全体の数(84089)	1240	1.00 1.00 1.00	1.00 1.00 1.00	1.00 1.00 1.00
size 200 min_count 2 全体の数(88696)	1790	0.98 0.99 0.98	1.00 0.98 0.99	0.97 0.87 0.92
size 200 min_count 5 全体の数(84089)	1213	1.00 1.00 1.00	1.00 1.00 1.00	1.00 1.00 1.00

なお, 化学物質名かそうでないかを区別する二値分類ではロジスティック回帰が用いられることもあるが，本実験では「1」を予測できるデータがなく適合率が 0 となり，F 値を計算できなかった．ロジスティック回帰は直線で分離できるような線形モデルには適しているため, 今回の事例には合わなかったと考えられる．

決定木の手法では最小単語出現数を 5 と設定した場合，どの次元数でも F 値が 1.0 となった．ランダムフォレストの手法では決定木の結果と同様，最小単語出現数を 5 と設定した場合，どの次元数でも F 値が 1.0 になり，加えて最小単語出現数が 2 の 50 次元でも 1.0 となった．一方，LightGBM の手法では，最小単語出現数を 5 に設定した場合，どの次元数でも F 値が 1.0 となったのは他の 2 法と同じであるが，最小単語出現数が 2 の場合は F 値が 0.9 前後であった．

Word2Vec パラメーターが次元数 50，最小単語出現数 2 において，決定木の手法で，木の深さ（デフォルト None の場合，ノードはすべてのリーフが 1 になるまで展開される）とノードを分割する必要な最小サンプルサイズ（デフォルト 2）を変化させた結果を表 16 に示す．いずれの最小分割サイズにおい

ても木の深さを 50 と設定した場合，F 値は最大となり，その後深くしても値は変わらなかった．木の深さは 50 が妥当と考える．最小分割サイズでは顕著な違いは見られなかった.

表 16　決定木の木の深さと最小分割サイズによる適合率, 再現率, F 値の比較（上段 適合率/中段 再現率/下段 F 値）

最小分割サイズ＼木の深さ	10	20	50	100	None
2	0.95 0.55 0.70	0.98 0.86 0.92	0.98 0.98 0.98	0.98 0.98 0.98	0.98 0.98 0.98
10	0.92 0.55 0.69	0.97 0.86 0.91	0.97 0.96 0.97	0.98 0.95 0.96	0.96 0.97 0.97
20	0.92 0.56 0.69	0.95 0.85 0.90	0.95 0.94 0.95	0.94 0.94 0.94	0.95 0.92 0.93

決定木，ランダムフォレスト，LightGBM の 3 通りの機械学習において，ランダムフォレストの手法が最もよい F 値が得られた. Word2Vec ベクトルの次元数は 50, 100, 200 と大きくなるにしたがって良好な結果が得られた．最小単語出現数では 5 に設定すると，すべての手法で F 値が 1.0 となった.

ランダムフォレスト，LightGBM は決定木を応用したモデルであるため，同様な傾向の結果となったと考えられる．F 値が BioCreative では 0.9 程度に対し，1.0 は高い結果となったが，これは 50 公報という少ないデータ数が木構造による学習に対して適応したと考えられる．BioCreative のように 20,000 程度と大規模なデータの場合は，木構造による分離方法では限界があると考えられる.

機械学習では, 試行するごとに全データをランダムに学習データとテストデータを 4:1 になるように計算時のパラメータでランダムに設定しているため，試行ごとに異なる結果が得られ, 取り出せなかった物質名の特徴を特定することはできなかった.

さらに 50 公報の単語に 507 公報の単語ベクトルを付与した場合の結果を表 17 に示す.

表 17　50 公報，507 公報による F 値の比較（上段 50 公報/下段 507 公報）

	テストデータの個数	決定木	ランダムフォレスト	LightGBM
size 50 min_count 2	1718 2106	0.98 0.86	1.00 0.91	0.89 0.68
size 50 min_count 5	1297 1714	1.00 0.93	1.00 0.96	1.00 0.82
size 100 min_count 2	1824 2095	0.99 0.85	0.99 0.91	0.92 0.71
size 100 min_count 5	1240 1633	1.00 0.93	1.00 0.95	1.00 0.85
size 200 min_count 2	1790 2056	0.98 0.86	0.99 0.90	0.92 0.72
size 200 min_count 5	1213 1685	1.00 0.93	1.00 0.96	1.00 0.86

10 倍強の 507 公報を取り出して，50 公報の単語ベクトルを 507 公報の単語ベクトルで置き換えた結果，抽出率は低くなった．しかしながら，大規模な単語群でベクトルを作っても，F 値でも最低 0.68 とある程度の結果がでた．この違いは今回の 50 公報と 507 公報は特許公報化学分野をランダムに選択しているものの，50 公報の選択では数が少なかったことや，分野に偏りがあり，文章の表現パターンが少なかったため，50 公報では過学習になったとも考えられる．大規模なものになると多様な文書表現が出現し，類似の意味を持つベクトルの値が大きく異なるなどベクトルが明確に区別できなくなる可能性がある．分野の範囲を広げ，かつより大規模な文章の収集を行うことにより，今回試みたベクトルの置き換えが，抽出率に影響を与えるかどうかは今後も検討される必要があると考える．

第 4 章　総合考察

データ準備，前処理，化学物質名の切り出し処理，単語のベクトル化，機械学習の一連の流れにおいて, 失敗分析を行って, 今回用いた材料, 方法を変更，改善すべき点など検討する必要があると考える.

データ準備では, 化学物質名が出現する確率が高いと思われる，国際特許分類をしぼってコーパスを作成しており，数も 50 公報と少ない．多様な日本語文章から化学物質名を抽出するためには, 採択する国際特許分類の見直し，コーパスの数の増加が望まれる. なお, 特許公開公報は明細書, 特許請求の範囲，要約によって内容が分かれている．さらに明細書は技術分野，背景技術，発明の概要，図面の簡単な説明，発明を実施するための形態，実施例と分かれているため, 記載場所によって化学物質名の抽出の傾向が変わると考えられる．それについても今後検討が必要であろう.

日本語文章中に出現する化学物質名の抽出結果の評価を行うため, 化学物質名をタグ付けしたコーパスを特許公開公報の化学分野の電子データを利用して作成した. コーパスは分離処理の過程で, 多様な意味で用いられている「水」を含む語や「アルカリ金属（Ｎａ、Ｋ等）塩」などの複合語については外すなどタグ付けを変更した．このようにコーパスの作成は，人的に行うためゆれが生じる. より高い精度で作成するにはコストがかかるが複数人による方法などの検討が必要であろう．ただ，BioCreative IV で人間のアノテーター間の合意比率が 91%であったことを考えると本研究で用いたコーパスの精度は十分高いともいえよう.

どのように低コストでコーパスを作成するかは大きな課題である．BioCreative V で用意された CHEMDNER コーパスは，医薬化学特許の発明の名称と要約に化学物質名を手作業でアノテーションしたコーパスを作成している．特許公報は各国で同様の内容が公開されているものがあるため，CHEMDNER コーパス[65]を日本語に翻訳すれば日本語コーパスとして用意できる．CHEMDNER コーパスの作成に用いた公報番号は国際公開公報 WO で始まるものは約 1/4 であり，その一部分は日本語でも公開されている．具体的には，WO2014092061 HYDANTOIN DERIVATIVE は日本語で特許第 5951799 号ヒダントイン誘導体として公開されており，CHEMDNER のコー

パスを翻訳して日本語のコーパスとして利用できる. 公報が対応していなくても，CHEMDER コーパスを「化合物名の字訳（日本語名）」[66]により日本語に置き換えることも検討できると考える.

前処理の形態素解析では，MeCab と日化辞のデータをもとに作成したユーザー辞書を用いた．ユーザー辞書は 15654 個である．今回用いた 50 個の特許公開広報に現れた化学物質名については対応できたが, より大規模なものに出現する全ての化学物質名にうまく適用できるとは限らない. 生命科学系データベースアーカイブの科学技術用語形態素解析辞書[67]の日化辞辞書を用いるとさらに細かい形態素解析が可能であると考えられるが, 大規模なコーパスについてはその点についての検討も必要である思われる.

正規表現による置換を用いた分離処理は 8 割の化学物質名を切り出せたことから有効と考えたが，残された 2 割の化学物質名についても取り出せるよう改善することが最も大きな課題である．できなかった残りの 20%のデータの抽出には付加語辞書（誘導体，組成物），単位辞書（重量%，モル）など，物質名に付加する辞書を準備する必要があると考える．また，取り出せなかった化学物質名には複合語や助詞が含まれている場合が多い. これらは分離処理のほかに，係り受けを考慮した言い換えなどの処理が考えられる.

文章中の単語について Word2Vec 手法を用いて全体の単語群をベクトル化し，化学物質名かそうでないかのデータを加えて二値分類を行った．決定木，ランダムフォレスト，LightGBM の 3 つの機械学習手法で検討した結果，いずれも高い F 値が得られた．これは，Word2Vec の条件，機械学習の手法による違いがあるが, 機械学習モデルを用いた化学物質名の自動抽出が一定の効果があると示唆するものである.

機械学習は，二値分類で主に利用されているアルゴリズムを用いている．決定木，ランダムフォレスト，LightGBM についてはその分岐の過程を分析をする必要がある.

大規模文章（507 公報）を用いたとき，ランダムフォレストでは F 値が最低で 0.91 であった．一方，50 公報のランダムフォレストの値は 1.00 であった．2 割の切り出せていない化学物質名があり，それらも似たようなベクトルを持っていることを考えると, 小規模コーパスでは過学習の影響が考えられる. コーパスのサイズをさらに拡大することにより，過学習を避けることができ，

この一連の方法による自動抽出の可能性を示唆するものである.

今回の実験は,形態素解析と連結により取り出すことができた全体の化学物質名に対して，余計な文字を除いて同定することができた 80%のデータに対する成果である.余計な文字が付いているために化学物質名としてラベル化できなかった残りの 20%のデータの影響も今後検討する必要があると考える.

ルールベースと機械学習で取り出せなかった化学物質名を比べた結果,ルールベースでは文字の構成に特徴があれば認識できるが,英数字だけの羅列であると認識が難しい．機械学習は文字の特徴ではなく，文構造に依存するため,妥当性の高い認識が行えていると考える．ただ，抽出率は上げるために，ルールベースと機械学習を併用して用いるという方法も考えられる.

今回は決定木，ランダムフォレスト，LightGBM という 3 通りの機械学習を用いたが、いずれも決定木を基本とした方法である． これらの方法で一定の精度が得られたが，大規模なコーパスを用いた場合は 507 公報の事例の様に精度が落ちることが予想される．CRF や HMM，MEMM などの機械学習の適応についても今後検討する必要があろう.

また機械学習には，大規模なデータを収集し，コーパスの必要のない教師なし学習によって固有表現抽出を行う BERT[68]という選択肢もあるので,その適用について検討してみることも必要であろう.

特許公開公報において化学物質名の記載は論文よりも書き手に委ねられている傾向が強く,「酢酸エチル」を「酢エチ」と表記するなど，同一物質にもかかわらず異なる表記となっている場合もある.化学物質名かどうかという判定では問題ないが,今後,抽出した化学物質名をファクトデータベースに取り入れるには異表記問題を解決する必要がある.

第 5 章　結論

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. 日本語文章の特徴を考えると，文章の形態素解析による語の切り出し，切り出した語の連結による化学物質名を含む単語の作成, 連結して得られた単語群からの化学物質名の識別という段階が必要であった.

化学物質名を識別する方法についてルールベースと機械学習による方法を検討した. ルールベースでは化学物質名を構成する文字に着目し出現頻度を調べたところ, 化学物質名には通常の文章にはあまり出現しない文字が含まれることがわかった．機械学習では文章中の単語をベクトル化することによって，化学物質名のベクトルと化学物質名以外のベクトルを分類することがある程度可能であることが明らかになった.

日本語文章の形態素解析による語の切り出しの手順は, 特許公開公報の文章の形態素解析を行い，形態素に分離し，記号，接頭詞，名詞が連続していた場合，それらを連結した．次に化学物質名に付着する余計な文字列の分離のために，一連の正規表現による置換を行った．これにより，タグ付けした化学物質名のうち約 80%の化学物質名を同定することができた.

ひとかたまりの単語が化学物質名か否かを判定する方法として, 化学物質名の特徴を考慮したルールベースによる識別を行った.

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，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.

もう一つの方法として, 先行研究から主流となっている機械学習を用いることを考え，文章中の単語について Word2Vec 手法を用いて全体の単語群をベクトル化し，化学物質名かそうでないかのデータを加えて二値分類を行った. 決定木，ランダムフォレスト，LightGBM の 3 つの機械学習手法で検討した結果，いずれも高い F 値が得られた.

50 公報の単語ベクトルを 507 公報の単語ベクトルで置き換えた場合，F 値は下がる傾向が見られたが一定の高い値であった.

文章から固有名詞を取り出す研究は古くから行われてきたが, 日本語の文章,

その中でも化学分野における固有名詞を抽出する研究はまだ少ない. 化学分野において, その中核となるのは化学物質名であり, 本研究では日本語文章の化学物質名にしぼって, その抽出法に関する検討と提案を行った.

日本語と英語の化学物質名の表記を比較し, 先行研究の進んでいる海外の文献に基づき, コーパスの必要性があるとの認識から特許公開公報から日本語のコーパスを作成し, 抽出方法としてルールベースと機械学習を検討し, 機械学習ではある程度の成果が見られた. なお, 抽出方法を検討する前には, 化学物質名として同定する過程が必要で, 形態素解析後の語のまとめ, 正規表現による置換により, 抽出方法を検討する語の準備を行った.

本研究では化学物質名の抽出について述べているが, 抽出後には, 化学物質名の特徴でも示したように, 異表記問題を解決する必要がある. さらに化学物質名とともに, その構造, 機能, 製造方法, 化学反応, 用途などの多様な属性を抽出することにより, ファクトデータベースの自動構築に向けて準備ができる. 今回生成した意味ベクトルは, 文章において前後の関係を考慮しているため, 化学反応, 用途などにも応用が考えられる. 化学物質の新規用途の探索にも期待できる.

化学物質名抽出法ではルールベースと機械学習の方法を検討し, 機械学習の手法ではある程度の成果が見られた. 化学物質名の抽出はファクトデータベース構築に至る第一歩であり, 化学物質の属性, 化学反応, 用途などの抽出などの課題がある. 大規模なデータベースが構築できれば, 新規物質の開発, 医薬品の探求, 翻訳などさまざまな分野で応用が期待される. 海外の BioCreative のようなワークショップで関連するタスクを設定し, ファクトデータベース自動構築の手順が確立されることが期待される.

網羅的なデータベースを迅速に構築するには, コンピュータの性能の進化, データの増加, 抽出法の開発が望まれる. 日本では新エネルギー・産業技術総合開発機構「化学・情報科学の融合による新化学創成に向けて」で提言されるように, 後押しされるようになってきた. NLM でも 2021 年に, Chem, Pubmed 全文文献における NER のための新しいリソースを提供している[69].

日本においても, 特許公開公報のように英語では得られない文献からデータを抽出してファクトデータベースを構築するためには, 日本語文章における化学に関する NER 研究のさらなる発展が望まれる.

謝辞

本研究に関しまして，中山伸一教授には，文章からの化学物質名の抽出という大規模な作業に対して化学物質名の定義や抽出の対象，手法に道筋をたてていただき，順序だてて取り組むことができました．芳鐘冬樹教授，真榮城哲也教授には研究方針，手法に関しまして助言をいただきました．皆様に心から感謝申し上げます．また本研究にあたり，「日本化学物質辞書 Web」，「Google Colaboratory」，「github」など各種サイトを使用させていただきました．提供してくださった方々にお礼申し上げます．

文献リスト

[1] 小野寺夏生. ファクトデータ，ファクトデータベースあれこれ. 情報知識学会誌, Vol. 27, No. 3, pp. 275-280, 2017.

[2] Q. Wang, *et al*., *Database*, pp.1–18. 2016.

[3] 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.

[4] 新エネルギー・産業技術総合開発機構，Connected Industries 推進のための協調領域データ共有・ AI システム開発促進事業, https://www.nedo.go.jp/activities/ZZJP_100157.html

[5] Eltyeb, Safaa; Salim, Naomie: "Chemical named entities recognition: a review on approaches and applications", Journal of Cheminformatics, pp.6-17, 2014.

[6] IUPAC Nomenclature of Organic Chemistry http://www.acdlabs.com/iupac/nomenclature (2021.10.10 参照)

[7] 日化辞（日本化学物質辞書）Web　https://jglobal.jst.go.jp/info/nikkaji (2021.10.10 参照)

[8] PubChem　https://pubchem.ncbi.nlm.nih.gov/ (2020.10.10 参照)

[9] Grishman, R.; Sundheim, B. Message Understanding Conference-6: A Brief History. In Proceedings of the 16th conference on Computational linguistics - Volume 1[10] Klein C: Information Extraction from Text for Improving Research on Small Molecules and Histone Modifications, Ph.D. thesis. Bonn, Germany: Universitäts-und Landesbibliothek; 2011.

[11] Humphreys K, Gaizauskas R, Azzam S, Huyck C, Mitchell B, Cunningham H, Wilks Y: University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In Proceedings of the Seventh Message Understanding Conferences (MUC-7); 1998.

[12] Budi I, Bressan S: Association rules mining for name entity recognition. In Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference; 2003:325-328.

[13] Bikel DM, Schwartz R, Weischedel RM: An algorithm that learns what's in a name. Mach Learn, 34:211-231, 1999.

[14] Borthwick A: A maximum Entropy Approach to Named Entity

Recognition. Ph. D. thesis, New York University: New York University; 1
[15] Chieu HL, Ng HT: Named entity recognition: a maximum entropy approach using global information. In Proceedings of the 19th International Conference on Computational linguistics-Volume 1 pp. 1-7, 2002.
[16] Ayodele TO: Types of machine learning algorithms. 2010, Internet: http://www.intechopen.com/articles/show/title/ types-of-machine learning-algorithms.
[17] Mansouri A, Affendey LS, Mamat A: Named entity recognition approaches. Int J Comp Sci Netw Sec, 8, pp. 339-344, 2008.
[18] Campos D, Matos S, Oliveira JL: Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. ; 2012.
[19] Nadeau D: Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision; 2007.
[20] Garfield, E. An Algorithm for Translating Chemical Names to Molecular Formulas. J. Chem. Doc., 2 (3), pp. 177−179, 1962.
[21] Zamora, E. M.; Blower, P. E., Jr Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistics Techniques. 1. Lexical and Syntactic Phases. J. Chem. Inf. Model. 24 (3), pp. 176−181, 1984.
[22] Reeker, L. H.; Zamora, E. M.; Blower, P. E. Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions. Proceedings of the first conference on Applied natural language processing (ANLC '83); Santa Monica, CA, February 1−3, pp 109−116, 1983.
[23] Hodge, G. M.; Nelson, T. W.; Vleduts-Stokolov, N. Automatic Recognition of Chemical Names in Natural-Language Texts; Presented at the 197th National Meeting of the American Chemical Society, Dallas, TX, April 7−9, 1989; paper CINF-17.
[24] Ai, C. S.; Blower, P. E., Jr; Ledwith, R. H. Extraction of Chemical Reaction Information from Primary Journal Text. J. Chem. Inf. Model., 30 (2), 163−169, 1990.
[25] Kemp, Nick.; Lynch, Michael. 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.

[26] 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.

[27] Zhang, Yaoyun, et al.: "Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning", Database, pp.1-10, 2016.

[28] Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Mulligen EM, Kleinjans J, Kors JA: A dictionary to identify small molecules and drugs in free text. Bioinformatics, 25:2983–2991, 2009.

[29] Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, StoehrP: EBIMed—text crunching to gather facts for proteins from Medline. Bioinformatics, 23:e237–e244, 2007.

[30] Narayanaswamy M, Ravikumar KE, Vijay-Shanker K: A biological named entity recognizer. Pac Symp Biocomput, 427, 2003.

[31] Tkachenko M, Simanovsky A: Named entity recognition: Exploring features. Proceed KONVENS 2012, 118-127. http://www.oegai.at/ konvens2012/proceedings/17_tkachenko12o/.

[32] Wang H, Zhao T, Tan H, Zhang S: Biomedical named entity recognition. Int J Mach Learn Cybern. 9(3):373–82, 2018.

[33] Huber T, Rocktaschel T, Weidlich M, Thomas P, Leser U: Extended Feature Set for Chemical Named Entity Recognition and Indexing. In BioCreative Challenge Evaluation Workshop vol. 2; 2013:88.

[34] Corbett P, Batchelor C, Teufel S: Annotation of Chemical Named Entities, BioNLP 2007: Biological, translational, and clinical language processing, pp. 57-64, 2007.

[35] Rocktaschel T, Weidlich M, Leser U: ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28:1633-1640, 2012.

[36] Wallach HM: Conditional random fields: An introduction. Tech Rep (CIS),22, 2004.

[37] Lafferty J, McCallum A, Pereira FCN: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data; 2001.

[38] McCallum A, Freitag D, Pereira FCN: Maximum entropy Markov models

for information extraction and segmentation. In Proceedings of the Seventeenth International Conference on Machine Learning. pp. 591-598, 2000.

[39] M. Krallinger *et al*., *Journal of Cheminformatics*, 7(Suppl 1):S1, 2015.

[40] M. Krallinger *et al*., *Journal of Cheminformatics*, **7**(Suppl 1):S2, 2015.

[41] Krallinger, Martin et al., "Information Retrieval and Text Mining Technologies for Chemistry", Chem. Rev., Vol.117, No.12, pp.7673-7761, 2017.

[42] V Yadav, S Bethard A survey on recent advances in named entity recognition from deep learning models COLING 2018 C18-1182 arXiv:1910.11470, 2019.

[43] Awan, Z., Kahlke, T., Ralph, P. and Kennedy, P. Chemical Named Entity Recognition with Deep Contextualized Neural Embeddings. DOI: 10.5220/0008163501350144 In Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019), pages 135-144, 2019.

[44] iiWAS2019: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services December 2019 Pages 617–621https://doi.org/10.1145/3366030.3366104

[45] 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.

[46] 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. 2421-2433, 1998.

[47] 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, 1997.

[48] 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.

[49] 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.

[50] 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.

[51] https://sites.google.com/site/extendednamedentity711

[52] 邊土名朝飛, 野中尋史, 小林暁雄, 関根聡. 外部知識源を使用した Wikipedia からの化合物情報抽出. 言語処理学会第 25 回年次大会発表論文集,

pp. 791-794, 2019.
[53] 特許庁 インターネット利用による公報発行サイト https://www.publication.jpo.go.jp/ (2021 年 12 月 5 日参照)
[54] http://taku910.github.io/mecab/ (2018 年 9 月 17 日参照)
[55] https://dbarchive.biosciencedbc.jp/jp/nikkaji/download.html (2018 年 9 月 17 日参照)
[56] 田中るみ子, 中山伸一. 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. 情報知識学会誌, Vol. 29, No. 3, pp. 238-246, 2019.
[57] 廖春栄. 生化学物質名称のつけ方. 三共出版, 1988.
[58] 畑一夫. 書く人と読む人のための化合物－情報検索に備えて. *CICSJ Bulletin*, Vol.14, No.4, Aug, 1996.
[59] T. Mikolov *et al*., **2013**, *arXiv preprint* arXiv:1301.3781.
[60] L. Luo *et al*., *Bioinformatics*, **34**, Issue 8, pp.1381–1388, 2018.
[61] 松野省吾, 水木栄, 榊剛史, 日本語大規模 SNS+Web コーパスによる単語分散表現のモデル構築, 2019 年度人工知能学会全国大会（第 33 回）
[62] 谷川原綾子, 佐藤哲太, *Journal of Nippon Hoshasen Gijutsu Gakkai (Japanese journal of radiological technology)*, **76**, pp.1118-1124, 2020.
[63] URL of gensim: https://radimrehurek.com/gensim/
[64] 機械学習アルゴリズム選択ガイド https://blogs.sas.com/content/sasjapan/2017/11/21/machine-learning-algorithm-use.
[65] https://biocreative.bioinformatics.udel.edu/resources/biocreative-v/proceedings-biocreative5/ (2021 年 12 月 9 日参照)
[66] http://nomenclator.la.coocan.jp/chem/jiyaku/jiyaku.htm(2021 年 12 月 9 日参照)
[67] https://dbarchive.biosciencedbc.jp/jp/mecab/data-3.html (2021 年 12 月 19 日参照)
[68] Jacob Devlin, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 [cs.CL], 2018.
[69] Rezarta Islamaj, et al. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. Scientific data, 8(1), 2021.

全研究業績のリスト

[1] 田中るみ子, 中山伸一「特許公開公報に出現する化学物質名の分析」ケモインフォマティクス討論会予稿集 第 40 回, 2017.

[2] 田中るみ子, 中山伸一「特許公開公報文章からの化学物質名の切出しと選別法の検討」ケモインフォマティクス討論会予稿集 第 41 回, 2018.

[3] 田中るみ子, 中山伸一「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」 情報知識学会誌, Vol. 29, No.3, 238-246, 2019.

[4] 田中るみ子, 中山伸一「特許公開公報文章からの化学物質名の抽出」Journal of Computer Chemistry, accepted, 2021.

付録

用語・略語

（用語）

切り出し	日本語文章から単語を取り出すこと
形態素	言語として意味が取れる最小単位
識別	化学物質名を他の単語と区別すること
単語	一定の意味を持ち、文法的にも一定の機能を持つ最小の言語単位
抽出	日本語文章から単語を取り出し，化学物質名を他の単語と区別すること

（略語）

ATC	automatic text categorization
CAS	Chemical Abstracts Service
CER	chemical entity recognition
CHEMNDER	chemical compound and drug name recognition
CRF	conditional random field
DARPA	Defense Advanced Research Projects Agency
DBMS	database management systems
DT	decision tree
FDA	Food and Drug Administration
HMM	hidden Markov model
IPC	international patent classification
IUPAC	International Union of Pure and Applied Chemistry
InChI	international chemical identifier
JSON	JavaScript Object Notation
KEGG	Kyoto Encyclopedia of Genes and Genomes
LSTM	Long Short Term Memory
MEMM	Maximum Entropy Markov Models
ML	machine learning
MUC	message understanding conferences
MedDRA	Medical Dictionary of Regulatory Activities
NCI	National Cancer Institute
NDA	new drug application

NE	named entities
NER	named entity recognition
NLM	National Library of Medicine
NLP	natural language processing
PCT	Patent Cooperation Treaty
PMC	PubMed Central
POS	part-of-speech
RDF	resource description framework
RF	random forest
SL	supervised learning
SLN	SYBYL line notation
SMILES	simplified molecular-input line entry system
SVM	support vector machine
UL	unsupervised learning
UMLS	Unified Medical Language System
URI	Uniform Resource Identifier
USAN	United States Adopted Names
USPTO	United States Patent and Trademark Office
UTF-8	8-bit Unicode Transformation Format
WIPO	World Intellectual Property Organization

付表　コーパス作成に用いた特許公開公報リスト

番号	公開番号	発明の名称	住所	出願人
1	2016-129515	条件的に不死化された長期幹細胞ならびにそのような細胞を作製する方法および使用する方法。	米国	ナショナル　ジューイッシュ　ヘルス
2	2016-129862	水処理装置	大阪	東洋紡株式会社
3	2016-129882	有機凝結剤	京都	三洋化成工業株式会社
4	2016-129977	帳票	東京	凸版印刷株式会社
5	2016-130183	レピドクロサイト型チタン酸塩及びその製造方法、それを含有する無機複合材、樹脂組成物並びに摩擦材	大阪	大塚化学株式会社
6	2016-130193	水素生成装置およびそれを用いた燃料電池システム並びにその運転方法	大阪	パナソニックＩＰマネジメント株式会社
7	2016-130203	急結剤及びその製造方法	東京	デンカ株式会社
8	2016-130213	遷移金属窒化物および遷移金属窒化物の合成方法	米国	シックスポイント　マテリアルズ，　インコーポレイテッド
9	2016-130234	機能剤	東京	白井松新薬株式会社
10	2016-130249	Ｃ５ａＲアンタゴニスト	米国	ケモセントリックス，インコーポレイティド
11	2016-130271	粘着剤組成物、粘着剤層、粘着シートおよび光学フィルム	大阪	日東電工株式会社
12	2016-130281	衝撃緩衝用部材の製造方法および衝撃緩衝用部材	大阪	株式会社カネカ
13	2016-130291	難燃性ポリカーボネート樹脂組成物	大阪	帝人株式会社
14	2016-130301	加圧式汚泥脱水機用洗浄剤および加圧式汚泥脱水機の洗浄方法	京都	三洋化成工業株式会社
15	2016-130311	エアロゾルジェット印刷のためのソルダーマスク組成物	米国	ゼロックス　コーポレイション
16	2016-130321	樹脂成形体用材料、及び樹脂成形体の製造方法	東京	三菱化学株式会社
17	2016-130331	溶銑物流計画方法および溶銑物流計画装置	東京	ＪＦＥスチール株式会社
18	2016-130341	マグネタイト鉱石を用いた焼結鉱原料の製造方法	兵庫	株式会社神戸製鋼所
19	2016-130351	円筒形材料およびその製造方法	東京	株式会社徳力本店
20	2016-130361	缶用鋼板及び缶用鋼板の製造方法	東京	ＪＦＥスチール株式会社

21	2016-130372	樹脂組成物、樹脂シート、積層シート及び発泡壁紙	東京	株式会社トッパン・コスモ
22	2016-130783	偏光子保護フィルム用ポリエステル樹脂組成物	東京	東レ株式会社
23	2016-130861	合わせガラス板上に表示画像を生成するための方法、装置およびその装置を備えた建物、自動車、航空機、ヘリコプターまたは船舶	フランス	サンーゴバン　グラス　フランス
24	2016-131193	積層板の製造方法、及び多層回路基板	東京	宇部エクシモ株式会社
25	2016-131244	樹脂フィルム、支持体付き樹脂フィルム、プリプレグ、金属張積層板及び多層印刷配線板	東京	日立化成株式会社
26	WO2014/020939	陽極酸化ポーラスアルミナ、アルミナスルーホールメンブレンおよびそれらの製造方法	神奈川	公益財団法人神奈川科学技術アカデミー
27	WO2014/021084	耐熱性樹脂複合体およびその製造方法、ならびに耐熱性樹脂複合体用不織布	岡山	株式会社クラレ
28	WO2014/021205	新規乳酸菌	大阪	株式会社カネカ
29	WO2014/021257	グラフェンとカーボンナノチューブからなる複合フィルムの製造方法	長野	国立大学法人信州大学
30	WO2014/021316	ランダムマットおよび繊維強化複合材料成形体	大阪	帝人株式会社
31	WO2014/021351	微生物検出法及び微生物検出キット	東京	森永乳業株式会社
32	WO2014/021388	ポリアクリル酸（塩）系吸水性樹脂粉末を用いた吸水剤及びその製造方法	大阪	株式会社日本触媒
33	WO2014/021419	硬化性樹脂組成物	東京	横浜ゴム株式会社
34	WO2014/021459	合わせガラス用中間膜及び合わせガラス	大阪	積水化学工業株式会社
35	2016-521114	線維化疾患治療において有用な分子標的及び化合物、並びにこれらの同定方法	ベルギー	ガラパゴス・ナムローゼ・フェンノートシャップ
36	2016-521125	新規なバチルス株及び組成物	米国	エンヴェラ　エルエルシー
37	2016-521195	セラミックフィルタ	米国	パイロテック　インコーポレイテッド
38	2016-521222	硬化接着剤シートを含む積層体の製造方法	米国	スリーエム　イノベイティブ　プロパティズ　カンパニー

39	2016-521241	ヒドロクロロシラン生産における腐食及びファウリングを低減する方法	米国	アールイーシー　シリコン　インコーポレイテッド
40	2016-521251	多面体オリゴマー状シルセスキオキサンナノ結晶安定化リガンド	米国	ナノシス・インク.
41	2016-521262	癌のための併用療法	米国	イーライ　リリー　アンド　カンパニー
42	2016-521295	炭素繊維含有樹脂から炭素繊維を回収するための熱分解システム及び方法	ドイツ	イーエルジー　カーボン　ファイバー　インターナショナル　ゲーエムベーハー
43	2016-521305	シリコーン及び脂肪酸アミドスリップ剤を有するポリマー組成物	米国	ダウ　グローバル　テクノロジーズ　エルエルシー
44	2016-521316	能動的に位置合わせされるファインメタルマスク	米国	アプライド　マテリアルズ　インコーポレイテッド
45	2016-521374	多層ミラーアセンブリ	イタリア	ソルベイ　スペシャルティ　ポリマーズ　イタリー　エス．ピー．エー.
46	2016-131495	尿臭消臭剤	東京	ライオン商事株式会社
47	2016-131516	ターミネータ配列とプロモータ配列とを順次備えた線状二本鎖ＤＮＡ	山口	国立大学法人山口大学
48	2016-131541	トランスジェニック非ヒト哺乳動物及びその用途	愛知	学校法人藤田学園
49	2016-131902	ポリアクリル酸（塩）系吸水剤の製造方法	大阪	株式会社日本触媒
50	2016-131932	排気ガス処理装置及び基板処理装置	東京	住友化学株式会社