外部知識源を使用したWikipediaからの化合物情報抽出

邊土名 朝飛[1]　野中 尋史[1]　小林 暁雄[2]　関根 聡[2]
[1]長岡技術科学大学　[2]理研 AIP
s1773348@stn.nagaokaut.ac.jp　nonaka@kjs.nagaokaut.ac.jp
{akio.kobayashi, satoshi.sekine}@riken.jp

1 はじめに

化合物の構造化された情報は，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．「関根の拡張固有表現階層 [1]」では，化合物の属性について種類・原材料・製造方法・別名・用途・特性の 6 つを定義しているが，これらの中でも原材料と製造方法はパターンで抽出することはほぼ不可能であり，さらに文脈を考慮する必要もあるため抽出が難しい属性である．

化合物の構造化データベースとしては既に様々なデータベースが存在し，PubChem[2]，ChEBI[3] などが有名である．しかし，これらのデータベースには化合物の原材料や製造方法に関する情報が含まれていない，もしくは少ないのが現状である．例として PubChem には Hazardous Substances Data Bank[4] から化合物の製造方法がリンクされているが，このデータベースが対象としているのは有害化学物質のみであり，カバー率に問題がある．また，これらのデータベースの内容は英語で記述されており，日本語における化合物の原材料・製造方法の構造化はより進んでいないというのが現状である．

この課題を解決するため，本研究では日本語版 Wikipedia の化合物記事を対象とした化合物の原材料および製造方法の抽出を行い，構造化データを作成することを目的とする．

原材料および製造方法を抽出するにあたり，固有表現抽出において高い抽出性能を示している深層学習モデル [5] を適用することを考える．しかし、化合物情報の抽出に深層学習モデルを適用した場合，以下の理由により分散表現がうまく獲得できず抽出精度が低下する恐れがある．

- 化合物の種類が極めて多い
- トレーニングデータ中での出現頻度が非常に少ない，もしくは出現していない化合物名が大量に存在する

また，原材料や製造方法の抽出にあたっては文脈等の文の特徴が重要であるため，特定の化合物名の影響を受けてしまうことは避けなければならない．

そこで，本研究では外部知識源を使用して化合物名を抽象化することでこの問題に対処する．

2 関連研究

ある種類の単語を抽象化することにより機械学習モデルの性能を向上させた研究として，関ら [6] の統計翻訳モデルが挙げられる．関らは，未知語となる可能性の高い単語のカテゴリとして地名に注目した．まず，外部知識源を用いて地名対訳辞書を作成し，次に作成した辞書を用いてコーパス中の地名を”PLACE”に置換，抽象化する手法を提案している．この手法により，コーパスに含まれていない未知の地名であっても翻訳が可能となり，さらに文中の適切な位置に地名が配置されるようになったことで翻訳モデルの精度が向上したことが示されている．

化合物名についても，種類数が膨大かつ新たな化合物も日々大量に生まれていることから，未知語となる可能性が高い単語カテゴリといえる．外部知識源から化合物名辞書を作成し，トレーニングデータ内に存在する化合物名を置換・抽象化することは，抽出モデルの性能向上にも有効に働くことが期待できる．

3 提案手法

本手法のベースモデルとして，固有表現抽出タスクにおいて高い精度を示している Bi-LSTM+CRF モデル [5] を用いる．しかし，このモデルに化合物名を含む文をそのまま入力として与えた場合，出現頻度が非常に低い化合物名から得られた低品質な分散表現や，特定の化合物名の分散表現の影響により，原材料や製造方法が出現しやすい文の特徴を獲得することが難しくなってしまう．また，トレーニングデータの中に含

まれている化合物名の種類が少なければ，実際に抽出を行う際に化合物名の多くが未知語となり，さらに抽出精度が低下することが懸念される．そこで，外部知識源から収集・構築した化合物名辞書で化合物名を置換し抽象化した文を，Bi-LSTM+CRF モデルの単語側の入力に与える．これにより，原材料や製造方法が出現しやすい文の特徴を学習できるようになると考えられる．また，トレーニングデータに含まれていない化合物名であっても，辞書内にある化合物名であれば置換によって「化合物名」ということが明示されるため，未知語の数が減り抽出精度が向上することが期待できる．

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．例えば，硝酸の Wikipedia 記事には以下のような文が存在する．

1. 二酸化窒素を水（温水）と反応させると硝酸と一酸化窒素が発生する
2. 硝酸は硫酸中では塩基として挙動しプロトン化を受け、脱水によりニトロイルイオン (nitroyl / NO2+) を生成する。

1 の文は硝酸の製造方法であり，原材料も含まれている．一方，2 の文はニトロイルイオンの生成に関する文章であり，抽出対象となる硝酸の原材料や製造方法は含まれていない．そこで，このような文を区別できるようにするために，タイトル化合物名およびその同義語を”[title-compound]”，その他化合物名を”[compound]”と置換する．

本手法により，記事タイトルとなっている化合物名の原材料および製造方法の抽出精度を高めることができると考えられる．

3.1 Bi-LSTM+CRF

抽出モデルには Lample ら [5] の Bi-LSTM+CRF モデルを用いる．モデルには単語系列と各単語を構成する文字系列を入力として与える．単語側の入力には化合物名が置換され分かち書きされた文を，文字側の入力には置換前の化合物名を含む文字系列を与える．これにより，抽象化を行いつつ化合物名らしい表現を獲得することができ，未知の化合物名にも頑強になると考えられる．

3.2 化合物名辞書の作成・置換

記事中の化合物名を置換する際に用いる化合物名辞書は，WikiData[7]，PubChem，ChEBI から収集・構築する．まず，Wikipedia の各化合物記事にリンクされている WikiData の項目から，化合物名の日本語・英語の別称，PubChem CID，ChEBI ID を取得する．次に，取得した PubChem CID と ChEBI ID を元に，各データベースから同義語情報を取得する．そして，取得した化合物名の同義語情報を用いて，記事タイトルの化合物名およびその同義語を”[title-compound]”に置換する．同義語情報も活用することで，略称や通称など記事タイトルとは異なる化合物名が使用されている場合にも対処することが可能となる．さらに，タイトル化合物以外の化合物名を置換する際には，上記のデータベースの他に日化辞辞書 [8] を用いる．タイトル化合物以外の化合物名は”[compound]”と置換する．
使用した化合物データベースの規模を表 1 に示す．

表 1: 収集した化合物データベースの規模

リソース名	化合物名の数
WikiData	35,010
ChEBI	24,782
PubChem	18,274
日化辞	82,922

4 実験

本研究では，原材料と製造方法それぞれの属性ごとにモデルを作成し，実験と評価を行った．

4.1 データ

モデルの学習と評価には，森羅プロジェクト [9] で公開されている構造化データおよび対応する Wikipedia 化合物記事を使用した．前処理として，対象となる Wikipedia 記事の本文を取得し，文単位に分割した．次に，化合物名辞書を適用した MeCab で形態素解析したのち，構造化データを元に IOB タグを付与した．また，先述した化合物名辞書を用いて化合物名を置換した単語系列も作成した．実験では，構造化データに対応する記事 598 件のうち，500 件をトレーニングデータ，98 件をテストデータとした．

表 2: 抽出結果

属性	評価方法	手法	TP	FP	FN	適合率	再現率	F 値
原材料	完全一致	置換無し	63	251	97	0.2006	0.3938	0.2658
		置換有り	89	207	71	0.3008	0.5563	0.3904
製造方法	完全一致	置換無し	3	52	91	0.0545	0.0319	0.0403
		置換有り	11	79	83	0.1222	0.1170	0.1196
	部分一致	置換無し	35	32	74	0.5224	0.3211	0.3977
		置換有り	72	61	51	0.5414	0.5854	0.5625

4.2 Bi-LSTM+CRF のパラメータ

単語の分散表現と文字の分散表現の次元数はそれぞれ 100 と 25，単語レベルおよび文字レベルの Bi-LSTM のユニット数はそれぞれ 100 と 25，CRF 層へ通じる全結合層の次元数は 100 とした．また，単語レベルの Bi-LSTM 層へのドロップアウト率は 0.5 に設定し，最適化アルゴリズムには Adam[10] を用いた．

5 結果

評価指標として，完全一致での適合率，再現率，F 値を用いた．また，抽出フレーズが長くなる傾向にある製造方法については，句点や末尾（「〜によって得られる。」など）の有無などの僅かな抽出ミスが多数確認されたため，部分一致も正答とした評価も行った．

結果を表 2 に示す．化合物名の置換ありと無しの結果を比較すると，原材料と製造方法どちらの属性についても適合率・再現率ともに性能が向上しており，F 値では原材料で約 13 ポイント増加，製造方法の部分一致評価では約 17 ポイント増加と 10 ポイント以上精度が向上していた．製造方法の部分一致での評価結果を見ると，適合率は 2 ポイントほどの増加に留まっているものの再現率が約 26 ポイント増加していた．原材料の結果を見ると，正しく抽出された件数が増加したとともに誤抽出の件数が減ったことで，再現率だけでなく適合率の方も 10 ポイント以上向上していた．

6 考察

6.1 有効性の検証

原材料の抽出結果から，化合物名を置換することで誤抽出されなくなったデータを分析した．誤抽出されなくなったデータは 162 件あり，原材料ではない化合物名や未知語，途切れて抽出された化合物名がほとんどであった．以下に，誤抽出されなくなったデータと抽出元の文の一例を示す．

ケリダム酸

ケリダム酸（英語、Chelidamic acid）とは、4-ヒドロキシピリジン-2,6-ジカルボン酸（英語、4-hydroxypyridine-2,6-dicarboxylic acid）のことである。

サッカリン

1884 年にファールバーグがサッカリンと名づけ、レムセンに無断で数か国で製造法に関する特許を取得した。

※赤文字は置換前に抽出されていた箇所

置換前にケリダム酸の記事から誤抽出されていたのは「4-ヒドロキシピリジン」「4-hydroxypyridine-2,6」であり，どちらもケリダム酸の原材料ではない上に，化合物名が途中で切れてしまっている．また，サッカリンの記事から誤抽出されていたのは「ファールバーグ」という人名であり，これはトレーニングデータには含まれていない未知語であった．一方で，置換後に正しく抽出されるようになったデータは 37 件あり，これらは全て化合物名であった．原材料でない化合物名の誤抽出が減少していることから，置換によって化合物の生成に関する文の特徴を獲得できたと考えられる．また，化合物名でない未知語が抽出されなくなったのは，未知語を抽出するよりも置換された化合物名を抽出する方が原材料である可能性が高いと抽出モデルが学習したためだと考えられる．化合物名を中心に抽出するようになったことについては，正しく抽出されるようになった原材料の全てが化合物名であることからも示唆される．

さらに，辞書に含まれていなかった長い化合物名の原材料も正しく抽出されるようになったことが確認できた．例えば，置換前には「3-（トリメチルシリロキシ」と抽出していたものが，置換後には「3-（トリメチルシリロキシ）-1-（トリブチルスタンニル）プロピン」と正しく抽出されるようになった．このような結果が得られた要因として，Bi-LSTM+CRF モデルの文字側の入力には元の化合物名を与えて学習したことで，数値や記号などが含まれる化合物名特有の表現をうまく獲得できたことが考えられる．

次に，化合物名を置換することで 1 つも属性値が抽出されなくなった文を取得し，これらの文の傾向を目視で確認した．原材料の抽出モデルでは，属性値が抽出されなくなった文が 125 文存在し，うち 105 文は化合物の定義や用途など，化合物の生成に関係しない文が 105 文あった．製造方法の抽出モデルでは，属性値が抽出されなくなった文が 29 文あり，うち 16 文が化合物の生成に関係しない文だった．この結果から，置換を行うことで化合物の生成とは関係のない文からの誤抽出が減少したことが分かった．これは，化合物名がタイトル化合物か他化合物かの 2 種類に抽象化されたことで各化合物名の分散表現からのノイズの影響が小さくなり，化合物の生成に関わる文の特徴を学習することが可能になったためだと推測される．

6.2 エラー分析

提案手法の改善の方向性を考えるため，化合物名の置換後も誤抽出されたデータについて分析した．

置換後の誤抽出データを確認すると，タイトル化合物の生成に言及していることを明示するためにタイトル化合物名と他化合物名を区別して置換を行ったのにも関わらず，タイトル化合物ではない化合物の生成に関わる文からの誤抽出が目立った．ここで，トレーニングデータ（500 記事，7435 文）の記事を確認すると，原材料が存在する 1106 文のうちタイトル化合物が存在しない文が 662 文あった．また，製造方法が存在する 520 文のうちタイトル化合物が存在しない文が 322 文存在した．これらの文を見ると，いわゆる主語抜き文や複数文にわたって生成プロセスが記述されているような文章が多く，タイトル化合物と他化合物どちらに言及している文なのかを見分けることは困難なデータであることが分かった．

7 おわりに

本研究では，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が確認できた．今後の課題として，他化合物の生成に関する文からの誤抽出を減らすために，前後の文から言及されている化合物を推定することが挙げられる．

参考文献

[1] 関根の拡張固有表現階層 -7.1.1-. <https://sites.google.com/site/extendednamedentity711>.

[2] The pubchem project. <https://pubchem.ncbi.nlm.nih.gov/>.

[3] Chemical entities of biological interest. <https://www.ebi.ac.uk/chebi/>.

[4] Hazardous substances data bank. <https://toxnet.nlm.nih.gov/newtoxnet/hsdb.htm>.

[5] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270. Association for Computational Linguistics, 2016.

[6] 関拓也, 山本和英. 統計的機械翻訳における地名の汎化の影響. 言語処理学会第 15 回年次大会, pp. 200–203, 2009.

[7] DennyVrande č i ć , MarkusKr ö tzsch. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, Vol. 57, pp. 78–85, 2014.

[8] 日化辞辞書. <https://dbarchive.biosciencedbc.jp/jp/mecab/data-3.html>.

[9] 関根聡, 小林暁雄, 安藤まや. Wikipedia 構造化プロジェクト「森羅 2018」. 言語処理学会第 25 回年次大会, 2019.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2014.