The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

文書からの化合物名抽出のためのサブワード有効性調査

Using Subword Sequence BiLSTM-CRF Model for Compound Name Extraction

浦澤合 *1　関根裕人 *1　乾孝司 *1　岩倉友哉 *2
Go Urasawa　Hiroto Sekine　Takashi Inui　Tomoya Iwakura

*1筑波大学大学院　/理研 AIP-富士通連携センター
University of Tsukuba/RIKEN AIP-FUJITSU Collaboration Center

*2富士通研究所/理研 AIP-富士通連携センター
Fujitsu Laboratories/RIKEN AIP-FUJITSU Collaboration Center

In this paper, we investigate of using subword sequences for compound name extraction problem. Five variety of subword sequence generators (SYMBOL, SP, BPE, BPE-DICT, and BPE-PMI) were used in the investigation. Last two of these, BPE-DICT and BPE-PMI, are originally proposed in this work. BPE-DICT is a variation of BPE which has a dictionary-based restriction. BPE-PMI introduces the PMI measure instead of word frequency count. The experimental results showed that subword sequence information improved the extraction performance. The F-measure value of BPE-DICT is 86.74 which is best score in all conditions of our experiments.

1. はじめに

人間の言葉で書かれた文書から自動的に構造化データを抽出するタスクを「情報抽出」という．抽出するデータの種類に応じてデータの特徴，文書形式などが異なる．したがって抽出データの分野により情報抽出をおこなうのに最適な抽出手法や考え方が異なるので抽出データに合わせた手法を考える必要がある．本研究では化学化合物に関する情報抽出について考える．

化学分野の研究ではさまざまな場面で化合物データベースが利用され，日本化学物質辞書 [5] や PubChem[4]，ChEMBL[1] など各種データベースが提供されている．現在，これらの化合物データベースは論文や特許を人手で読み解くことで作成されている．しかし化学化合物に関する論文，特許は数え切れないほどの数が存在するのでそれらから必要なデータを人手で抽出することは非常にコストの高い作業である．そこで解析技術と組み合わせたデータベース作成支援が求められている．

化合物抽出用のソフトウェアが存在するが，精度の点で実用レベルに至っているとは言い難い．化学化合物の抽出が実用レベルでないのは，化学化合物の命名規則である IUPAC 命名規則 [2] があるにも関わらず化合物の多様な表記方法があることが理由として挙げられ，略称，通称，化学式など多くの表記を持つ．例として化合物「フェニルアラニン」は「Phe-OH」,「L-Phe-OH」,「L-Phenylalanine」,「(S)-2-Amino-3-phenylpropionic acid」などで表現されるが，これで全ての表記ではない．また化学の分野において新しい化合物が頻繁に報告されることも理由として挙げられる．他には複合語の存在である．複合語とはある化合物を部分的に含む化合物のことである，これらにより化合物の正確な抽出が困難となる．

上記で述べたように，化学化合物は未知語が発生しやすい分野であるため，未知語に対応する処理が重要となる．サブワードは単語と文字両者の中間的な特性を持つため単語情報を保持しつつ，未知語に対応できると考えられる．そのため本研究は様々な分割方法でサブワードを獲得し，どのようなサブワード系列が化学化合物抽出において効果が見られるか検討する．

連絡先: 浦澤合，筑波大学大学院システム情報工学研究科，
g.u@mibel.ca.tsukuba.ac.jp

本稿の構成として，第 2 章で関連研究について述べる．第 3 章ではサブワード獲得方法について述べる．第 4 章では評価実験について述べる．最後に，第 5 章では予備調査を踏まえた考察と今後の指針を述べる．

2. 関連研究

化学分野の論文や特許から化合物を抽出する研究 [8][9][6] は盛んに行われている．機械学習を利用した手法 [8] が高精度の結果を残すものとして以前から知られていたが，近年ではニューラルネットワークを利用した手法 [9] が多く提案され化合物抽出において素晴らしい結果をもたらしている．

機械学習を利用した Lu ら [8] は文字や単語から獲得できる情報を CRF の素性として利用した．また単語のクラスタリングを素性として利用することで精度や再現性を高めようとした．次に Ling ら [9] は化合物抽出のニューラルネットワークを利用した手法によく見られる BiLSTM に attention 機能を追加した手法を提案し，CHEMDNER task[6] では現在もっとも精度の高い手法となっている．以上のように単語および文字の情報を利用した手法は数多く提案されているが，本研究のようにサブワードを利用した手法は提案されていない．

3. サブワード

3.1 固有表現抽出におけるサブワード

化合物名抽出を行う際には系列ラベリング問題として定式化することが一般的である．系列ラベリング問題として考えた場合に単語単位の系列を仮定すると，未知語が発生しやすくなり，また，抽出したい化合物と処理上の単位である単語との間で境界が一致しない問題を引き起こす．この問題への対策として，文字単位系列を用いることが考えられるが，この場合，系列長が長くなり計算量が増加する．また，単語がもっていた意味情報を利用することができないといった新たな問題が発生する．単語の使用と文字の使用はそれぞれに利点と欠点があり，両者はトレードオフの関係にあると言える．サブワードはこの両者の中間的な特性をもっており，サブワードを考慮した系列を仮定することで，単語と文字の両者の欠点を補うことができ

Algorithm 1 BPE

1: $DICT \Leftarrow$ 辞書データ
2: $VOCAB \Leftarrow$ 語彙 (初期は空)
3: テキストを文字に分割する
4: **while** $VOCAB$ が指定語彙数に達するまで **do**
5: 隣り合う全ての文字トークンのペアに対して，それらを結合して新たな語彙候補を作成する．($VOCAB$ に存在するものは一文字として扱う)
6: 語彙候補の中で，出現頻度の最も高い候補を $VOCAB$ に追加する．
7: **end while**

Algorithm 2 BPE-DICT

1: $DICT \Leftarrow$ 辞書データ
2: $VOCAB \Leftarrow$ 語彙 (初期は空)
3: テキストを文字に分割する
4: **while** $VOCAB$ が指定語彙数に達するまで **do**
5: 隣り合う全ての文字トークンのペアに対して，それらを結合して新たな語彙候補を作成する．($VOCAB$ に存在するものは一文字として扱う)
6: 結合前の左右の要素がどちらも $DICT$ に登録されていない語彙候補の中で，出現頻度の最も高い候補を $VOCAB$ に追加する．
7: **end while**

ると考えられる．サブワードとは，ある単語の部分文字列のことである．例として「magnesium」という単語の場合,「ma」,「mag」,「magn」,「si」,「sium」などがサブワードになる．

3.2 サブワード獲得方法

本研究では単語から得られるサブワードとして以下 5 種類のサブワードを試みた．このうち，BPE-DICT と BPE-PMI は本研究で提案するサブワード獲得方法である．

- 記号などが存在する際にその記号で単語を分割するもの (SYMBOL),
- SentencePiece(SP),
- Byte Pair Encoding(BPE),
- 辞書制約付き Byte Pair Encoding(BPE-DICT),
- PMI による Byte Pair Encoding(BPE-PMI).

3.2.1 SYMBOL

SYMBOL は単語中に記号などが存在する際に，その記号で単語を分割するサブワード分割方法である．一般的な単語は記号を単語中に含まないので，単語は分割されず，単語そのままであることが多い．単語が組み合わせられた複合語や，長い単語には記号が含まれることが多く，分割される．

3.2.2 SP

SentencePiece[3][7] を実行することでサブワード系列を獲得する．

3.2.3 BPE

Byte Pair Encoding は Sennrich ら [10] が提案したサブワード分割方法である．BPE は原文をすべて文字に分割し，1 文字 1 語彙から始まる．隣り合う文字のペアに対して，それらを連結して新たな語彙の候補とする．この際すでに語彙に含まれているものは 1 文字として扱う．連結した際に最も出現頻度が高くなるサブワードを選び語彙に追加する．この手続きを決められた語彙数に達するまで繰り返し語彙結合ルールを学習することでサブワード分割を行う．

3.2.4 BPE-DICT

BPE-DICT は辞書制約付きの BPE である．基本的な手続きは通常の BPE と同じであるが，連結する前の左右の要素どちらも化学化合物辞書に存在しない語彙候補の中で出現頻度が最も高いものを語彙に追加する．また今回，化学化合物辞書として利用したのは PubChem データベース [4] で約 3 億個の化合物を含んでいる．

3.2.5 BPE-PMI

通常の BPE は出現頻度が高いサブワードを新たな語彙として追加するが，BPE-PMI は出現頻度ではなく Pointwise Mutual Information(PMI) が高いサブワードを新たに語彙に追加する．学習データを参照することで各サブワードについて，化合物の構成要素 (クラス 1)，構成要素でない (クラス 0) を割り当て，各サブワードとクラス 1 間の PMI を求め，出現頻度に置き換えて BPE を行う．式 (1) に PMI の定義式を示す．ここで，$P(SW)$ はあるサブワードが出現する確率，$P(C=1)$ はある要素が化合物の構成要素である確率，$P(SW, C=1)$ はあるサブワードが出現した際にそれが化合物の構成要素である確率である．

$$PMI(SW, C=1) = \log 2 \frac{P(SW, C=1)}{P(SW)P(C=1)} \quad (1)$$

4. 評価実験

4.1 実験設定

前節で述べた手法によって得たそれぞれのサブワード系列が化学化合物抽出にどの程度有効であるか観察した．データセットは CHEMDNER tsak における学習データ 3,500, 開発データ 3,500, 評価データ 3,000 を利用した．これは PubMed から化合物について書かれた論文の abstract を 10,000 件集め，人手でアノテーションをされたものである．合計で 84,355 の化合物エンティティが存在し，それらの重複を省くと 19,806 となる．化学化合物の固有表現抽出モデルとして Bidirectional LSTM-CRF[9] を使用し，これは単語と文字の LSTM を持つ．本研究では固有表現抽出モデルの単語 LSTM をサブワードに

表 1: サブワード別出力例：学習データ内に存在する化合物

method ＼化学化合物	docosahexaenoic acid	nitric oxide	glutathione	superoxide
SYMBOL	docosahexaenoic acid	nitric oxide	glutathione	superoxide
SP:chem6k	docosahexaeno ic acid	nitr ic oxide	glutathione	super oxide
BPE:32k	docosahexaenoic acid	nitric oxide	glutathione	superoxide
BPE-DICT:32k	docosahexaenoic acid	nitric oxide	glutathione	su p eroxide
BPE-PMI:32k	docosahexaenoic acid	nitric oxide	glutathione	superoxide

表 2: サブワード別出力例：学習データ内に存在しない化合物

method ＼化学化合物	isocorilagin	diasesartemin	tetrahydropalmatine	ritanserin	polyphosphoinositides
SYMBOL	isocorilagin	diasesartemin	tetrahydropalmatine	ritanserin	polyphosphoinositides
SP:chem6k	isoc or il ag in	di a s es ar te m in	tetrahydro p al mat ine	rit an serin	poly phosphoinositide s
BPE:32k	is ocor il agin	di as es artem in	tetrahydro palmatine	rit anserin	poly phospho inosi tides
BPE-DICT:32k	isoc or il agin	di as es artem in	tetrahydro pal matine	rit an serin	p oly p hos p hoinosi tides
BPE-PMI:32k	isoc ori lag in	di as esar te min	tetrahydro palmatine	rit ans erin	polyphospho inositi des

表 3: 利用したモデルのパラメータ

epoch	200
batch size	100
単語分散表現	50
文字分散表現	30
単語 LSTM の隠れ層	100
文字 LSTM の隠れ層	50
initial rate	0.015
dropout	0.5

置き換えて用いる．また今回使用したモデルのパラメータを表 3 に示す．単語分散表現，文字分散表現はそれぞれ 50 次元，30 次元とし，LSTM の隠れ層では 100 次元，50 次元とした．表中の「単語」は実際にはサブワードである．

サブワード獲得方法別の設定を述べる．SentencePiece の学習には上記と同じ学習データを利用した．学習データの全テキストを語彙数 32,000，ユニグラムで学習させたもの (SP:32k) と学習データのタグづけされた化学化合物部分のみを語彙数 6,000，ユニグラムで学習させたもの (SP:chem6k) がある．次に BPE,BPE-DICT,BPE-PMI について説明する．これら 3 つも同様に先と同じ学習データを利用したが，BPE,BPE-DICT は学習データのテキスト部分のみを語彙獲得に利用し，BPE-PMI は学習データのテキスト部分とタグ部分を語彙獲得に利用した．また 3 つそれぞれに対して 8,000，16,000，32,000 の語彙数で学習させた．また BPE-DICT では辞書引きを行う手続きに仮候補の文字列が 3 文字以上という制限を加えたもの (BPE-DICT-char3) と制限なしのもの (BPE-DICT) がある．この制限は制限なし BPE-DICT が獲得したサブワードを観察した際に，多くの 1 文字サブワードが残った．その結果に対して調整を行う目的で取り入れた．

4.2 実験結果と考察

表 4 に実験結果を示す．性能を F-measure の値で比較すると，最も良い性能であるのは語彙数 32,000 で制限なしの BPE-DICT である．これの F-measure は 86.74 であり，ベースラインである単語区切りの 86.32 に 0.4 上回っている．これは通常の BPE と比較しても性能が良いことから辞書制約付きがある BPE は化合物抽出において効果があると言える．

またシンプルなサブワード分割方法である SYMBOL もベースラインの F-measure を上回っており，これはサブワードが単語よりも化合物抽出で良い影響を持っていると言える．BPE-PMI の F-measure は通常の BPE と比べて低い．したがって各サブワードと化合物の構成要素クラス間との PMI をもとに BPE を行うよりも出現頻度をもとにした通常の BPE の方が今回の実験では良いサブワード分割ができていると言える．

表 4: サブワード別実験結果

method	Precision	Recall	F-measure
単語	86.62	86.03	86.32
SYMBOL	85.10	88.01	86.53
SP:32k	87.56	85.39	86.46
SP:chem6k	78.90	69.53	73.95
BPE:8k	87.12	84.98	86.04
BPE:16k	87.47	85.99	86.72
BPE:32k	84.38	79.38	81.80
BPE-DICT:8k	87.74	84.32	86.00
BPE-DICT:16k	87.37	85.70	86.53
BPE-DICT:32k	87.14	86.33	86.74
BPE-DICT-char3:8k	87.17	84.33	85.73
BPE-DICT-char3:16k	86.40	86.39	86.39
BPE-DICT-char3:32k	88.10	85.10	86.62
BPE-PMI:8k	85.51	82.45	83.95
BPE-PMI:16k	86.86	82.76	84.76
BPE-PMI:32k	78.69	71.03	74.66

SP:chem6k と BPE-PMI:32k の性能が他と比較すると極端に低い．SP:chem6k と BPE-PMI:32k ともにサブワード分割の特徴として化学化合物が分割されることが少なく，残りやすいことが挙げられる．これは学習をともに学習データの化学化合物部分に集中して行うからと考えられる．化学化合物とそれ以外の違いを他のサブワード分割方法では区別できているとも言える．

表 1 と表 2 に各サブワード分割方法で得られた化学化合物のサブワード出力例をいくつか示す．表 1 の化学化合物は学習データに化合物部分としてタグ付けされているのでほとんどのサブワード分割方法が分割を行わず単語を維持している．反対に，表 2 の化学化合物は学習データに化合物部分としてタグ付けされていないため多くのサブワード分割方法で分割が行われる．また単語や記号でサブワード分割を行うシンプルな分割方法である SYMBOL では表 2 の化合物すべてが抽出するこ

とができなかった．しかし実験結果で最も F-measure が高いサブワード分割である BPE-DICT:32k では表 2 の化合物を抽出することができた．したがって，学習データに化合物部分としてタグづけされていない化合物でもサブワードを利用することで抽出することを可能とし，DICT-BPE:32k は良いサブワード分割を行なっていると言える．

5. おわりに

今回，様々なサブワードを Bidirectional LSTM の入力系列として利用し化学化合物抽出を行い，各サブワード獲得方法別の結果を観察し，サブワードが単語よりも化学化合物抽出において効果があることを示した．

結果としては BPE-DICT が最も高い性能を残したが，各サブワード獲得方法すべてに同様のパラメータを適用したので BPE-DICT が化学化合物抽出に最も適しているとは一概にも言えない．さらに今回語彙数として 8,000，16,000，32,000 を使い実験を行なった，それに対して実験結果や結果の傾向が異なるものとなったが，これはサブワードを利用する上で語彙数を決定することは化学号物抽出の性能に大きな影響があると言える．今後は各サブワード獲得方法に適したパラメータの調査やサブワード分割方法とそれに適した語彙数の決定について行う必要がある．また以上のことから得られる化学化合物抽出に適したサブワードをサブワードのみの入力系列ではなく，単語系列などと組み合わせた実験などを行なっていきたい．

参考文献

[1] ChEMBL. https://www.ebi.ac.uk/chembl/.

[2] Color books-iupac international union of pure and applied chemistry. https://iupac.org/what-we-do/books/color-books/.

[3] Github-google/sentencepiece: Unsupervised text tokenizer for neural network-based text generation. https://github.com/google/sentencepiece.

[4] The pubchem project. https://pubchem.ncbi.nlm.nih.gov/.

[5] 日本化学物質辞書 web—j-global 科学技術総合リンクセンター. https://jglobal.jst.go.jp/info/nikkaji.

[6] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. Chemdner: The drugs and chemical names extraction challenge. *Journal of cheminformatics*, Vol. 7, No. 1, p. S1, 2015.

[7] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[8] Yanan Lu, Donghong Ji, Xiaoyuan Yao, Xiaomei Wei, and Xiaohui Liang. Chemdner system with mixed conditional random fields and multi-scale word clustering. *Journal of cheminformatics*, Vol. 7, No. S1, p. S4, 2015.

[9] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, Vol. 34, No. 8, pp. 1381–1388, 2017.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.