化学ドメインにおける教師無し固有表現抽出

辰巳 守祐 [1]　　進藤 裕之 [1]　　松本 裕治 [1,2]
[1] 奈良先端科学技術大学院大学　先端科学技術研究科
[2] 理化学研究所　革新知能統合研究センター
{tatsumi.shusuke.tk0, shindo, matsu}@is.naist.jp

1　はじめに

化学物質名の固有表現抽出（Named Entity Recognition，以下 NER）では，(1) 新しい未知の化学物質が日々生み出されている，(2) 辞書やコーパス作成において専門分野の知識が必要，といった今まで幅広く研究されてきた人名や組織名に対する NER とは異なる特徴・課題がある．

問題 (1) への対処方法として，文脈情報を用いた分散表現を用いることで，辞書に記載されていない未知固有表現も NER 可能な手法が考えられてきた．分散表現を抽出する為の時系列データ処理単位として，文，単語，サブサード，文字が考えられる．そこで，本研究の Research Question1 を次のようにする．「Research Question1: 化学物質ドメインでの分散表現抽出において，どの処理単位が最も有効か」実験では分散表現抽出の処理単位を文字ベースとサブワードベースで比較し，文字ベースが有効であることが明らかになった．

問題 (2) への対処方法として，大規模な生データとある程度の規模の辞書がある前提で，Distant Supervision による擬似アノテーションコーパス作成手法が考えられてきた．ただ，Distant Supervision で得られる擬似アノテーションコーパスはノイズが生じやすい問題がある．そこで，本研究の Research Question2 を次のようにする．「Research Question2: Distant Supervision で生成された擬似コーパスからどのようにしてノイズを取り除くか」実験結果より，提案手法が固有表現の抽出性能向上に有効であることを示した．

本稿の貢献は以下の通り
・Research Question 1 に対して：分散表現に文脈情報を用いて，NER を行なった．化学物質名の抽出において，文字ベースの方がサブワードベースよりも性能が良いことが明らかになった．
・Research Question 2 に対して：提案手法がベースラインを上回った．特に，未知固有表現の割合が高い状況下での NER に有効であることが分かった．

表 1: 分散表現の比較

モデル	単語表現 抽出単位	時系列処理	単語表現 事前学習
ELMo	単語	BiLSTM	WordLM
BERT	単語	Transformer	MaskedLM, Next sentence prediction
Flair	文字	BiLSTM	CharLM

2　関連研究

2.1　文脈情報を用いた分散表現

文脈情報を用いた分散表現に関する研究として，BiLSTM による単語ベースの言語モデルを用いた ELMo[1] やこれを文字ベースで行った Flair[2] が知られている．また，Transformer による Masked 言語モデルと Next sentence prediction を用いた BERT[3] も注目されている．

化学物質名は文字数が長くなる傾向にあり，それらの接頭辞や接尾辞が，例えば“acid”のように特徴的な文字列であることが多い．これらの特徴が NER に有用であると仮説を立て，文脈表現を時系列文字ベースで扱う Flair を本研究で使用する．

Flair は 1 文全体の情報を BiLSTM により分散表現に反映させる．具体例として，図 2 では‘Washington’の分散表現抽出過程を示す．初めに，forwardLSTM により‘ Washington ’直後のスペースの隠れ層状態を取得する．この隠れ層状態は文頭から‘ Washington ’直後スペースまでの文脈情報を保持している．次に，backwardLSTM により‘ Washington ’直前のスペースの隠れ層状態を取得する．この隠れ層状態は文末から‘Washington’直前スペースまでの文脈情報を保持している．最後に，これらの 2 つの隠れ層状態を連結し，‘ Washington ’の分散表現とする．これらの抽出過程により，‘ Washington ’分散表現は 1 文全体の文脈情報を反映させた表現となる．

2.2　Distant Supervision

人手によるデータ作成を必要としない Distant Supervision として, 1 単語につき複数のラベル付けを認めることで，擬似アノテーションノイズに対応した

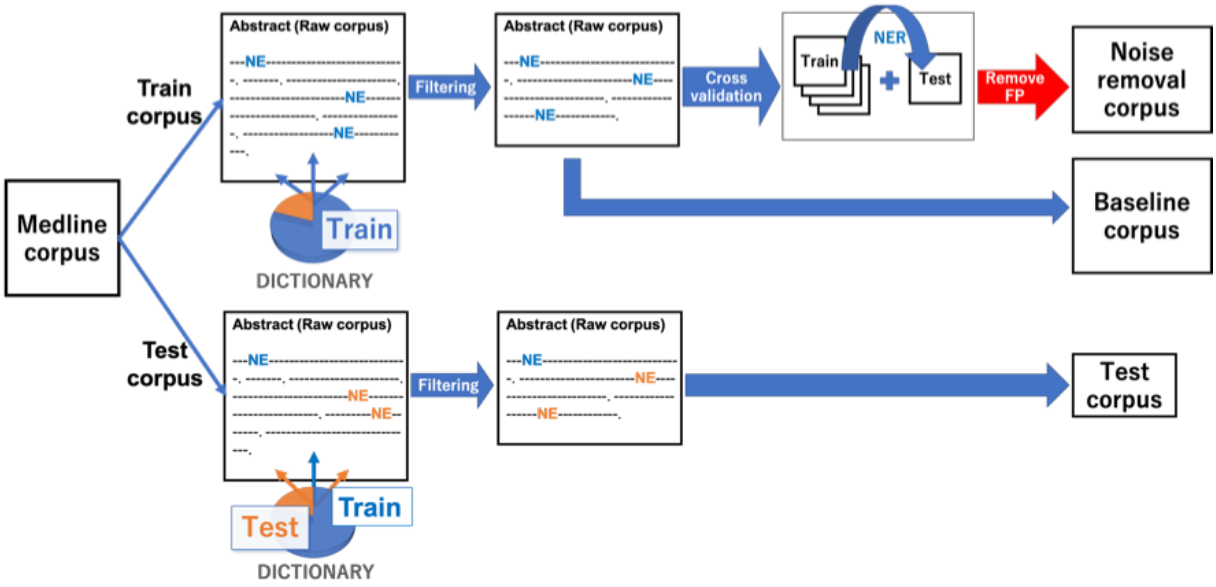

図 1: 擬似コーパスのノイズ除去過程．　初めに，論文 Abstract データセットを Train, Test に分割し，　それぞれカバレッジの異なる辞書による辞書マッチアノテーションを行う．これにより，Test データに未知固有表現を出現させることができる為，辞書記載，未記載両方の物質名抽出が求められる実タスクを再現できる．　そして，Train データからノイズ除去データを生成し，除去前のベースラインと比較する．

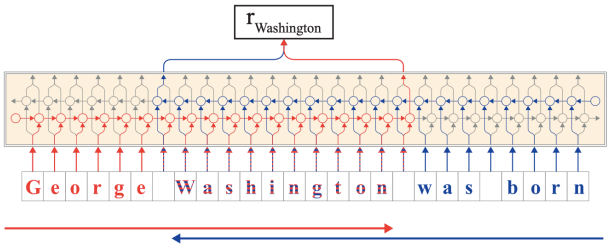

図 2: Flair の外観．‘ Washington ’の分散表現抽出例．BiLSTM により，文脈情報を用いて分散表現を抽出する．　赤色矢印が forwardLSTM，青色矢印が backwardLSTM．図は文献 [2] から引用．

Fuzzy-LSTM-CRF[4] や単語間のつながりでラベリングを行う AutoNER[4] がある．

3　提案手法

3.1　サブワード Flair

化学ドメインでの分散表現抽出において，どの処理単位が最も有効かを検証する．具体的には，文字ベースとサブワードベースの比較を行う．Flair は既存設定では文字ベースで処理を行う．しかし，化学物質名は長くなる傾向があり，更に接頭辞や接尾辞が特徴的な文字系列であることが多い．そこで，サブワードベースの系列処理の方が効果的であると仮説を立てた．ベースラインを Flair，提案手法をサブワード Flair として比較実験を行う．

表 2: Distant Supervision の比較

モデル	ラベル	ノイズへの対処方法
辞書マッチ	IOBES	無し
Fuzzy LSTM CRF	IOBES	曖昧な Token はマルチラベル 全ラベルパターン CRF
AutoNER	Tie Break	Token 間繋がりラベリング (step 1) Token スパン推定 (step 2) Entity タイプ推定
提案手法	IOBES	Dataset distillation 学習済み NER による False Positive 文の除去

サブワード Flair は前処理時にコーパスを SentencePiece で区切っておき，Flair 事前学習，NER 学習ではサブワードを 1 単位として学習するように Flair を修正する．評価は，文字，サブワード独立で Flair の事前学習を行い，それらの NER モデル F 値で行う．

3.2　擬似コーパスのノイズ除去

辞書マッチによる Distant Supervision の問題は辞書未記載の化学物質名（未知固有表現）を O とラベリングすることである．具体的には，実際は固有表現で B, I とラベリングされるべき化学物質が，辞書未記載な為に擬似アノテーションでは O とラベリングされる場合がある．このようなノイズに対処する為，上記事例をコーパスから除去する方法を提案する．具体的には，学習済み NER に推論させて False Positive（コーパスの正解ラベルが O，NER の予測ラベルが B, I の場合）となる単語が上記事例に該当する可能性が高いた

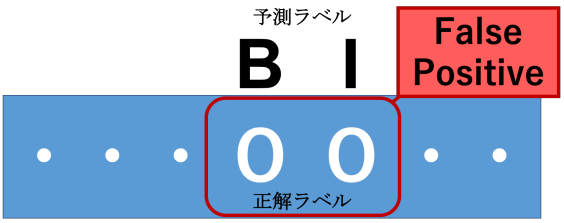

図 3: False Positive な単語例．NER モデルの予測は BI ラベル（固有表現）だが，正解は O ラベル (固有表現ではない)．このような単語はノイズになると考え，除去する．

め，これを除去する．補足説明として，False Negative（コーパスの正解ラベルが B,I，NER の予測ラベルが O の場合）な単語の除去は不要．なぜなら，擬似アノテーションコーパスの正解ラベルが常に正しいとは限らないが，正しいと仮定すれば False Negative な単語は固有表現であり，コーパスのアノテーションとして正しいからである．

3.2.1 ノイズ除去方法

まず，論文 Abstract データセットを Train, Test に分割し，それぞれ辞書マッチ方式によるアノテーションを行う．この時，Train と Test ではカバレッジの異なる辞書を使用する．(詳細は 4.2.1)

次に，Train データで cross validation を行う．その際，各回独立で学習済み NER モデルによる推論を行う．更に，各回の推論結果の内，False Positive な単語（正解は固有表現ではないが，モデル予測値が固有表現．図 3）を含む文は NER の学習においてはノイズになると考え，除去する．

ベースラインのデータをノイズ除去前データ，提案手法のデータをノイズ除去後データとし，それらを独立に用いて NER の学習を行う．

3.2.2 Test データの未知固有表現

問題 (1) で述べたように，化学ドメイン NER では辞書に記載されている物質名だけでなく，辞書未記載の物質名も抽出する必要がある．これに近い状況下で NER の評価をするため，Test データに辞書未記載の単語（未知固有表現）を出現させる．

4 実験

4.1 分散表現の抽出単位比較

本実験は，分散表現の事前学習とそれらを用いた NER の 2 部構成で，スパン F 値で NER を評価する．分散表現抽出の事前学習について，処理単位の比較対象である Flair（文字，又はサブワード）と，GloVe の 2 手法を用意し，NER ではこれらを連結させて使用する．

4.1.1 実験詳細

(1) 分散表現の事前学習

比較対象である文字 Flair とサブワード Flair の事前学習を行う．ベースラインである文字 Flair[1]の文字数は英数字記号などを含めた 275 文字．また，提案モデルは Flair の処理単位を文字ベースからサブワードベースに修正したものを使用．具体的には，コーパスをサブワードに区切り，文字ベースから前処理で区切られたサブワードベースで学習するように Flair を修正する．コーパス区切りには SentencePiece[2]を使用する．SentencePiece のサブワード辞書は，Flair 事前学習で使用するものと同じ学習データから SentencePiece で語彙数指定することで作成する．指定する語彙数は 1,000 サブワード，8,000 サブワードの 2 パターン．Flair の学習は，Loss の増減が見られなくなるか，Epoch 数約 2,000 程度まで行う．

また，補助的な分散表現として GloVe[5] を使用する．

(2) NER の学習

(1) で学習した Flair と GloVe の連結分散表現を用い，NER 学習を ChemdNER[6] で 150 epoch 行う．

4.1.2 実験設定

使用データについて，分散表現（Flair, GloVe）の事前学習データは Medline，NER の学習データは ChemdNER を使用．分散表現は Flair(文字，又はサブワード), GloVe（100 次元）．NER モデルは文献 [7] の BiLSTM + CRF.

4.1.3 結果と考察

表 3 に実験結果を示す．

Observation 1: F 値に関しては，文字ベースの方が高く，サブワードベースより優れた結果となった．

Observation 2: サブワードベースの方が NER 1 epoch 当たりの学習時間が短かった．これはサブワードベースで処理することで文字ベースの時よりも 1 単語中の Token 数（文字数，サブワード数）が減少した為と考えられる．

Observation 3: Flair 事前学習において，文字ベースでは 1,600 epoch 程度で学習収束が見られた．一方，サブワードベースについては，約 2,000 epoch では学習収束に至らなかった．このことから，サブワードベースの方が Flair 事前学習により多くの Epoch を要することが分かった．これは，文字からサブワードへの移行により，ソフトマックス層の次元数が増加した為と思われる．

[1] https://github.com/zalandoresearch/flair
[2] https://github.com/google/sentencepiece

表 3: 分散表現の抽出単位比較

Flair	Token	LM PPL	Epoch	Time min/epoch	F 値
Sub1k	1,000	4.63	2,200	4.53	72.21
Sub8k	8,000	6.88	2,000	4.07	73.81
Char	275	2.06	1,600	8.78	83.64

4.2 擬似コーパスのノイズ除去

4.2.1 実験詳細

(1) 辞書マッチアノテーション

論文アブストラクト（Medline）と化学物質名辞書（CTD，MeSH）の辞書マッチアノテーションを行う．3.2.2 で述べたように，現実に近い状況下での評価の為，Test データに未知固有表現を出現させる必要がある．そこで，辞書単語を 8 対 2 に分割し，8 割を用いて Train データのアノテーションを行う．残りの 2 割は，Test データにのみ使用する．Test データは残りの 2 割のみを使用してアノテーションしたもの（Test 1）と，辞書単語全体を使用したもの（Test 2）の 2 データセット用意する．Test 1 は Test データ中に出現する固有表現全てが未知固有表現であり，Test 2 は 2 割が未知固有表現となる．

また，上記の実験設定に応じて，評価は Recall で行う．これは，辞書マッチで全ての物質名を固有表現とラベリングするのは不可能であること．更には，Test 1 ではアノテーションに辞書の一部しか用いず，Test データにおいて固有表現ではなくても実際は固有表現である単語の出現が考えられる為である．

最後に，Train, Test 双方において，固有表現以外の単語割合が高くなり過ぎるのを防ぐ為，固有表現とラベリングされた単語が出現しない文を除外する．

(2) Train データによるノイズ除去

Train のみで Cross validation を行う．学習は 5 epoch．各学習済み NER モデルが誤って Positive と推論した単語は，本実験ではノイズと考え，このような単語を含む文を除去して提案 Train コーパスを作成する．

(3) ノイズ除去コーパスによる NER の学習

ベースラインを False Positive 除去前，提案コーパスを除去後のコーパスとする．ベースラインと提案コーパスそれぞれで NER モデルの学習を 15 epoch 行う．

4.2.2 実験設定

使用データは，分散表現の事前学習データ，NER の学習データ共に Medline，化学物質辞書には CTD，MeSH を使用する．なお，本実験では NER の学習においては計算機のメモリ制約から Medline 全体の 5 % 相当である 15 万文を用いる．辞書の規模は，CTD 183,481 単語，MeSH 238,232 単語，合計で約 40 万単語を使用．使用する分散表現は Flair（文字）と GloVe（100 次元）を連結したもの．NER モデルは実験 4.1 と同様．

表 4: Baseline と提案手法比較

モデル	Test 1			Test 2		
	Pre	Rec	F1	Pre	Rec	F1
Baseline	30.43	27.13	28.69	94.15	79.54	86.23
提案手法	34.30	32.90	33.58	92.52	80.70	86.21

4.2.3 結果と考察

表 4 に実験結果を示す．

Observation 1: Test1, 2 共に，提案手法がベースラインを Recall で上回った．

Observation 2: 特に，未知語固有表現の割合が高い Test1 において優れており，ベースラインより 5 ポイント高い Recall 結果が得られた．

5 おわりに

本稿では，化学ドメインにおける教師無し NER を分散表現の抽出単位比較，擬似コーパスのノイズ除去の 2 点で検討した．Research Question 1 に対して実験 4.1 を行ったところ，文字ベースの方がサブワードベースよりも分散表現事前学習の収束が早く，NER の F 値も高くなった．これにより，化学物質名の抽出において，文字の方が処理単位として優れていることが分かった．Research Question 2 に対しては実験 4.2 を行い，提案手法による，NER の Recall 向上を示した．特に，未知固有表現の割合が高い状況下での向上が大きかった．これにより，未知の化学物質名 NER に提案手法が有効だと分かった．今後の課題として，Distant Supervision の大規模データ実装や，更に性能の高い False Positive 除去方法の検討などが挙げられる．

参考文献

[1] Peters Matthew E., Neumann Mark, Iyyer Mohit, Gardner Matt, Clark Christopher, Lee Kenton, and Zettlemoyer Luke. Deep contextualized word representations. In *NAACL*, 2018.

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Shang Jingbo, Liu Liyuan, Ren Xiang, Gu Xiaotao, Ren Teng, and Han Jiawei. Learning named entity tagger using domain-specific dictionary. In *EMNLP*, pp. 2054–2064, 2018.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[6] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, Vol. 7, No. 1, p. S2, 2015.

[7] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.