遠距離教師あり固有表現抽出における辞書マッチの誤りの考慮

小林 滉河 † 若林 啓 ‡

† 筑波大学大学院図書館情報メディア研究科 ‡ 筑波大学図書館情報メディア系
†s1511506@klis.tsukuba.ac.jp ‡kwakaba@slis.tsukuba.ac.jp

1 はじめに

固有表現認識 (Named Entity Recognition; NER) とは，文章中から人名・地名・組織名といった固有表現や時間表現などの語句を抽出する自然言語処理技術のことである．近年，NER は深層学習を用いることで，人手による素性選択を必要とせずに高い抽出性能を持つモデルが数多く提案されている [4, 5]．しかし，これらの NER モデルの多くは人手によってラベル付けが行われた大量のアノテーションコーパスを教師データとして必要とする．そのため，化学や医学といった専門的な知識が必要となる分野において，教師データを作成する際のアノテーションコストは高く，教師データを十分に確保することが難しいという問題がある．

アノテーションコーパス以外に使える資源は二つあり，一つ目は辞書である．化学や医学といった専門分野では，専門用語を収集している辞書が入手可能である場合が多い．辞書を利用した最も簡単な固有表現抽出手法として，文字列マッチングにより対象となるテキスト中の固有表現を抽出する方法がある．この方法では，テキスト中に辞書に含まれる固有表現と一致する文字列を見つけたとき，それを固有表現としてみなし，辞書中のどの固有表現にも当てはまらなった単語は固有表現ではないとする手法である．しかし，この方法では辞書に含まれる固有表現しか抽出出来ないことに加え，本文中の固有表現の一部だけが辞書に含まれていた場合，間違った抽出が行われるといった問題が存在している（図 1）．

二つ目に利用できる資源として，生コーパスが挙げられる．生コーパスとは，アノテーションがなされていないテキストのことである．このようなテキストは学術記事や論文から低コストで収集できる．

これら二つの資源を利用した NER モデルとして，辞書の固有表現を元に，文字列マッチングによって教師データを作成する遠距離教師あり NER モデル [1, 2]

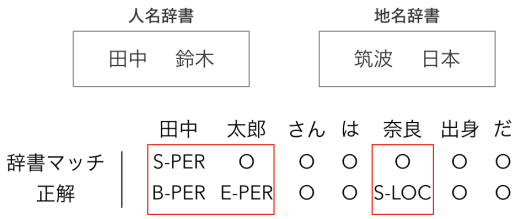

図 1: 辞書によるラベリングの失敗例

や PU 学習による NER モデル [6] 等が提案されている．

遠距離教師あり NER モデルの問題点として，表記ゆれや略称等の原因によって文字列マッチングによるラベリングが行われなかった単語の中にも固有表現は含まれているため，Recall が低下するという点が挙げられる．この問題については Jie ら [3]，辰巳ら [8] が取り組んできている．しかし，先程述べた文字列マッチングによる固有表現抽出と同様に，間違ったラベリングによる Precision の低下を無視することはできない．

本研究では，複数のモデルによる学習とラベル付けを繰り返し行い，辞書マッチにおけるラベリングの誤りを除去する手法を提案し，このような辞書マッチの誤りの考慮によって，NER における Precision の向上の可能性について検証する．

2 先行研究

辞書マッチングによって生成された教師データは，辞書に含まれる固有表現以外の単語にラベルが存在しない．このような教師データのことを一般的に部分的アノテーションコーパスと呼ぶ．部分的アノテーションコーパスを学習可能な NER モデルとして，Fuzzy-LSTM-CRF について説明する．Fuzzy-LSTM-CRF は Lample ら [4] の Bi-LSTM-CRF の CRF 層に変更を加え，ラベル列の一部が欠損している場合においても，学習が行えるように一般化したモデルである．ま

ず通常の CRF について説明する．CRF では単語列 $X = (x_1, \cdots, x_n)$, ラベル列 $\mathbf{y} = (y_1, \cdots, y_n)$ に対して，以下のようなスコア関数 $s(X, \mathbf{y})$ が定義される．

$$s(X, \mathbf{y}) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \quad (1)$$

ここでの $A_{y_i, y_{i+1}}$ はラベル y_i からその次のラベル y_{i+1} への遷移スコア，P_{i, y_i} はラベル y_i への重みである．

そして，CRF は唯一の正解であるラベル列 $\mathbf{y}$ への確率を最大化するように学習を行う．損失関数は以下のように表すことができる．

$$\begin{aligned} L(\mathbf{y}|X) &= -\log p(\mathbf{y}|X) \\ &= -\log \frac{e^{s(X,\mathbf{y})}}{\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}} \\ &= \log \sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})} - s(X, \mathbf{y}) \end{aligned} \quad (2)$$

ここで，Y_X は単語列 X のすべての取りうるラベル列を意味する．しかし，部分的アノテーションコーパスでは，一つの文章に対して複数の有効なラベル列が含まれる可能性がある．そのため可能なすべてのラベル列の合計確率を最大化するように学習を行う．数式では次の通りに表現できる．

$$\begin{aligned} L(\mathbf{y}|X) &= -\log \frac{\sum_{\tilde{y} \in Y_{possible}} e^{s(X,\tilde{y})}}{\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}} \\ &= \log \sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})} - \log \sum_{\tilde{y} \in Y_{possible}} e^{s(X,\tilde{y})} \end{aligned} \quad (3)$$

$Y_{possible}$ は部分的アノテーションによって与えられた取りうる有効なラベル列を意味する．また推論では，ビタビアルゴリズムを適用してスコア関数式 (1) が最大になるようなラベル列を出力する．

別の部分的アノテーションコーパスを学習できるモデルとして，Jie ら [3] の手法について説明する．Jie らは人手で行うアノテーションには，固有表現の部分にのみタグの付与が行われ，固有表現ではないと示すタグ，つまり O タグが付与されないと仮定するのが自然であるという主張に基づき，O タグが存在しない教師データに対して学習を行える手法として次の方法を提案した．

1. 部分的アノテーションコーパスを二つのサブデータセットに分割する．このとき、タグが付いていない単語に対しては O タグを付与する．
2. NER モデルをそれぞれ分割したサブデータセット毎に学習する．
3. 学習した NER モデルを利用し，もう一方のサブデータセットに対して制約付きビタビアルゴリズムによるアノテーションを行う．制約付きビタビアルゴリズムとは，人手で付与したタグを必ず採用するようにビタビアルゴリズムを適用する手法である．
4. 新しくアノテーションされたサブデータセットを元に再度モデルを学習する．
5. 3, 4 を複数回繰り返す．
6. 更新されたサブデータセットの両方を結合し，これを元に学習を行う．

遠距離教師学習を用いた NER において，辰巳 [8] らは辞書マッチによる自動アノテーションについて二つの問題があると主張した．一つ目は，辞書マッチによって作成された NER の教師データにはノイズを含むという問題である．二つ目は，略称や表記ゆれにより，文字列マッチングがうまくいかず False Negative が大量に出現するという問題である．これら二つの問題を解決するために，辞書を用いてコーパスを拡張し，教師データの Recall を高めた．本研究では辞書マッチの Recall ではなく Precision の向上を目指している．

3 提案手法

Jie らが提案したモデルは，部分的アノテーションコーパスに対して学習を行うことが可能だが，人手による正しいアノテーションが付与されることを前提とした学習方法になっている．本研究では，Jie の手法がアノテーションの失敗を考慮出来ていないことに着目し，サブデータセットに対してのアノテーション時に制約付きビタビではなく，通常のビタビアルゴリズムを利用した．また通常のビタビアルゴリズムを用いると，学習途中に辞書には含まれるが，生コーパスには出現頻度が少ないような固有表現が失われてしまう可能性が高い．そのため，提案手法ではモデルによるタグ予測の結果に対して，辞書マッチを行う．この場合，以下の三つのパターンが考えられる．

1. モデルによるタグ予測では固有表現と認識されないが，辞書マッチでは固有表現と認識される．

2. モデルによるタグ予測では固有表現とみなされるが，辞書マッチでは固有表現ではない．

3. モデルによるタグ予測で固有表現かつ，辞書マッチでも固有表現と認識される．

今回の実験では，図 2 のように，1 の場合のみ辞書マッチによるラベル付けを採用し，それ以外の場合にはモデルによるタグ予測を採用することにした．

図 2: モデルによる予測と辞書マッチ

4 実験

4.1 データセット

BC5CDR BC5CDR は 15,935 つの Chemical と 12,853 つの Disease，二種類の固有表現を含むデータセットであり，全 15,000 記事で構成されている．辞書は Shang ら [7] が公開しているデータセットを利用した．

CoNLL-2003 CoNLL-2003 は新聞記事を対象とした，固有表現抽出データセットであり，LOC, ORG, DATE, MISC の四種類の固有表現がアノテーションされている．教師データからアノテーションを削除したものを生コーパスとして利用した．CoNLL-2003 は，テスト用のデータセットが二つ用意されている．辞書は，このテスト用データセットのうち，評価に利用しなかった方から，固有表現を抽出し作成した．

4.2 比較対象

以下の四つの手法を比較対象として実験を行った．

辞書マッチング 辞書から直接文字列マッチングによって，固有表現を抽出する手法である．今回は単語に対して複数の固有表現がマッチングした場合，マッチングした中で最も長い固有表現を選択する最長一致法による文字列マッチングを採用し，実験を行った．

BiLSTM-CRF 生コーパスに対して辞書マッチによる自動アノテーションを行い，アノテーションが付与されなかった単語は O タグを付与して，Lample[4] らの Bi-LSTM-CRF モデルを学習する．

Fuzzy-LSTM-CRF BiLSTM-CRF と同様に生コーパスに対して辞書マッチによるラベル付与を行ったものを教師データとして学習を行う．ただしアノテーションが付与されなかった場合，全てのタグが付与される可能性があると仮定して学習を行う．

Jie's method Jie らの手法を辞書による自動アノテーション後の教師データに対して適用する．また提案手法同様，最終的なアノテーション結果に対して辞書マッチを行ったモデルについても確認を行う．

BiLSTM-CRF のハイパーパラメータに関しては Lample ら [4] のものを，提案手法と Jie's method のハイパーパラメータは Jie ら [3] と同様のものを採用した．

4.3 実験結果

実験結果は表 1 の通りである．全体的に Fuzzy-LSTM-CRF の Precision が低いという結果になった．このような結果が生じた原因は，辞書マッチによって生成された教師データ中には O タグが存在しないため，推論時に O タグの可能性をほとんど考慮しなくなっていたからだと思われる．両データセットに対して，Jie's method より提案手法のほうが Precision が高いという結果になった．これはビタビアルゴリズムが，教師データの生成時に起きた辞書による間違ったアノテーションが複数回に渡る相互アノテーションにより正しいタグになったからだと考えられる．しかし，Recall が低下したため，F1 値としては Jie's method よりも低い結果となった．また Jie's method に対して辞書マッチを適用しても，性能がほとんど変わらなかったのは制約付きビタビによるラベリングの影響で，辞書に含まれる固有表現を取り逃すことが少なかったからだと思われる．

各データセットについてみると，CoNLL-2003 では提案手法の Precision は辞書マッチングより高いものになった．理由として CoNLL-2003 の実験で利用した

表 1: 各モデルにおける性能比較

手法	BC5CDR			CoNLL-2003		
	Precision	Recall	F1	Precision	Recall	F1
辞書マッチング	**87.8**	61.7	72.4	49.2	29.3	33.8
BiLSTM-CRF	88.4	61.8	72.6	72.5	39.2	50.9
Fuzzy-LSTM-CRF	11.2	**85.5**	19.3	18.1	**80.0**	26.4
Jie's method	81.4	73.7	77.3	77.7	49.1	57.5
Jie's method + 辞書マッチング	81.4	73.8	**77.3**	77.7	49.1	**57.5**
提案手法	83.3	69.3	75.6	78.7	47.8	57.0
提案手法 + 辞書マッチング	83.4	71.7	77.1	**78.7**	47.8	57.0

辞書は自動生成によって作成されたものであり，LOC と ORG を間違えてタグ付けを行うようなミスが多く見られたからだと思われる．このように辞書マッチングの精度が不安定な場合，Precision を重視した目的において提案手法は有効であると考えられる．ただし，BC5CDR のような辞書マッチングがある程度機能しているケースにおいては，提案手法は Precision, Recall の両面において有効性がないと言える．

5 おわりに

本稿では，遠距離教師あり固有表現抽出手法において，辞書マッチの誤りを考慮することでパフォーマンスの向上ができるかどうかを検証した．その結果，元の辞書マッチの性能が高い場合には提案手法は Precision, Recall の両面において有効性がないが，低い場合においては Precision の向上がみられた．

5.1 謝辞

本研究の一部は，JSPS 科研費（課題番号 16H02904，19K20333）および JST CREST（#JPMJCR16E3）AIP チャレンジの助成によって行われた．

参考文献

[1] Jason A. Fries, Sen Wu, Alexander Ratner, and Christopher Ré. Swellshark: A generative model for biomedical named entity recognition without labeled data. *CoRR*, Vol. abs/1704.06360, , 2017.

[2] Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. Unsupervised aspect term extraction with b-LSTM & CRF using automatically labelled datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 180–188, 2017.

[3] Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. Better modeling of incomplete annotations for named entity recognition. In *Proc. NAACL*, p. 729–734, 2019.

[4] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proc. NAACL*, pp. 260–270, 2016.

[5] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proc. ACL*, pp. 1064–1074, 2016.

[6] Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proc. ACL*, pp. 2409–2419. Association for Computational Linguistics, 2019.

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

[8] 守祐辰巳, 啓介後藤, 裕之進藤, 裕治松本. 辞書を用いたコーパス拡張による化学ドメインの distantly supervised 固有表現認識. Technical Report 7, 奈良先端科学技術大学院大学, 2019.