E-022

固有表現抽出のためのパープレキシティを用いた外部情報検索
External information retrieval with Perplexity for Named Entity Recognition

美野 秀弥† 後藤 淳† 山田 一郎†
Hideya MINO Jun GOTO Ichiro YAMADA

1. はじめに

固有表現抽出は，文章中の人名，地名，組織名などの固有表現を抽出して「人名」，「地名」，「組織名」などのクラスに分類する自然言語処理における基盤技術の一つであり，質問応答システムや匿名化技術などの応用先がある．本論文では，固有表現抽出タスクに外部情報（外部データ）を活用する手法に着目し，固有表現抽出の精度向上に有効な外部データを取得する手法を提案する．提案手法は，外部コーパス中のデータから，対象データとの意味的類似度と言語モデルの評価に用いるテストセットパープレキシティ値の２つを組み合わせたスコア計算式を用いて固有表現抽出モデルに入力する外部データを取得する．提案手法により取得した外部データを固有表現抽出の対象データと結合して固有表現抽出モデルに入力し，このモデルを学習する．提案手法の効果を確認するために，生物医学のデータセット NCBI[5]とソーシャルメディアのデータセット WNUT-17[4]の２種類のコーパスを用いて評価実験を行った．実験結果より，提案手法は従来手法と比較して高精度に固有表現を抽出できることが確認された．

2. 先行研究

近年，ルールベースの固有表現抽出手法[6,11]に加え，ニューラルネットワークを用いた固有表現抽出手法[9,15]が盛んに研究されている．あらかじめ固有表現が付与された学習データを増やすことで精度は向上するが，人手によるデータ作成のコストがかかるという課題がある．この課題に対して，固有表現を付与したデータを増やさずに，外部データを用いて固有表現抽出の精度を向上する手法[13,16]がある．

Yamada ら[16]は，対象データの周辺の文を用いて固有表現抽出の精度が向上することを示した．しかし，学習時と推論時ともに周辺の文を用意する必要があるため，利用条件が限定される．

Wang ら[13]は，周辺の文の代わりに，固有表現抽出の対象データをクエリとして意味的に関連する文章を外部コーパスから選択して用いる手法を提案し，固有表現抽出の精度が向上することを示した．しかし，固有表現抽出に有効なデータが対象データと意味的に近い文であるという仮定を元にした手法であり，外部コーパスの中に意味的に近い文がなかった場合を想定していない．

本稿では，Wang ら[13]の手法をベースにして，より適切な外部データの取得手法を提案する．

† NHK 放送技術研究所　NHK Science & Technology Research Laboratories

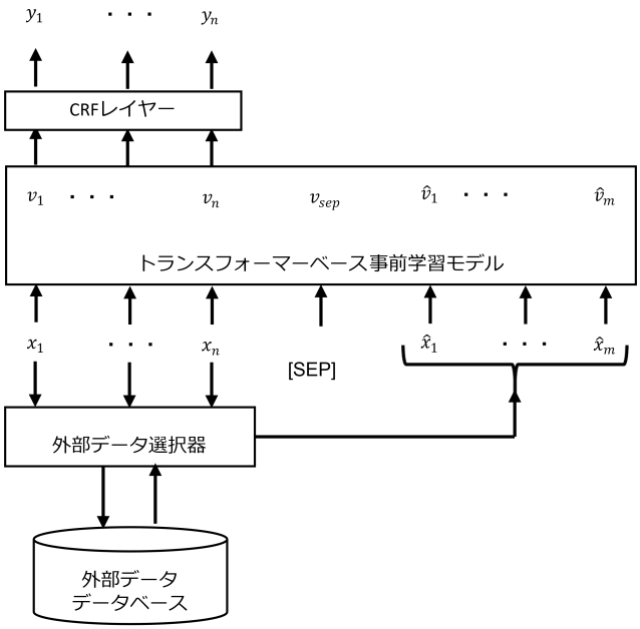

図１：外部データを活用した固有表現抽出器

3. 提案手法

本稿では，外部データを活用した固有表現抽出手法において，有効な外部データが意味的に類似した文と言語モデルが近い文であることと仮定し，固有表現抽出の際に用いる外部データを選択するための新たな外部データ検索手法を提案する．従来手法が用いていた固有表現抽出の対象データと外部コーパスの各データとの類似度S_{score}に加え，固有表現抽出するコーパスで学習した言語モデルを用いて外部コーパスの各データのテストセットパープレキシティ値P_{score}を用いた下記の計算式を提案する．

$$N_{score}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \alpha S_{score}(\boldsymbol{x},\hat{\boldsymbol{x}}) + (1-\alpha)P_{score}(\hat{\boldsymbol{x}}) \quad (1)$$

α は両者の重みを制御するハイパーパラメータである．固有表現抽出の対象データの文を$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$とし，$n$ は対象データのトークン長を示す．外部データとして用いる文は$\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$，$m$ は外部データのトークン長を示す．外部データのテストセットパープレキシティ値P_{score}は固有表現抽出対象のコーパスであらかじめ計算した言語モデルを基に下記で計算する．

$$P_{score}(\hat{\boldsymbol{x}}) = \sqrt[m]{\prod_{i=2}^{m} \frac{1}{p(\hat{x}_i|\hat{x}_1, \ldots, \hat{x}_{i-1})}} \quad (2)$$

コーパス	学習	開発	テスト	タグ数	平均長 (文字数)
NCBI	5,424	923	940	2	25.01
WNUT-17	3,394	1,009	1,287	7	18.48

表１：コーパスの統計情報

提案手法により外部コーパスの各データにスコアが付与され，スコアの高い順に固有表現抽出モデルに用いるデータの候補とする．提案手法は，外部コーパスの中に類似度が高いデータを多く含む場合には類似度が高く，かつ言語モデルが近いデータを選び，外部コーパスの中に類似度が高いデータがない場合には類似度のみに頼らずに言語モデルが近いデータを優先的に選ぶことが期待できる．

図１に外部データを活用した固有表現抽出器を示す．固有表現抽出モデルには，Akbik ら[1]が用いた CRF レイヤーを付け加えたトランスフォーマーベースの事前学習済みモデルを用いる[1]．学習時と推論時ともに，事前学習モデルには，入力文と外部データを，セパレータ（[SEP]）を介して結合して入力する．入力データは事前学習モデルによりベクトル化（$v = (v_1, v_2, \dots, v_n), \hat{v} = (\hat{v}_1, \hat{v}_2, \dots, \hat{v}_m)$）され，CRF レイヤーには$v = (v_1, v_2, \dots, v_n)$を入力する．CRF レイヤーは，固有表現抽出結果として，入力トークン毎に固有表現を表すタグまたは固有表現ではないことを示すタグ（$y = (y_1, y, \dots, y_n)$）を出力する．

4. 評価実験

4.1 実験設定

提案手法の効果を確認するために，下記２種類の固有表現抽出タスクのデータセットを用いた実験を行った．

１． 生物医学のデータセット：NCBI [5]

２． ソーシャルメディアのデータセット：WNUT-17 [4]

各データセットの統計情報を表１に示す．用いる２種類のタスクには，ジャンルの違いに加えて固有表現タグの数（表１のタグ数の列）の違いがある．

外部データを選択する対象の外部コーパスには，本論文で用いた NCBI と WNUT-17 のコーパスと固有表現抽出タスクとして公開されているソーシャルメディアのデータセット WNUT-16 [12]，ニュースのデータセット CoNLL++ [14]，生物医学のデータセット BC5CDR[8]の５つのコーパスを１つにまとめたコーパスを用いた．外部コーパスのデータ数は 28,365 文であった．

文の意味的類似度S_{score}を計算する手法には，Wang ら[13]が用いた BERTScore [17]を用いた．固有表現抽出モデルに入力する外部データには，提案手法で最も高いスコアのデータ１文を用いた．

単語のベクトル化には，NCBI では Bio-BERT [7]を，WNUT-17 では XLM-RoBERTa [3]をそれぞれ用いた．固有表現抽出モデルの学習時は，バッチサイズ 4 で学習し，AdamW [10]を用いて最適化した．学習率は5×10^{-6}とした．CRF レイヤーの学習率は 0.05 とした．モデルは最大 200 エポックまで学習し，各エポック終了後の開発データの精度が最も高かったモデルを評価に用いた．式(1)のハイパーパラメータαについては[0.1, 0.25, 0.5, 0.75, 0.9]の値で開発データを用いた事前実験を行い，最良の精度となった値として NCBI は$\alpha = 0.9$，WNUT-17 は$\alpha = 0.50$を用いた．学習，推論ともに，NVIDIA A100 を用いた．

ベースラインには，外部データを用いないモデルと，意味的類似度のみを用いて外部データを選択した結果を用いたモデルを用いた．また，提案手法の効果の信頼度を上げるために，外部データとして外部データ群の中からランダムに選択したデータを用いて学習したモデルをベースラインに加えた．さらに，先行研究で State-of-the-art を達成している CL-KL [13]もベースラインに加えた．

評価尺度には，NCBI には F1 を，WNUT-17 には Micro-F1 と Macro-F1 をそれぞれ用いた[2]．

手法	NCBI	WNUT-17	
	F1	Micro-F1	Macro-F1
CL-KL[13]	88.30	60.18	52.02
外部データなし	88.72	59.44	52.27
BERTScore	89.33	59.08	50.98
ランダム	87.42	60.04	51.78
提案手法	90.12	60.88	53.13

表２：実験結果

手法	NCBI		WNUT-17	
	学習	テスト	学習	テスト
BERTScore	52.5	49.5	56.4	7.0
ランダム	2.1	2.4	0.8	0.8
提案手法	52.6	38.8	36.5	6.0

表３：外部データにおける固有表現のカバー率

4.2 実験結果

実験結果を表２に示す．表中の値はランダムシードの異なる３つのモデルを学習して得られた結果の中央値である．

外部データを用いない手法と比較して提案手法の精度は Micro-F1，Macro-F1 ともに向上しており，外部データを用いることで固有表現抽出の精度が向上したことを確認した．ランダムな外部データを用いた手法と比較しても精度が向上しており，提案手法の効果が正則化によるものだけではないことを裏付けている．

BERTScore を用いて選択した外部データを利用した手法との比較においても，提案手法の精度が向上したことを確認した．テストセットパープレキシティを用いることで固有表現抽出の精度が向上すると考えられる．これは，外部データとして，対象データとの類似度だけでなく言語モデルが近いデータを用いた方がよいことを示唆している．

従来手法である CL-KL との比較においても提案手法の精度が高く，提案手法の有効性を確認した．

BERTScore を用いた手法は，WNUT-17 において，外部データなしの手法よりも低下した．この原因を考察するために，外部データに対象データ中の固有表現がどの程度含ま

[1] CRF レイヤーを用いることで出力するラベルの前後関係をより考慮した学習が期待できる．

[2] WNUT-17 は固有表現タグが２種類以上あるため，多クラス分類の評価尺度である Micro-F1 と Macro-F1 を用いた．

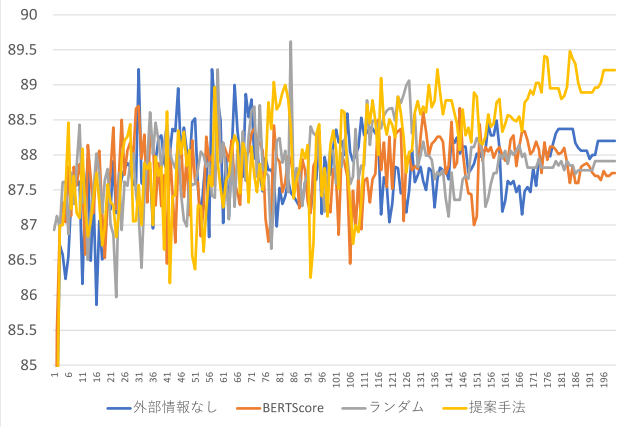

図２：学習時における開発データの精度の推移（NCBI）

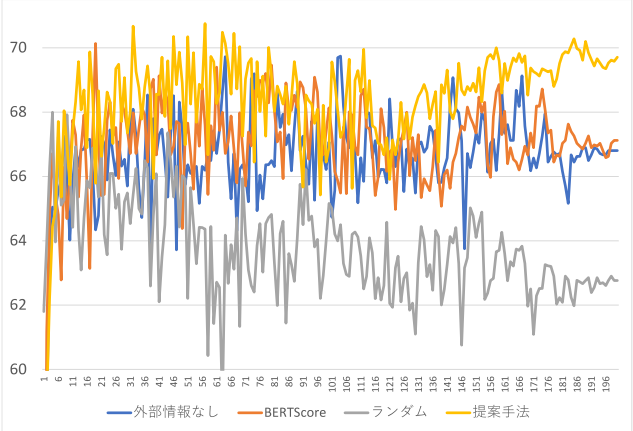

図３：学習時における開発データの精度の推移（WNUT-17）

れているかを調査した．表３は，各コーパスの学習データとテストデータのそれぞれの外部データに対象データ中の各固有表現が含まれているかどうかの割合（カバー率）を示している．表３より，学習データとテストデータとの間でカバー率が異なっており，WNUT-17 ではこのギャップが大きいことが分かる．学習時と推論時間のデータのギャップは exposure bias[2]と呼ばれており，これが精度に影響している可能性がある．提案手法では，WNUT-17 において，このギャップが抑制されている．

また，学習エポックごとの開発データの精度の推移のグラフを図２と図３にそれぞれ示す．提案手法は，NCBI, WNUT-17 ともに学習の後半においても精度が低下しないことを確認した．

学習時間については，NCBI では外部データなしで 6.5 時間，提案手法は 8 時間かかった．WNUT-17 では外部データなしで 6 時間，提案手法は 8 時間を要した．提案手法は外部データを用いており，入力データ長が長くなったために学習時間が増加したと考えられる．

5. おわりに

本稿では，従来の外部データを用いた固有表現抽出手法の精度を向上させるために，類似度とテストセットパープレキシティの２つの尺度を用いたスコア計算式を用いて有効な外部データを取得する手法を提案した．生物医学とソーシャルメディアの２種類の固有表現抽出データセットを用いて評価実験を行い，提案手法の効果を確認した．

提案手法では固有表現抽出タスクに有効な外部データが対象データとの類似度と対象ジャンルと言語モデルが近いデータであると仮定したが，その他にも有効な特徴があると考えられる．今後，データの観察などを通して固有表現抽出に有効な外部データの有効な特徴を見つけていきたい．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究（課題 225）により得られたものです．

参考文献

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 1171–1179. Curran Associates, Inc.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440– 8451, Online. Association for Computational Linguistics.

[4] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

[5] Rezarta Islamaj Dog˘an, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics, 47:1–10.

[6] Ji-hwan Kim and P.C. Woodland. 2000. A rule-based named entity recognition system for speech input. In Proceedings of the International Conference on Spoken Language Processing, pages 521–524.

[7] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

[8] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas CWiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database: The Journal of Biological Databases and Curation, 2016.

[9] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering.

[10] Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In International Conference on Learning Representations.

[11] Satoshi Sekine and Chikashi Nobata, “Definition, dictionaries and tagger for extended named entity hierarchy,” in Proc. 4th Int. Conf. Lang. Resour. Eval., 2004, pp. 1977–1980.

[12] Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity

recognition shared task. In Proceedings of the 2nd Workshop on Noisy User-generated Text, pages 138–144, Osaka, Japan.
[13] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning. In the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics.
[14] Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. CrossWeigh: Training named entity tagger from imperfect annotations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5154–5163, Hong Kong, China. Association for Computational Linguistics.
[15] Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2145–2158.
[16] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6442–6454, Online. Association for Computational Linguistics.
[17] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations.