データ拡張による固有表現抽出の不確実性推定

橋本 航　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
{hashimoto.wataru.hq3, kamigaito.h, taro}@is.naist.jp

掲載号の情報

32 巻 3 号 pp. 829-858.
doi: https://doi.org/10.5715/jnlp.32.829

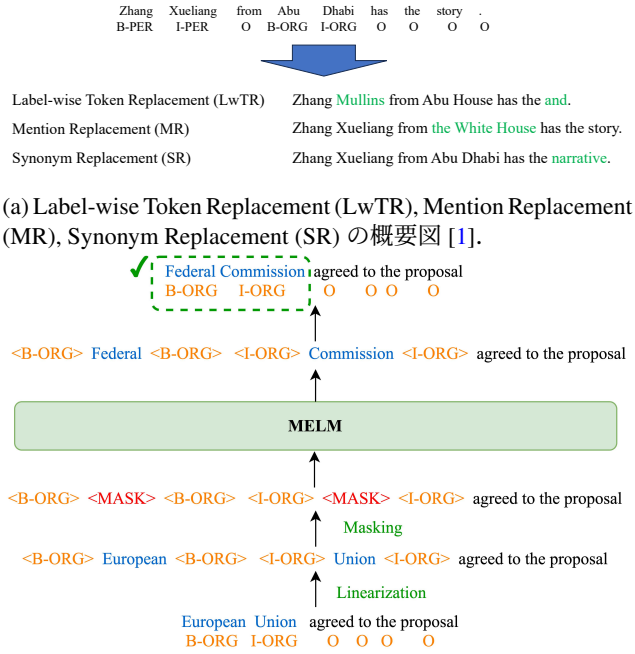
概要

近年の固有表現抽出 (Named Entity Recognition; NER) の発展は，BERT や DeBERTa などの事前学習済み言語モデル (Pretrained Language Models; PLMs) の発展に支えられている．しかし，Deep Neural Networks (DNNs) および PLMs は予測の不確実性や確信度をしばしば正しく推定できない傾向にあることがわかっている．この問題により，医療や金融などの高い確信度での予測の誤りが致命的となる領域では DNNs および PLMs の適用が制限される．既存アプローチとして，複数回の確率的推論から予測の不確実性を求める方法があるものの，計算コストが高い．

そこで，本研究ではデータ拡張に着目する．データ拡張によって生成された多様なデータを訓練データ集合に追加したうえでモデルを学習することで，モデルが入力に対してロバストになるため，不確実性推定の改善につながることが期待される．さらに，データ拡張は訓練時間は増やすものの，モデル構造や推論過程を変えるわけではないため，推論時間の増加は起こらない．本研究では，図 1 に示すような，4 つのトークン置換およびエンティティ置換に基づく NER のデータ拡張手法が不確実性推定性能に与える影響を，ジャンル横断および言語横断の設定の双方から網羅的に調査した．

実験結果から主に以下の知見が得られた．第一に，NER におけるいくつかのデータ拡張は，ドメイン内設定[1)]では一貫して不確実性推定性能の改善につながることが判明した．一方で，ドメイン外における不確実性推定性能の改善は限定的であることもわかった．第二に，データ拡張によって生成された文のパープレキシティが低いほど，データ拡張サイズを増やすと不確実性推定性能は改善される傾向にあることが判明した．

1） 本研究では訓練データとテストデータの性質 (ジャンルおよび言語) が同じであることをドメイン内，異なることをドメイン外として扱っている．



(a) Label-wise Token Replacement (LwTR), Mention Replacement (MR), Synonym Replacement (SR) の概要図 [1].

(b) Masked Entity Language Modeling (MELM) の概要図 [2].

図 1: 本研究で用いるデータ拡張手法.

参考文献

[1] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 3861–3867, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[2] Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. MELM: Data augmentation with masked entity language modeling for low-resource NER. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2251–2262, Dublin, Ireland, May 2022. Association for Computational Linguistics.