固有名詞置換による共参照解析データセットの拡張

富村 勇貴 [1]　上垣外 英剛 [1]　渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学大学
{tomimura.yuki.tw1, kamigaito.h, taro}@is.naist.jp

概要

共参照解析は共参照情報のアノテーションが高コストであるため、深層学習の際に利用できるデータセットが限られているという問題がある．そこで本研究では，共参照解析において学習データ内の共参照クラスターに含まれる固有名詞を別の固有名詞に置換することでデータセットを拡張し，拡張データで学習した共参照解析モデルの性能を向上させることを目的とする．実験の結果，人物名の固有名詞を性別と特殊な例を考慮してランダムに置換する方式で拡張したデータセットで学習したモデルがベースラインモデルの性能を僅かに上回り，訓練データにおいて未知の固有名詞に対しての予測精度が向上していることがわかった．

1　はじめに

共参照解析とは，自然言語処理におけるタスクの一つである．文書内のある単語列同士が現実において同じ実体（エンティティ）を指している場合，それらの単語列は共参照関係であるとされ，共参照解析は文書内の共参照関係を全て検出することを目的とする．共参照解析モデルの学習に用いられるデータセットは多くの場合手動で共参照チェーン情報をアノテーションされた大規模コーパスである CoNLL-2012[1] が用いられている．しかし，この手動アノテーションは非常に高コストであり，共参照解析モデルの学習に利用できるデータセットが限られているという問題がある．そのため，現行の共参照解析モデル [2, 3, 4] は CoNLL-2012 の特性により様々なバイアスがかかっていると考えられる．頑健なモデルを構築するためには，そういったバイアスの影響を小さくする必要がある．そこで，データ拡張によりこのバイアスの影響を小さくする方法を提案する．

関連研究では CoNLL-2012 には出現する男性代名詞が全体の 80%以上を占めていることからジェンダーによるバイアスが生じていると考えられ，人物名を E1，E2 といった匿名記号に変え，代名詞の性別を反転した補助データセットを作成することでジェンダーバイアスの影響を小さくするといったもの [5] がある．しかし，この方法で拡張したデータセットを用いて学習したモデルの全体の予測性能は低下している．その理由の一つには，人物名を記号化することで扱う文書の内容が不自然になるというノイズの要素が考えられる．

そこで，本研究ではデータ拡張手法の一つである **Type Swaps**[6] を基に，CoNLL-2012 に含まれている固有名詞に付与されているエンティティタイプを利用して，固有名詞を別の固有名詞に置換することで匿名化をせずにデータセットの拡張を行う手法を提案する．エンティティタイプ毎に無秩序に置換を行ったデータットで学習した場合，モデルの精度は大きく低下するという結果となった．その理由は置換の際に，共参照解析は先行詞と照応詞との意味的な乖離が発生する場合，共参照関係を検知することが困難になってしまうためであると考えられる．そこで，固有名詞をより細分化したエンティティタイプに分類し，同タイプの名詞の中で置換を行うことでこの問題を解決した．実験の結果，人名を性別毎に分類し，**God** やアメリカ大統領といった文書において特別な意味を持つ単語は置換しないという工夫を行った置換データセットで拡張したデータセットで学習を行ったモデルの精度はベースラインモデルの性能を約 0.7 ポイント上回った．

2　CoNLL-2012

CoNLL-2012[1] は放送会話，ニュース，雑誌記事，ニュースワイヤー，聖書，電話会話，ウェブテキストなどの様々なドメインの文書で構成された大規模多言語コーパスである．本研究では，英語テキストデータセットのみ対象に扱う．

CoNLL-2012 には固有名詞に対して PERSON（人名），NORP（国籍，または宗教的・政治的団体）な

どのエンティティタイプの情報がアノテーションされている．これらのエンティティタイプ毎に，各固有名詞が訓練データ内の全共参照クラスターに含まれている時の回数の内上位 4 位までを表にしたものが表 1 である．置換の回数は同じ文書内で同じ名詞を置換する場合は 1 回とカウントしている．

表 1 CoNLL-2012 訓練データ内の共参照クラスターに含まれる固有名詞のエンティティタイプ別出現回数（上位 4 位まで）

エンティティタイプ	出現回数
PERSON（人名）	6068
GPE（国，都市，州など）	4414
ORG（企業，機関など）	3969
NORP（国籍，政治的団体など）	2078

表 1 から分かるように，最も多く共参照クラスター内に出現する固有名詞は PERSON（人名）であり，次点で ORG（企業，機関など），GPR（国，都市，州など）と続く．そこで，出現回数の多いほど置換した際の影響力が大きいと考えられるため，今回の実験ではこれらの内上位 3 位までの PERSON，ORG，GPE に対して置換を行うことにした．また，固有名詞の有無に関わらず，訓練データセット全体に含まれる共参照クラスターの数は 35,143 個なので，全 PERSON に対して置換を行った場合，全共参照クラスターの内，約 17%を置換することになる．

3 提案手法

本研究ではデータ拡張手法の一つである **Type Swaps**[6] を用いる．**Type Swaps** は抽出的質問応答データセットにおいて，回答となる Mention を **human**，**country** といったタイプ毎に分類し，各 Mention を同タイプの異なる Mention に置換することでデータ拡張を行うといった手法である．本研究では共参照解析を扱うため，ある固有名詞が共参照クラスターに含まれている場合に，アノテーションされたエンティティタイプによって **Type Swaps** を行う．また，実験初期に共参照解析における置換に対する問題点があることが発覚したため，それぞれのタイプに応じて工夫を加えた置換を行う．

PERSON に対する置換の工夫 PERSON（人名）タイプを持つ固有名詞に対する置換は性別に関する問題と，特別な人物名詞に関する問題がある．

まず，性別に対する問題は，先行詞である固有名詞（**John**，**Cathy** など）に対して，照応詞は he や she などの性別に応じた代名詞であることがほとんどであるため，she で照応される先行詞が一般に男性として認知される名前に置換されてしまう場合，それらが共参照関係であると検知されづらくなるといった問題が発生する．そこで，与えられた名前に基づいて、その名前が最も一般的に使用されている性別を統計的に予測するオンラインサービス **Genderize.io**（https://genderize.io/）を用いて，PERSON タイプの固有名詞を Male，Female，Family Name or Unidentifiable の 3 タイプに細分化し，Male と Female についてそれぞれ置換を行う．この工夫により置換回数は 6068 回から 4163 回になった．

もう一つの問題点は，CoNLL-2012 がニュースや聖書によって構成されているため，**God**，**Jesus** やアメリカ大統領の名前が，非常に多くの回数訓練データセットに登場するということである．（**God** は 2594 回，**Jesus** は 1755 回，**president** は 3057 回）これらの名詞は照応詞が **God**，**president** でほぼ固定化されているため，他の名詞に置換すると共参照関係と検知され辛くなるなど，名詞自体が持つ意味が文書において特別であることが置換の弊害となる．そこで，これらの特別な名詞は置換しないといった工夫を行った．この工夫により置換回数は 4163 回から 3145 回になった．

GPE に対する置換の工夫 GPE（国，都市，州など）タイプを持つ固有名詞に対する置換における問題点は国，都市，州といった異なる文脈，照応詞で扱われる固有名詞が全て GPE タイプで纏められていることである．そのため，China（国）→ Kansas（アメリカの都市）といった置換が行われてしまい，文脈に沿わず，照応詞との整合性も取れない置換になることが多い．そこで，GPE タイプを国毎の都市名に分け，同じ国内の都市名，州名を置換するとようにするといった工夫を行った．この工夫により置換回数は 4414 回から 2537 回になった．

ORG に対する置換の工夫 ORG（企業，機関など）に対しても同様の問題点があり，企業，機関の種類が様々であるために，PERSON や GPE と同様に照応詞との整合性が取れないといった問題がある．しかし，ORG においてはその分類が PERSON や GPE と比較してあまりに複雑である（国際組織，株式会社，ニュース，委員会，銀行，病院など多岐に渡る）ために，同様にサブクラスに分けるといった手法を取ることが難しく，残念ながら今回の実験では ORG に対して特に工夫は行わなかった．

表 2　各置換データセットと拡張データセットによって学習したモデルの性能評価

	$coref_p$	$coref_r$	$coref_f$	$mention_r$
基本データセット（ベースラインモデル）	78.9	78.7	78.8	96.6
置換データセットのみ				
PERSON（工夫無し）	79.4	71.8	75.4	93.3
PERSON（性別）	79.6	78.3	78.9	96.5
PERSON（性別+God）	79.1	79.3	79.2	96.5
GPE（工夫無し）	78.3	72.2	75.1	92.5
GPE（国別都市置換）	78.8	79.0	78.9	96.8
ORG（工夫無し）	79.3	71.8	75.3	94.0
+ 基本データセットと合わせての学習				
PERSON（工夫無し）	77.6	72.5	75.0	93.4
PERSON（性別）	79.0	78.9	78.9	97.2
PERSON（性別+God）	79.8	79.3	**79.5**	97.2
GPE（工夫無し）	79.4	76.6	78.0	94.1
GPE（国別都市置換）	79.2	79.1	79.1	97.3
ORG（工夫無し）	79.5	77.0	78.2	95.0

4　実験

各エンティティタイプについて提案手法により置換を行ったデータセット単体と，置換したデータセットを元のデータセットを組み合わせた拡張データセットとを用いて深層学習を行い，学習したモデルを比較する．

データセット　学習には CoNLL-2012 の訓練用データ（2769 文書），検証用データ（342 文書），評価にはテスト用データ（347 文書）を用いる．

事前学習済みモデル　事前学習済みモデルには SpanBERT-large [7] を使用する．SpanBERT はテキストのスパンをより良く表現し予測するために設計された事前学習方法であり，共参照解析のようなスパン選択タスクに適している．

学習に使用したモデル　学習に使用したモデルは Coarse-to-fine Inference モデル [4] で，このモデルは Text-to-Text 形式に変換するといった，データセットに特別な操作を加えずに扱える共参照解析モデルにおいて最先端の性能を持つことが知られている．

4.1　実験設定

実験設定の詳細は付録 A にて記載する．

4.2　評価指標

共参照解析モデルの評価に使われる一般的な指標は主に B^3，MUC，$CEAF$ があり [8]，これらの指標により共参照解析における精度（precision 値）と再現率（recall 値）とそれらの調和平均（F1 値），加えて Mention（言及）を検出できているかを測る Mention recall 値をこれらの指標の平均値を用いて評価する．

4.3　実験結果

実験の結果を表 2 に記す．表 2 によると，基本データセットを工夫無しにエンティティタイプ別にランダムに置換して作成したデータセットで学習したモデルはいずれも精度が大幅に低下している．特に再現率において精度が下がっている理由は，おそらく工夫無しの置換データセットは 3 節で述べた問題点により，先行詞と照応詞の実体に大きな乖離があるために，先行詞と照応詞のペアワイズスコアに関する学習に支障をきたしているためであると考えられる．この結果に対して，PERSON タイプの置換では，エンティティタイプをより細分化した性別を考慮した置換データセットと，さらに **God** やアメリカ大統領といった文書において特別な意味を持つ固有名詞に対しては置換しないという工夫を加えた置換データセットでは，どちらも精度が基本データセットとほぼ同じになるほど改善している．**GPE** に対する置換も同様に，国別の都市毎に置換したデータセットでは精度が改善している．そのため，データ拡張のために安易に別の固有名詞に置換を行うことは逆効果であり，よりエンティティタイプを

細分化するなどの工夫により，先行詞と照応詞の埋め込みベクトルがより近いものになるように置換することが重要であると考えられる．

また，置換データセットと元のデータセットを合わせて学習を行った場合，置換データセットのみの場合と比較して全体の傾向としては精度が向上している．しかし，工夫無しの置換データセットを用いている場合は元のデータセットのみの場合よりも精度が落ちているため，やはり先行詞と照応詞の共参照関係が崩れているデータセットを作成することはモデルの精度に悪影響であることが分かる．しかし，工夫ありの拡張データセットはいずれも精度がベースラインモデルを僅かに上回っており，エンティティタイプ毎の固有名詞置換によるデータ拡張が共参照解析において有効である可能性を示している．最も結果が良かったのは，PERSON タイプの固有名詞を性別のサブタイプに細分化し，特殊な名詞は置換しないという工夫を加えた拡張データセットであり，ベースラインモデルと比較して+0.7 ポイントの精度向上が見られた．

4.4 拡張データセットによる未知のエンティティに対する頑健性

表 2 から，精度が向上した拡張データセットは共通して Mention Recall 値に向上が見られることから，データ拡張によりモデルの頑健性が向上していると期待される．

未知のエンティティに対する実験 テストデータの共参照クラスターに含まれるエンティティ（PERSON，GPE）が、訓練データに含まれないものがあるという条件で、テストデータを絞り込むと PERSON タイプの場合 347 文書→ 167 文書に，GPE タイプの場合 347 文書→ 87 文書に絞り込まれた．これらの絞り込んだテストデータの文書は，訓練データに含まれない未知のエンティティを持つ共参照クラスターを含んでいるため，これらの文書に対して性能を評価することで，モデルの頑健性が向上しているかを確かめることとする．実験の結果を表 3 に纏めた．表 3 によると，まず未知のエンティティを含むテストデータに対しては全体的に精度が低下するという傾向が見られる．また，性別＋ God などの名詞は無置換の工夫を加えた拡張データセットは精度が大きく向上しているものの，期待に反して他二つの F 値は変わらないか，下がっている．そのため，今回の実験でデータ拡張を行うことで特に未知のエンティティに対する共参照解析の精度が向上するという明確な結論を支持するには不十分である．ただし，拡張データセットはいずれの場合も Mention Recall 値が向上しているため，データ拡張によって，より Mention を検出するようにモデルが学習するようになるという知見が得られた．

表 3 訓練データにおいて未知のエンティティを含むテストデータに対するモデルの性能評価

	$coref_p$	$coref_r$	$coref_f$	$mention_r$
基本データセット（テストデータ：PERSON）	76.0	75.9	75.9	96.4
基本データセット（テストデータ：GPE）	76.9	79.0	77.9	96.2
+ 基本データセットと合わせての学習				
PERSON（性別）	76.0	76.0	76.0	96.9
PERSON（性別+God）	77.1	76.7	**76.9**	97.0
GPE（国別都市置換）	77.1	77.9	77.5	96.6

5 おわりに

本研究では，アノテーションが高コストである共参照解析の代表的なデータセットである CoNLL-2012 を，固有名詞を同エンティティタイプの名詞に置換することで拡張することを行った．その結果，先行詞と照応詞との意味的な距離が乖離しないようにエンティティをより細分化されたサブタイプに分類するか，置換に制限を掛けること (3 節) で拡張したデータセットで学習したモデルは精度が僅かに向上した．また，今回の研究では CoNLL-2012 から置換候補の固有名詞を抽出しているため，データセット内の語彙は拡張されていない．しかし，訓練データにおいて未知のエンティティを含むテストデータに対しては全体的に精度が落ちることが実験で判明したため，外部のデータベースなどを用いて置換先の単語の語彙の拡張を行うことの有効性の検証を次の課題とする．

参考文献

[1] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue, editors, **Joint Conference on EMNLP and CoNLL - Shared Task**, pp. 1–40, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[2] Vladimir Dobrovolskii. Word-level coreference resolution. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7670–7675, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Bernd Bohnet, Chris Alberti, and Michael Collins. Coreference resolution through a seq2seq transition-based system. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 212–226, 2023.

[4] Luke Zettlemoyer Kenton Lee, Luheng He. Higher-order coreference resolution with coarse-to-fine inference. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, p. 687–692, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[5] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[6] Jonathan Raiman and John Miller. Globally normalized reader. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 1059–1069, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[7] Yinhan Liu Daniel S. Weld Luke Zettlemoyer Omer Levy Mandar Joshi, Danqi Chen. Spanbert: Improving pre-training by representing and predicting spans. July 2019.

[8] Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 632–642, Berlin, Germany, August 2016. Association for Computational Linguistics.

A 付録 (Appendix)

4 節での実験設定をここに記す.

- Transformer Model: SpanBert-large-cased[7]
- Model Architecture: Course-to-Inference[4]
- Max Length: 512
- Feature Size: 20
- Max Span Width: 30
- Transformer Dimension: 1024
- Batch Size: 1
- Number of Epochs: 40
- Patience: 10
- Learning Rate for Transformer Layer: 1×10^{-5}
- Initial Learning Rate: 3×10^{-4}
- Maximum Number of Antecedents: 50
- Spans per Word: 0.4
- Coarse to Fine: True
- Inference Order: 2