［研究助成（C）］

単語単位ラベルを考慮したテキストデータ拡張手法の研究

2237007

研究代表者 （助成金受領者）	同志社大学　大学院 文化情報学研究科	博士後期課程	寺　本　優　香
共同研究者	同志社大学　文化情報学部	教　授	波多野　賢　治
	名古屋大学　数理・データ科学 教育研究センター	准教授	駒　水　孝　裕

［研 究 の 目 的］

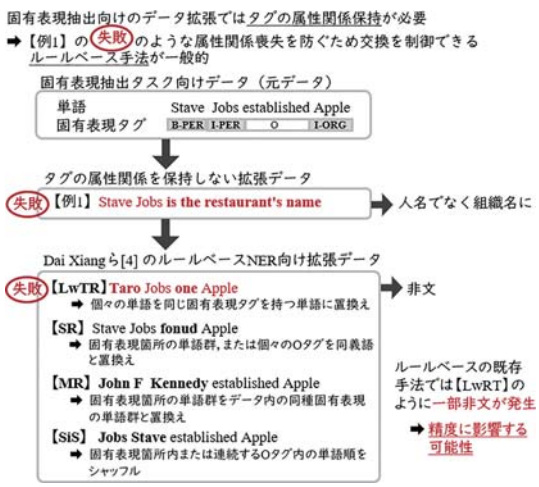
固有表現抽出とは，文中から人名や地名，組織名，日付表現，薬品名，遺伝子名といった特定の属性を持つ単語を抜き出す自然言語処理の基礎技術であり，医療・創薬分野での情報抽出やチャットボットなど様々な分野で利用されている。

このタスクでは教師ありの機械学習モデルが頻繁に利用されており，近年は深層学習モデルが優れた性能を示している。しかし，その学習には大規模かつ高品質なデータが必要であり，特に分野固有の固有表現を扱う場合は，専門知識を持つアノテータによるデータ作成が不可欠であり，人的コストが課題となっている。

固有表現抽出タスクで中心的役割を果たしてきたデータセットに CoNLL2003 があるが，その固有表現カテゴリ（人名・地名・組織名）は比較的粗い粒度であり，分野やアプリケーションによってはより細かい粒度が求められる。近年では，Ding らによって提案された Few-NERD のように，政治家や作家など細粒なカテゴリに分けたデータセットも登場している。今後も使用目的に応じて多様な粒度・種類の固有表現カテゴリが求められることは自明である。

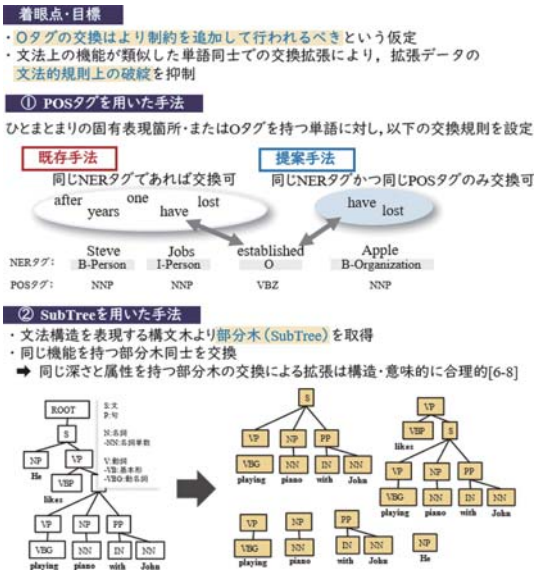
このようなデータ準備のコストを削減する方法の一つにデータ拡張がある。データ拡張は，小規模な学習データに対して加工や合成などの処理を施し，新たなデータを生成して学習に用いる技術である。固有表現抽出タスクでは，単純なルールに基づく交換・挿入・削除といった手法が主流であり，これは系列ラベリングにおいて単語の並びの構造的整合性を保持する必要があるためである。本手法の目的は，既存手法のルールベースを発展させ，固有表現抽出のデータ拡張において新たな置換可能条件を明らかにすることである。

［研究の内容，成果］



既存手法の交換対象は同種のトークン・単語境界範囲内か，同一文中に出現する他のセグメントに限定される。本研究では NLP で伝統的に用いられる構文木と格フレームを手掛かりと

し，より広範囲の境界線を探索する。前者の構文木は部分木ごとに文構造上のまとまりを取得でき，後者は固有表現と関連する規則性とみなされている。



構文木の検証では子ノードを二つ以上もつような部分木を交換対象とする。図の例では計3パタンの部分木が得られる。元データ全体の整合性を担保するため，そこから得られた用法の近い部分木，あるいは名詞同士を置換候補とした。これにより，固有表現箇所以外の交換についても実現が可能となる。上記の両手法について，既存の置換手法と置換候補の幅，データ品質を比較し検証する。

我々の実験において比較に採用する既存手法および提案手法の略字表記と簡単な説明を以下に記す。

MR：既存手法。固有表現を同種固有表現と交換

MR＋POS：MR において，交換条件に POS タグが同じことを追加

OLwTR：O タグを持つ個々の単語同士を交換

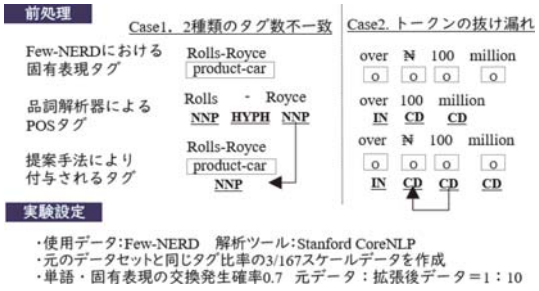
OLwTR＋POS：において交換条件に POS タグが同じことを追加

MR＋OLwTR：MR と OLwRT を同時に適用

MR＋OLwTR＋POS：MR＋OLwTR において交換条件に POS タグが同じことを追加

TreePos：交換条件に TreePos の深さ属性を追加

なお，解析器の違いなどによるセグメントの差に関する取扱い，使用データ，使用ツール，データ拡張時の実験設定については，以下の図に記す。



実験の結果，交換に POS タグを考慮した手法群では，もともとのデータセットにおいて，複数のトークンから成立する固有表現が多く含まれるカテゴリの Recall が向上した。使用したデータセットでは，主に紛争名，災害名などのカテゴリがこれにあたる。一方で，O タグにあたるトークンを変更した手法では，固有表現・一般名詞両方の用途がある単語に対する正答率が低下した。

具体的な例としては，「民主党」あるいは「民主的な（形容詞）」」の両方の意味を持つ Democratic という固有名詞がこれに該当する。一般名詞として使用される場合には，特定の文脈のみに出現していたが，ルールを設定したランダムな交換によりその規則性が無くなり，文脈情報が失われたと考えられる。こうした傾向は，とりわけ MR＋OlwRT＋POS を用いた場合に，Democratic のような複数の意味を持つ単語を含んだ固有表現箇所で頻繁に発生した。以下の表は，本研究の結果，得られた成果である。

提案手法は，ごく少数のデータを使用するようなカテゴリの分類では，既存手法よりわずかに精度が向上した。一方，オリジナルのデータが数十件程度確保できる場合には，既存手法によるデータ拡張のほうがより高精度であった。

method	proposed	macro-AC	macro-F1	macro-P	macro-R
Original-small		0.915	0.626	0.604	0.650
MR		0.902	0.562	0.528	0.601
MR+POS	✔	0.902	0.562	0.527	0.602
OLwTR		0.902	0.570	0.545	0.597
OLwTR+POS	✔	0.903	0.578	0.552	0.606
MR+OLwTR		0.901	0.560	0.529	0.594
MR+OLwTR+POS	✔	0.902	0.569	0.540	0.603
SubTree	✔	0.900	0.536	0.540	0.593

また，100 件程度の正解データを確保できる場合，提案手法・既存手法ともにデータ拡張はかえって精度の低下につながるということが明らかになった。このことから，生成系モデルなどである程度のデータ量が確保できる一般的なドメインの固有表現抽出タスクでは，現時点ではデータ拡張そのものがあまり現実的ではない可能性がある。一方で提案手法は，歴史的な書籍など少数のデータの利活用が必要なドメインにおいて有効である可能性が示唆される。

［成果の発表，論文など］

寺本優香，駒水孝裕，波多野賢治：“固有表現タグおよび POS タグによる交換制約付きデータ拡張手法”，第 15 回データ工学と情報マネジメントに関するフォーラム予稿集，No. 1b-6-4, March 2023.（学生プレゼンテーション賞 受賞）

寺本優香，駒水孝裕，波多野賢治：“POS タグおよび構文木の SubTree を用いた NER 向けデータ拡張手法”，東海関西データベースワークショップ 2023