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

寺本　優香†　　駒水　孝裕††　　波多野賢治†

† 同志社大学　〒 610–0394 京都府京田辺市多々羅都谷 1-3
†† 名古屋大学　〒 464-8601 愛知県名古屋市千種区不老町
E-mail: †teramoto@mil.doshisha.ac.jp, ††taka-coma@acm.org, †††khatano@mail.doshisha.ac.jp

あらまし　固有表現抽出（以下 NER）のタスクにおいて，深層学習モデルの有効性が示されている．大規模データの準備に，学習データを拡充するデータ拡張が用いられてきた．ルールに基づくデータ拡張は有効な解決方法である一方，系列ラベリング問題に分類されるタスクにおいては適用可能な手法が制限される．NER タグに基づき固有表現箇所を交換する既存拡張手法は，一定の効果を示した．しかし，データの大部分を占める非固有表現の NER タグ箇所に対しては，同様の交換により非文が発生する．本研究では，NER タグに加え POS タグに基づく固有表現ラベル交換規則の厳格化と，POS タグに基づく非固有表現の NER タグ箇所の交換を提案し，その効果を検証した．

キーワード　自然言語処理応用，情報抽出，固有表現抽出，データ拡張，POS タグ，品詞情報

1 はじめに

固有表現抽出とは文中から人名や地名，組織名，日付表現，薬品名あるいは遺伝子名といった，特定の属性を持つ一つまたは複数の単語を抜き出す自然言語処理分野の基礎技術である．固有表現抽出はテキスト解析の基礎的な技術であり，様々な分野で利用されている．例えば，医療分野・創薬分野における新たな病名や薬品名の検知および情報抽出 [1,2] や，顧客対応を自動化するチャットボットシステム [3] に利用されている．固有表現抽出タスクの分野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，昨今では深層学習モデルが優れた性能を示している．

このような機械学習モデルの学習には，大規模な学習データが求められる [4] が，大規模で高品質な学習データの確保は容易ではない．その主な理由は人的コストである．特に，分野固有の固有表現抽出タスクを教師あり学習モデルで実現するためには，その分野に精通したアノテータが学習データを作る必要がある．しかし，そのような人材は希少であり，そうした人材を十分に集め，大量の学習データを用意することは容易ではない．

大規模な学習データの例として固有表現抽出タスクの中で中心的な役割を果たしてきたデータセットが CoNLL 2003 [5] である．CoNLL 2003 の固有表現カテゴリは人名や地名，組織名といった粒度であり，あらゆるタスクにおいてこの固有表現カテゴリが広く利用されてきた．当初このデータセットは挑戦的で，困難なものであり，昨今の深層学習技術の発展により，固有表現抽出の性能限界を示したように思われていた．

一方で，固有表現抽出は様々な分野で異なる利用方法をされることがあり，固有表現の粒度も分野やアプリケーションによって大きく異なるという事実もある [6]．このこともデータ準備の困難さに拍車をかけている．

特に，CoNLL が想定する固有表現（人物や場所など）よりも更に細かい粒度が求められることは珍しくない．昨今提案されたデータセットとして，Few-NERD がある [7]．Ding らは，固有表現の多くは既存のデータセットよりさらに細粒な固有表現カテゴリに属している点を指摘し，細かな粒度を持つ Few-NERD データセットを提案している [7]．Few-NERD データセットでは人名の固有表現を表すタグを，例えば政治家と作家のようにさらに細かく分類している．このように今後も使用者の目的に応じて粒度・種類が様々な固有表現カテゴリが出現することは自明である [6]．

こうした学習データ準備のコストを削減するための方法として，少数データを元手に効率よくモデルを構築するアプローチ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部分的な加工や合成，モデルへの入力といった処理を施す．そうして新たに得られた異なるデータを追加することで，大規模な学習データを得る技術である．このとき，新たに得られたデータははじめに与えられた学習データをもとに作成されていることに注意する．固有表現抽出タスクにおけるデータ拡張は単純なルールに基づく交換・挿入・削除による手法が主流である [8]．これは，固有表現抽出タスクのような系列ラベリング問題では図 1 の例に示されるように単語の並びが持つ構造的な整合性を保持するためである．

先行研究 [8] においては，共通の固有表現タグに属する語同士の交換を非固有表現，すなわち文中における固有表現以外全ての箇所を意味する固有表現タグに対しても適応している．非固有表現に所属する単語は固有表現のものよりも多く，交換を行うことによりデータ拡張の結果図 1 の例に示すような非文が生成されることがある．

本研究での目的は，こうした非文を抑制する際，どのような制約が有効であるのかを基礎分析と観察により明らかにすることである．そのため，固有表現タグに加えて品詞情報を含む POS タグの種類により交換に制約を加え，非固有表現を加工するデータ拡張を実装し，その効果について検証する．

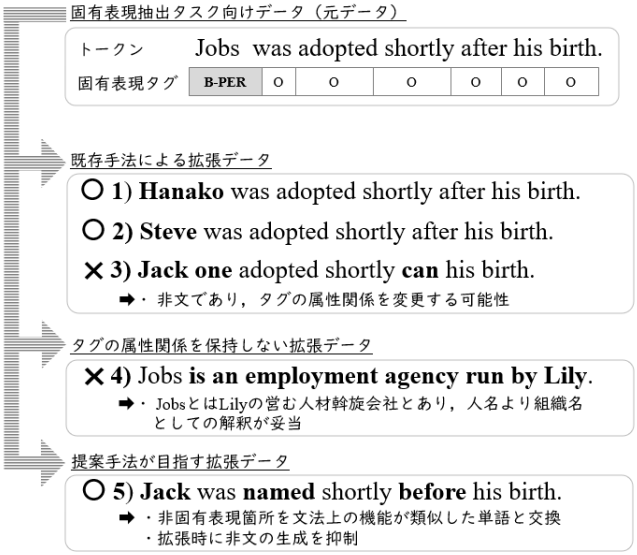

図 1　提案手法を用いた固有表現抽出タスク向けデータ拡張

2　基 本 事 項

本節では，本稿で用いる用語について解説する．

2.1　固有表現抽出タスク

固有表現抽出は，トークンの列で構成されるある文 $s = (w_1, w_2, \ldots, w_n)$ が与えられたとき，各トークンに対応する固有表現タグの列 $t = (t_1, t_2, \ldots, t_n)$ を予測するタスクである．固有表現抽出モデルは，文と固有表現タグの列のペアから予測モデルを学習する．図 1 に，このペアの例を示す．この例では，“Jobs was adopted shortly after his birth.” という文を空白区切りでトークン化した例である．それぞれのトークンに対応する “B-PER” や “O” が固有表現タグである．

固有表現タグを記述する際の規則には IOB や IOE，IOBE，IOBES などが存在する [9]．タグの具体例として上に挙げた形式では O (Outside: 固有表現の外部)，B (Begin：固有表現の冒頭)，I (Insite：固有表現の内部)，S (Single：1 個の単語からなる固有表現)，E (End：固有表現の最後) が主に定義されている．こうした固有表現タグの表記は図 1 に示す B-PER のように，主に固有表現のカテゴリを表す文字列と組み合わせて使用される [5]．また図 2 の organization-government のように，カテゴリ名そのものがタグとして定義されるデータセット [7] も存在する．いずれの場合も固有表現の箇所以外は O タグである．

本稿では固有表現タグとして使用されるタグを機能ごとに大きく 2 種類に分類する．

NE タグ　固有表現箇所の単語であることを表すタグは，どの表記方式を選択するかにより異なる．本稿では固有表現の箇所に振られるタグ全般を指して NE タグと呼称する．NE タグには，B，I，S，E で表されるタグや，具体的な固有表現のカテゴリ名を含む図 1 の B-PER，図 2 の organization-government といったタグが含まれる．

¬NE タグ　固有表現箇所の単語でないことを表すタグを，本稿では ¬NE タグと呼称する．多くの場合 ¬NE タグは O タグで表現される．

2.2　固有表現を含む文の構造

既存手法および提案手法においては，固有表現を含む文をいくつかの構造に分割して扱う．その際の用語についての解説を行う．以下では次の例を用いて，構造について説明する．ある文 s の各単語に対応する固有表現タグ $t = (t_1, t_2, t_3, \ldots, t_{10})$ に三つの固有表現 e_1, e_2, e_3 が含まれており，ぞれぞれ $e_1 = (t_1, t_2, t_3)$, $e_2 = (t_4)$, $e_3 = (t_9, t_{10})$ のような単体または複数の単語で表現する．

セグメント　Dai ら [8] の提案する手法では，同じ種類の固有表現タグを持つ連続した単語をセグメントと定義している．すなわち文 s は四つのセグメント (t_1, t_2, t_3), (t_4), (t_5, t_6, t_7, t_8), (t_9, t_{10}) からなる．図 1 の was adopted shortly after his birth や，図 2 の call to boycott および U.S. Government もそれぞれ一つのセグメントに該当する．

メンション　Dai らの提案する手法では，NE タグが付与された一つまたは複数の単語からなる固有表現を一つのメンションと定義している [8]．メンションの意味は固有表現と同等である．すなわち (t_1, t_2, t_3), (t_4), (t_9, t_{10}) はそれぞれ一つのメンションにあたる．図 1 において元データの例として示す Jobs や，図 2 の U.S. Government もそれぞれ一つのメンションに該当する．このように固有表現が他の固有表現を内包するケースを考えないとき，メンションはセグメントのより狭い概念であり，NE タグによるセグメントとも言い換えられる．

POS タグ　品詞解析の結果得られる品詞情報は POS(Part-of-Speech) タグにより管理される．例えば，自然言語処理用ライブラリである Stanford CoreNLP[1] を用いて品詞解析を行った場合，一文が品詞単位の単語に分けられ，各単語に NN（単数形または不可算の名詞），NNS（複数形の名詞），JJ（形容詞）といったタグが付与される．

3　関 連 研 究

本節では，最初に自然言語処理分野におけるデータ拡張手法の分類について解説を行う．またこのうち単純ルールベースに基づくデータ拡張手法を固有表現抽出タスクに応用した既存手法について述べる．

3.1　自然言語処理におけるデータ拡張手法の分類

Feng らは，データ拡張は画像処理分野において特に頻繁に用いられる手法であり，離散的なテキストデータに対し効果的な拡張を行うことは困難であるとしている [6]．Feng らは自然言語処理分野を対象とした手法も含むデータ拡張手法全般を俯瞰し，以下の 3 種類に分類している．

1：“CoreNLP” https://stanfordnlp.github.io/CoreNLP/ (最終閲覧日 2023/1/8)

a） ルールに基づく手法

ルールに基づく手法とは，事前に定められた規則に従い，データを変換し，新たなデータを生成する手法である．自然言語処理分野では，同義語あるいは同属性固有表現との置換．あるいは同一の文や文書中に出現する語および句に対するランダムな交換・削除・挿入といった処理によるシンプルなデータ拡張が有効とされる．また構文木から得られた構造をもとに交換や削除を行う手法も存在する．固有表現抽出タスクのような系列ラベリング問題においては，元のラベルシーケンスが保持する意味合いを保存しつつデータ拡張を行う場合にルールベース手法が広く用いられる．挿入・削除についてはデータ拡張時に元データに対しどの程度加工を加えるかの割合をパラメタとし，元の文に対する改変を一定の割合以下になるよう制御している．こうした手法はいずれも少量のデータを用いた学習においてモデルの性能を向上させることが報告されている．

b） Mixup による補完

複数の元データを合成するデータ拡張手法は古典的である．その中でも Mixup による補完では Beta 分布により混合比率を決定したうえで，複数の学習データだけでなくラベルの性質を合成するデータ拡張を行う．例えば画像分野における Mixup では，二種類の画像を透過処理したうえで重ね合わせ，両者の属性を混合された割合で併せ持つ新たなタグを付与する [10]．自然言語処理分野においては，単純にテキストデータやタグを混合させるだけでなく，言語モデルから得られた埋込みをもとに合成を行う手法も存在する [11].

c） モデルに基づく手法

モデルに基づく手法とは，学習済みのモデルを利用した文生成・パラフレーズ生成によるデータ拡張のことである [6]．自然言語処理分野の主な手法には RNN による文生成，seq2seq や翻訳機によるパラフレーズ生成が存在する．

3.2 固有表現抽出タスクにおけるルールベースのデータ拡張

自然言語処理分野の中でも単語単位のラベルを保持する固有表現抽出タスクの場合，ラベルの整合性を保持したままデータを加工することは難しく，Mixup による補完に該当する手法 [11, 12] が存在するものの，データ拡張は単純なルールベース手法が主流となっている．

Dai らは自然言語処理分野において用いられるルールベースのデータ拡張についてまとめたうえで，固有表現抽出分野の 2 種類のデータセットに対しそれらを適用した手法を提案している [8]．Dai らの提案するデータ拡張手法を次に示す．

a） Label-wise token replacement（LwTR）

NE タグ・¬NE タグの別に関わらず，個々の単語に対し置換えるかを二項分布でランダムに決定し，置換えによって生成された新たな文をデータに加えることで拡張を行う．置換え候補としては学習データ内に出現し，かつ同じ種類の固有表現タグを持つ別の単一単語を無作為に選択する．元データと新たに生成されたデータではタグのシーケンスが共通している点が特徴である．

b） Synonym replacement（SR）

単語を WordNet から検索された同義語と置換えデータ拡張を行う．拡張により，元データと生成後データで単語数が変化する場合が存在する．

c） Mention replacement（MR）

メンションに対して置換えを実行するかを二項分布でランダムに決定し拡張を行う．学習データ内で同じラベルを保持する他の固有表現箇所を無作為に置換え候補として選択する．

d） Shuffle within segments（SiS）

単語列を同じラベル集合ごとのセグメントに分割したうえで，ランダムにセグメント内に含まれる単語の順番をシャッフルし拡張を行う．このセグメントの境界は，構文解析により得られる句および節の階層関係に基づき決定される．またこの時シャッフルは二項分布に基づき実行するかどうかを決定する．なおこの手法ではラベルの順序は変更しない．

これらの手法は学習データが少数である場合に有効な拡張手法として提案されており，少数データを再現するため，元の学習データを前から順にそれぞれ 50 文，150 文，500 文取得した 3 種類のデータを準備し，提案手法ごとにデータを拡張している．拡張におけるパラメタは n と p の 2 種類で，グリッドサーチにより最良の精度を得られるパラメタを決定する．n は 1 文から新たに拡張する文の数であり，1, 2, 3, 6, 10 から手法の特性に応じて値の候補を選択したうえで探索を行う．p は交換の実行を決定する二項分布のパラメタであり 0.1, 0.3, 0.5, 0.7 からいずれかの値を取る．

3.3 本研究の位置づけ

Dai らがまとめたルールに基づくデータ拡張手法 [8] のうち，LwTR と MR の 2 手法は，同一のデータ内を交換候補取得の範囲とし，各トークンに付与された固有表現タグの種類によって交換を制御している．本研究で取り扱う内容は，ルールに基づくデータ拡張手法の中でも，固有表現タグに基づく交換に関するものであり，LwTR と MR をベースに発展させたものである．本研究では，NE タグに比べ多くの種類の単語が属する ¬NE タグの交換時に非文が多く発生する点に着目し，品詞情報により交換を制約することで非文発生抑制を試みる．

4 提 案 手 法

本節では，提案手法について説明する．はじめに，本提案手法の着想と設定する仮説について述べ，手法の概要について説明する．次に，用いるデータやツール，前処理の詳細を説明する．最後に，モデルの学習にあたり設定した各種パラメタについて述べる．

4.1 本研究の特色

Feng らは，モデルの精度向上を目的とした拡張されたデータは，元データと一致も乖離もしない状態が望ましいとしている [6]．すなわち，新しく生成されたデータが持つ特徴は，元データ群の分布と比較したとき，元データ群に含まれるあるデータと完全に一致せず，また外れ値でない状態が望ましい．

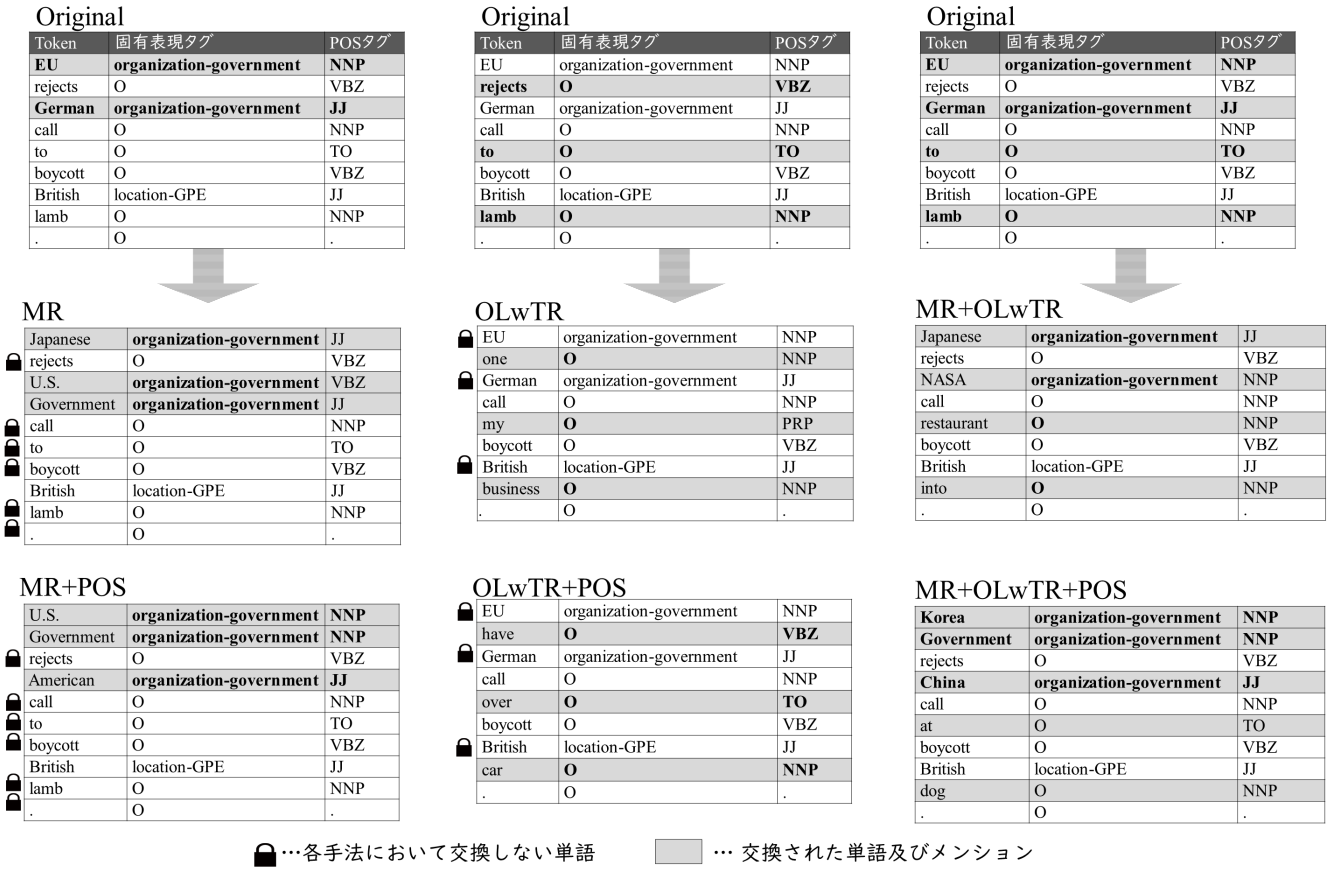

図 2　各データ拡張手法により生成されるデータの例

これについて，本研究では人間が見て不自然と判断するような非文を元データから乖離していると判断する．これに基づき，本研究では，非文を生成しないような制限を加えたデータ拡張手法を設計する．非固有表現はデータの大部分を占めるため，図 1 の 5) に示すようなタグの属性関係破綻を避けたデータ拡張手法を確立することで，より多くの学習データのバリエーションを獲得することが可能となる．

以上を踏まえ本研究では，各単語が持つ様々な文法情報を交換時の制約とすることで文法的な破綻の少ない学習データが生成可能であり，そうした拡張手法の改善がモデルの精度向上に貢献するという仮説を立てる．本研究では，¬NE タグの交換に非文が発生しやすいことに着目し，それらを抑制するために固有表現タグと品詞情報である POS タグの両方を交換制約とした拡張手法を提案する．どのような制約が有効であるのかを基礎分析と観察により明らかにすることがその目的である．

4.2　提案手法の概要

本研究では非文の発生を抑制することを目的としているため，メンション部分の交換に対しては LwTR のようなメンションを単位に分割するような置換えは行わず，MR のみを採用する．また，NE タグと ¬NE タグの性質の違いを観察するため，それぞれの交換のみを行った実験と，両方同時に行った実験を分けて行う．さらに NE タグを持つメンションの交換においても品詞情報が有効である可能性が否定できない．従って MR に対しても固有表現タグに加え POS タグを考慮した提案手法のデータ拡張を実装する．以上を踏まえ本研究では次の 5 手法を提案し，既存手法である MR との比較を行う．

MR+POS　既存手法 [8] である MR のメンション交換において，固有表現タグに加えて POS タグを制約として追加した手法．同じ固有表現カテゴリに属するメンションであっても，POS タグが同じ場合のみ交換が可能となる．

OLwTR　既存手法 [8] である LwTR を ¬NE タグである O タグに適用した手法．個々の ¬NE タグ同士の交換のさい，固有表現タグのみを制約としている．NE タグについては交換を実施しない．

OLwTR+POS　OLwTR における個々の ¬NE タグ同士の交換の際，固有表現タグに加えて POS タグを制約として追加した手法．NE タグについては交換を実施しない．

MR+OLwTR　NE タグに対しては MR，¬NE タグに対しては OLwTR を適用した手法．いずれの交換においても固有表現タグのみを制約としている．

MR+OLwTR+POS　MR+OLwTR における交換のさい，固有表現タグに加えて POS タグを制約として追加した手法．

4.3　使用データと前処理

本研究では固有表現カテゴリに属するデータが少数である場合に，POS タグを使う有用性を実験的に確認する．従って提案手法はあらゆるデータに汎用的に用いることが可能であるが，

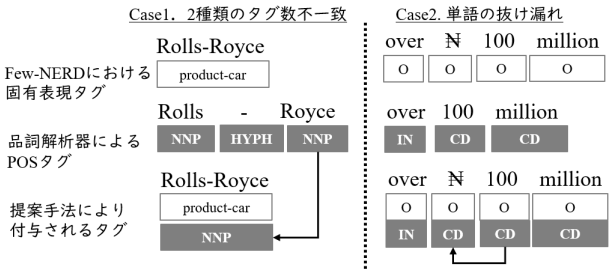

図 3　POS タグと固有表現タグの統合方法

今回は特にラベル粒度が詳細であり，少数データのみからなる固有表現カテゴリを含む Few-NERD データセット [7] を使用する．

本稿では，提案手法をすべての固有表現カテゴリのデータが少なくともひとつは存在するデータセットを対象とする．Few-NERD において最もデータ数が少ない固有表現カテゴリは *art_painting* であり，このタグを含む文の数は 167 しか存在しない．従って *art_painting* の固有表現タグを含むためには，全データ数の 1/167 スケールが最小のデータセットサイズとなる．本実験では，*art_painting* を必ず含み，各固有表現カテゴリの相対出現頻度が，全データ数のときと変わらないように，データセットを構築する．具体的には，作成するデータセットのスケール τ に対して，各カテゴリ c について，$N_c \times \tau$ 個の文をサンプリングする．ただし，N_c は，カテゴリ c の固有表現を含む文の数である．各固有表現カテゴリを含む文の数が過小にならないように，今回は 3/167 スケールで実験を行った．

POS タグの付与には Stanford CoreNLP [13] を用いる．ただし，Stanford CoreNLP により提供される POS タグの境界と，固有表現抽出向けのデータに付与された固有表現タグの境界が異なる場合が存在する．両タグともに，付与されている単語に対する交換の制約を決定する手掛かりであるため，タグの境界を統一する必要がある．本研究では Few-NERD が提供する固有表現タグによる単語の境界を保持し，POS タグをそれに合わせて変換する．変換の方法について図 3 に示す．

該当箇所において文字列は同じであるが，POS タグが固有表現タグよりも細粒となるケース（図 3：Case1）が valid データを含む学習用データ 150,589 文のうち 30,674 件で見られた．この問題への対処として複数付与された POS タグのうち最後のものを全体の POS タグとして採用している．また 150,589 中 8 件のみであるが，固有表現タグが POS タグよりも少なくなるケースが見られた．この原因は，a) 一つのメンション範囲に複数の POS タグが含まれる，b) 特殊文字の単語が抜け落ちている，のいずれかであった．a) に対しては，POS タグを固有表現タグに合わせて分割することで対応している．一方，b) については図 3 の Case2 に示す通り，抜け落ちた単語を補填したうえで直後の POS タグを新たに付与している．なお POS タグ・固有表現タグともにタグの境界線はいずれも単語を区切る空白の上にあったため，あるタグの境界が他のタグの範囲内に含まれるケースについては今回は想定していない．

品詞解析により分数を表す特殊記号が英数字に変換されるといった記号処理が行われ文字の表層が変化する場合が存在したが，固有表現抽出タスク向けデータとして扱う際は Few-NERD 上での表記に統一している．

4.4　少数学習データの再現と各種パラメタ

先行研究に倣い，データ拡張におけるパラメタとして，1 文から生成する文の最大数 n と，二項分布のパラメタ p を設定する．提案手法では，それぞれ $n = 10, p = 0.7$ とした．この値は，先行研究において設定されていたパラメタ n, p のいくつかの値のうち最大のものである．最大の値を設定した理由は，¬NE タグがデータセット全体における固有表現タグの 97.67% を占めるため，ルールベースでの交換を適用する範囲が非常に大きいためである．NE タグを対象とした交換では，出現数の低く，頻繁に交換を行っても同じ文を生成してしまう可能性が高い．それゆえ提案手法では，¬NE タグを積極的に交換するために，$p = 0.7$ として，交換確率を大きくとり，生成する文の数も $n = 10$ と多めに生成する様にしている．

また，本提案手法ではパラメタとして学習時の epoch 数は $e = 3$ とする．また，MR と MR+POS，OLwRT と OLwRT+POS，MR+OLwRT と MR+OLwRT+POS は図 2 に示した通り POS タグを条件として付与した場合としなかった場合の対応関係にあたる．このため，これらの組に関しては交換を発生させる箇所を統一することで，条件による結果の差を検証できるようにしている．

4.5　評 価 方 法

評価実験の設計について解説を行う．本実験で用いるデータは 4.3 に示す理由により 3/167 データスケールである．今回使用するデータセットでは，固有表現タグのカテゴリごとに含まれるデータ数に偏りがある．この偏りを考慮しなかった場合，データ数の大きいカテゴリの結果が全体に影響を及ぼす恐れがあるため，本研究ではカテゴリごとのデータ数に影響を受けない macro-F 値，macro-Precision，macro-Recall を用いる．これは，各カテゴリの F 値，Precision，Recall を算出したのち平均をとった評価指標であり，カテゴリごとに含まれるデータ数の偏りに影響を受けない．

また，MR と MR+POS，OLwRT と OLwRT+POS，MR+OLwRT と MR+OLwRT+POS は図 2 に示した通り POS タグを条件として付与した場合としなかった場合の対応関係にあたる．このため，これらの組に関しては交換を発生させる箇所を統一する．これにより条件による結果の差を検証できるようにしている．

5　結果と考察

先行研究および提案手法により拡張したデータをそれぞれ学習データとしてモデルを構築し，その精度を比較することでデータ拡張手法の評価を行う．

5.1　結　　果

表 1 は，本研究における評価実験の結果である．MR と MR-POS の差は，micro-F 値，micro-Precision，micro-Recall で

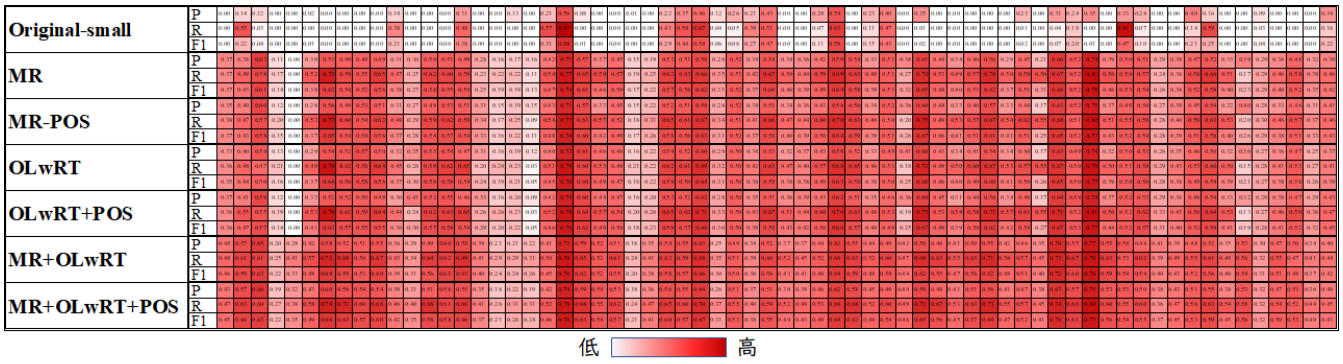

図 4　固有表現タグごとの P,R,F1 ヒートマップ

それぞれ-0.001,0,0 であった．一方，OLwRT と OLwRT+POS では 0.008，0.023，0.015 であった．また MR+OLwRT および MR+OLwRT+POS では 0.006，0.011，0.008 となった．

5.2　考　　察

本節では，少数データへの影響，POS タグによる制約の有効性，O タグ箇所交換における POS タグ制約の有効性についてそれぞれ

少数データへの影響

少数スケール群においてデータ拡張の有無を比較したとき，Original-small について，学習データ数が 14 件未満の固有表現タグについては，正答率が 0 となっていた．少数データの学習における精度が極端に低いことが，Original-small 全体の精度が低い原因となっている．このことから，関連研究および本提案手法は，ゼロショットあるいは few ショットに近いごく少数のデータのみが存在する場合に有効な拡張手法であることが分かる．逆に，ある一定以上のデータが集まった場合，データ拡張を行わなかったデータセットに対する固有表現抽出器のほうが，データ拡張により得られたデータセットに対する固有表現抽出器に比べて，精度が高い傾向にあった．

POS タグによる制約の有効性

POS タグを条件として加えた場合と加えなかった場合で比較をした場合，MR 以外の全ての手法において僅かに性能が上昇している．また，図 4 に示す通り，Original-small，MR，MR+POS，OLwRT，OLwRT+POS でも性能が上がりにくい一部のタグについて，本研究の提案手法である MR+OLwRT および MR+OLwRT+POS で精度が向上する傾向が見られた．

POS タグを条件として追加した考慮したモデルでは，それ以外の条件が対応している POS タグを考慮しないモデルに比べ，固有表現を複数単語と予測する傾向が見られた．紛争名，災害名といった固有表現が複数の単語を含みやすいタグにおいて Recall の値が向上している原因となっている．

O タグ箇所交換における POS タグ制約の有効性

¬NE タグに対する交換を行う際，POS タグの追加により交換に制約が与えられたモデータを用いたモデルの精度が高い．一方，精度向上の度合いは最大でも 0.013 でり，品詞情報を含む規則性は言語モデルにとってそれほど意外性がある規則性ではないことが示唆される．¬NE タグに対する交換拡張を行った後学習したモデルでは，固有表現としても一般名詞としても使用される単語からなる固有表現に対する正答率が低下している傾向が見られた．例えば Democratic (「民主党」あるいは「民主的な [形]」) は O タグの交換をしないモデルが正答する傾向にあった．

表 1　3/167 スケールデータの結果

method	proposed	macro-F1	macro-P	macro-R
Original-small		0.125	0.134	0.104
MR		**0.412**	0.493	0.444
MR+POS	✓	0.411	0.493	0.444
OLwRT		0.410	0.478	0.435
OLwRT+POS	✓	**0.418**	**0.501**	**0.450**
MR+OLwRT		0.461	0.537	0.495
MR+OLwRT+POS	✓	**0.467**	**0.548**	**0.503**

固有表現タグの性質と分類精度

さらに，MISC や other を含む固有表現タグについて，Original-small 以外の手法ごとの精度に大きな差が見られなかった．このタグには，他の固有表現タグに該当しない固有表現がすべて含まれてしまうという傾向がある．例えば，固有表現タグが person-musician，person-other の 2 種類であった場合，政治家の氏名である John F Kennedy は person-other に該当するが，person-politician という固有表現タグの定義が存在した場合は，person-other よりも person-politician のほうがふさわしい．このように含まれる固有表現の種類が，各タスク・ドメインにおける固有表現タグの設計に依存する．こうした固有表現タグの分類精度向上には，本研究で提案したものとは別のアプローチが必要であると考えられる．

6　おわりに

本研究では，固有表現抽出タスクのデータ拡張において，既存手法を発展させるべく O タグの拡張を組み合わせた方法を提案した．その際，拡張時の交換に固有表現タグに加えて POS タグによる制限を新たに設け，文法的に破綻した文が生成されにくくすることで，データ拡張により得られるデータの品質向

上を目指した．データ拡張の結果，交換箇所の品詞がほぼ同じである MR 以外の全ての手法において提案手法の精度が向上した．このことは，大規模な言語モデルによるアプローチにおいても，データ拡張に基礎的な文法情報を考慮した制約を課すことが有効であることを示唆している．そのような文法規則は他にも存在する可能性がある．以上より今後は品詞情報以外の文法情報を活用したデータ拡張手法を検討する．

謝　　辞

本研究の一部は，JSPS 科研費 JP21H03555 の助成を受けたものである．ここに記して謝意を示す．

文　　献

[1] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, Vol. 7, pp. 73729–73740, 2019.

[2] Nasi Jofche, Kostadin Mishev, Riste Stojanov, Milos Jovanovik, Eftim Zdravevski, and Dimitar Trajanov. Named entity recognition and knowledge extraction from pharmaceutical texts using transfer learning. *Procedia Computer Science*, Vol. 203, pp. 721–726, 2022.

[3] Pēteris Paikens, Artūrs Znotiņš, and Guntis Bārzdiņš. Human-in-the-loop conversation agent for customer service. In *Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings 25*, pp. 277–284. Springer, 2020.

[4] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics: COLING2018*, pp. 2145–2158, 2018.

[5] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. 2003.

[6] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988, 2021.

[7] Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. Few-nerd: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing:ACL-IJCNLP 2021*, Vol. 1, pp. 3198–3213, 2021.

[8] 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:COLING 2020*, pp. 3861–3867, 2020.

[9] Archana Goyal, Vishal Gupta, and Manish Kumar. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, Vol. 29, pp. 21–43, 2018.

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization.

[11] Rongzhi Zhang, Yue Yu, and Chao Zhang. Seqmix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:EMNLP 2020*, pp. 8566–8579, 2020.

[12] Rongzhi Zhang, Yue Yu, and Chao Zhang. Seqmix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:EMNLP 2020*, pp. 8566–8579, 2020.

[13] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, 2014.