系列ラベリングデータにおける CutMIX によるデータ拡張

田村 光太郎
株式会社ユーザベース
koutarou.tamura@uzabase.com
k.tamura.phd@gmail.com

概要

固有表現抽出モデルを構築するにあたり，訓練データとして，固有表現の位置をアノテーションする系列ラベリングデータが必要となる．しかし，データ構築の際のアノテーションでは，テキストの精読が必要となり，継続的に高品質な教師データを作成することが難しい．そのため，少量のデータを利用した効率的な学習として，テキストに対する拡張手法を適用することを試みた．ここではニューステキストにおける情報抽出タスクを題材として，画像データで利用される Mix-based の手法を適用しデータ拡張を行った．これらの拡張手法におけるデータ量やその強度での精度変化を調べ, モデルの精度向上を行った.

1 はじめに

固有表現抽出モデルの構築では，系列ラベリングされたデータが必要となる．しかし，系列ラベリングのアノテーションでは，対象となるテキストの精読が必要であり，質の良い大規模なデータを構築することに，非常に大きなコストがかかる．そのため，モデル構築の初期では少量データでの効率のよい学習が求められ，その一つの方法としてデータ拡張がある．

テキストデータにおけるデータ拡張の方法は，ルールベースの置換操作 [1, 2, 3, 4]，データを利用した合成 [5, 6, 7, 8, 9]，モデルによる生成 [10] の 3 つに分類される．その一つであるルールベースでの手法は，文章の一部を入れ替えることで意味情報にノイズを付加する基礎的な方法として主流となっている．これらは，少量データにおいて精度が有意に改善することが報告 [1] され，その有効性も認識されている．

最近では，画像認識タスクで利用される，Mix-based の拡張手法 [11] が，自然言語処理の領域でも応用 [5, 6, 7, 8, 9] されている．基本的な操作として，画像データにおける利用に類似して，2 つのテキストデータを一定の比率で混ぜ合わせ，教師データにはないデータを合成する方法である．主には，文書分類タスクにおいて利用されている．テキストデータにおいては，データにある 2 つの文章をある比率で分割し，つなぎ合わせる方法もあり，精度改善したことが報告 [8, 9] されている．しかし，文章を連結して拡張する方法はデータの半分は元データで構成されるため過学習の発生が懸念されること [8] や，逆に，多量のデータを同時に混ぜ合わせると，ラベルの足し合わせで過度な平滑化がおき，データの特徴が薄れてしまうことが懸念される．

しかしながら，固有表現抽出における系列ラベルがついたテキストデータにおいては，ラベル間の足し合わせは生じないため，多数のテキストの連結によるデータ拡張が有効であることが期待される．本研究では，これらの観点に基づきデータ拡張の手法を検討する．さらに，この拡張方法をニュース記事における情報抽出タスクに適用し，拡張の有効性を検証する．

第 1 章では，本研究の背景を述べた．2 章では，関連する先行研究について，3 章では，提案する手法，4 章では利用するデータセットの詳細について述べる．5 章では，有効性を検証するための実験設定を述べ，6 章ではその結果を，最終章では本研究をまとめる．

2 関連研究

2.1 Mix-based によるデータ拡張

Mix-based の手法は，画像データにおけるデータ拡張手法として，いくつかの方法が提案されている．主流な方法として，MixUP[11] と CutMIX[12] があり，それらについて本研究と関連のある部分について簡単に説明する．

MixUP

最も基本的な方法で，Zhang ら [11] により 2017 年に提案された方法である．1 件の拡張データは，既存のデータにある 2 つの画像データ x_1,x_2 とそれに対応するラベル y_1，y_2 から作成される．具体的には，選ばれた 2 つの画像データの数値情報を割合 λ, $1-\lambda$ で混ぜ合わせることで，1 件の拡張データ X,Y を得る．

$$X = \lambda x_1 + (1-\lambda)x_2$$
$$Y = \lambda y_1 + (1-\lambda)y_2$$

ここで，x_i は画像データを表す RGB のカラー行列や埋め込み表現などの数値情報，y_i は分類ラベルの One-hot 表現のベクトルなどである．

自然言語処理においては Guo[5] や Sun ら [6] が適用を試みている．テキストの埋め込み表現における MixUP を行うことで，既存の分類タスクで精度が改善することを示している．また，固有表現抽出においても Wu ら [7] が精度の向上が見込まれることを報告している．一方で，菊田ら [9] は，これら MixUP の方法でデータ拡張を行う際に，文書の埋め込み表現を事前に作る必要があることから，モデルの学習プロセスの順序に課題があることを指摘している．

CutMIX

画像の一部を切り出し，別の画像をあてはめて合成する CutMix[12] と呼ばれるデータ拡張方法がある．テキストデータにおける CutMIX 手法のアナロジーとして，文章の単語やトークン列を分割し，異なる分割同士を連結する方法が提案 [9, 8] されている．具体的には，井上らは，混合比率 λ を Cut rate と読みかえ，データ x_1 の文頭から λ の比率のトークンと，データ x_2 の文末から $1-\lambda$ の比率のトークンを取得し，双方から取得したトークン列を組み合わせることで拡張データを作成する方法を提案している．Cut rate によっては，拡張データの大半が元データと似ることになるため，過学習の発生が懸念されることを言及しているが，分類タスクの精度改善が一定程度期待できる結果を得ている．これは，CutMIX 手法がテキスト系タスクにおけるデータ拡張手法として有効であることを示唆している．

2.2 ニューステキストの情報抽出モデル

本研究では，拡張方法を適用する固有表現抽出タスクの例として，ニュース記事における情報抽出を題材とする．著者ら [13] は，ニュース記事の情報抽出モデルとして，ニュース記事から組織名を固有表現として抽出するモデルを提案している．ニューステキストに IOBE 形式のラベルを付与した教師データを用い，先行研究でベースとされていたモデルである BERT から LUKE[14] に置き換えたものを利用する．ここでの LUKE は，トークナイザーを MeCab に変更し，固有表現抽出タスクで単語境界をより自然な形でとらえるようにし，Wikipedia の 7 月 1 日時点の日本語記事ファイルで事前学習している [15]．この LUKE をベースとして，最終層に CRF 層を接続する構造とした．ここでの CRF 層において，IOBE 形式における禁止遷移に関しては，遷移コストを-100 に設定している．

3 提案手法

テキストデータにおける Mix-based の手法では，埋め込み表現の足し合わせのほかに，文章が連結される形での合成方法が提案されていることに注目する．特に，系列ラベルが付与されている文章の連結では，ラベル間の混合が生じないため，ラベルの足し合わせによる平滑化の懸念を避けられると考え，複数件のデータから拡張する次のような方法を拡張方法として提案する．

ここで，文は句点で区切られたもので，文が集まったものを文章とする．1 つの文章に含まれる文の数の分布 $P(L)$（以下，文数分布）と，全文章における L 文で構成される文章が固有表現を含有する割合 $p_0(L)$ のもと，次の方法で拡張データを作る．

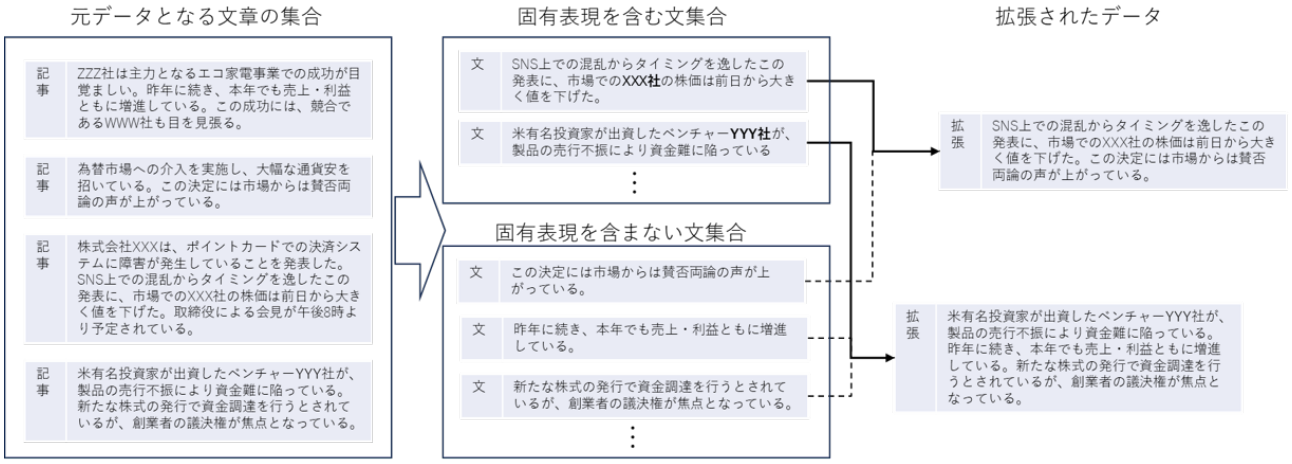

図 1　実データにおけるデータ拡張の方法の手順．（図中のテキストは架空の例）

実験設定

元データの文章の特徴を反映する CutMIX 手法

元データにおける L 文の文章の固有表現の含有割合 $p_0(L)$，記事の文数分布 $P(L)$ を設定する．

- 訓練データにおける文章データの全てのテキストを句点ごと，1 文ずつに分割する．
- 固有表現の含まれる文の集合 M と固有表現の含まれない文の集合 $\bar{M}$ に分ける．
- 文数分布 $P(L)$ に従う変数 L を決める．
- 下記の確率 p_M で集合 M から，$1-p_M$ で集合 $\bar{M}$ から文を選択することを L 回実施し，選ばれた文章を連結する．

$$p_M = 1 - (1 - p_0(L))^{1/L} \tag{1}$$

このアルゴリズムによって，元データにおける 1 文章に含まれる文数の分布 $P(L)$ や L 文の文章の固有表現の出現割合 $p_0(L)$ が保った拡張文が作られることになる．つまり，元データと拡張データの特徴が類似したものになり，拡張データの混在によりモデルの傾向が変わらないことを期待するものである．

4　利用するデータ

本研究で利用するデータは，NewsPicks が取得した 2023 年 1 月〜6 月までのニュース記事から，経済メディアを中心にサンプルした 1991 件の記事データである．記事本文の長さ（文字数）は平均 801 字で，最長 6341 字となる記事もある．このため，アノテーションコストや学習時の特定の記事のテキスト量でバイアスが発生することを避けるため，主要な内容が記述されるタイトルと本文冒頭 10 文までのテキストを利用する．この記事冒頭 10 文までのテキストに対して，系列ラベルを付与したデータを，本研究で利用する固有表現抽出の教師データとする．固有表現抽出のための系列ラベルとして，単語が企業名であるかとその出現位置を付与している．

5　数値実験

元データ 1991 件のデータのうち 500 件を検証データとし，1491 件を学習データとしてファインチューニングを行った．この学習データのデータ量を変え，拡張方式を適用することを 5 セット行っている．また，学習は 5epoch 行い，モデルの入力で 512 トークンをこえる文章については，512 トークンを超えないように文章を当分割している．このデータについて，前章の記述のとおりの図 1 に例示したように拡張を行い，実際のデータに則した $p_0(L)$) と，文数分布 $P(L)$ を利用する．最後に，元データの量に対し，拡張データの比率を表の割合で変えたときの混合データを学習の対象とする．

結果

検証データ 500 件に対して，各条件の精度（5 セットにおける平均）は表 1 のとおりとなった．各組み合わせに対して，固有表現の出現位置を正確に推論できたことをもとに正解とし，F 値を表に記載している．全体的に，少量データのときに精度の改善は大きいが，データ量の増大とともに精度改善の効果は薄れていくことが分かる．また，元データの量に依存して，適切な拡張データの量も変わってくる．特に，元データ量の大きいところでの精度向上の厳密な効果については検討の余地を残している．

表 1 各条件下における数値実験の結果．数値は，F1 で 5 セット行ったものの平均をとっている．

		オリジナルデータのサイズ			
		25%	50%	75%	100%
拡張比率	0%	73.15 ± 0.23	80.07 ± 0.27	80.40 ± 0.22	81.90 ± 0.11
	20%	76.63 ± 0.51	80.10 ± 0.28	**80.77 ± 0.19**	**82.20 ± 0.15**
	50%	77.39 ± 0.63	80.32 ± 0.38	80.70 ± 0.17	82.10 ± 0.13
	100%	76.85 ± 0.74	**80.40 ± 0.36**	80.54 ± 0.22	81.30 ± 0.25
	300%	**77.52** ± 0.92	79.87 ± 0.44	79.99 ± 0.26	80.17 ± 0.26

6 まとめと今後の課題

本研究では，ニュース記事中の組織名を抽出する固有表現抽出タスクに対して，Mix-based の手法をもとに文章を連結する拡張方式を提案し，モデルの精度改善が期待されることを検証した．

系列ラベルにおいては，データの合成による過度な平滑化をさけることができるため，元データに類似したデータの生成をさけることができることが示唆された．さらに，本提案手法では，拡張手法による元のテキストデータからの乖離を抑えることを検討し，元データの特徴を残したまま拡張する方法とした．それにより，精度改善が見込まれることが示唆され，本提案手法が固有表現におけるデータ拡張の方法として検討の余地があることが期待された．

今後の課題として，与えられたデータセットに対し，適切な拡張手法やその強度，比率などを推定するため，テキストデータの構造と拡張手法の関係について調べていく．

参考文献

[1] X. Dai and H. Adel. An analysis of simple data augmentation for named entity recognition. In **Conference: Proceedings of the 28th International Conference on Computational Linguistics**, 2020.

[2] J. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In **EMNLP-IJCNLP 2019 short paper**, 2019.

[3] 大林弘明. データ拡張を用いた固有表現抽出の精度向上. 言語処理学会 第 25 回年次大会 発表論文集, 2019.

[4] 山崎智弘. 系列ラベリングタスクのための単純なデータ水増し. 言語処理学会 第 29 回年次大会 発表論文集, 2023.

[5] et al. H. Guo. Augmenting data with mixup for sentence classification: An empirical study. In **arXiv:1905.08941**, 2019.

[6] et al. L. Sun. Mixup-transformer : Dynamic data augmentation for nlp tasks. In **COLING 2020**, 2020.

[7] et al. L. Wu. Robustself-augmentation for named entity recognition with meta reweighting. In **NAACL 2022**, 2022.

[8] 井上顧基ほか. Mixed-based データ拡張手法による文書評価問題に対する予測精度の向上. 第 37 回人工知能学会全国大会論文集, 2023.

[9] 新納浩幸菊田尚樹. Bert を用いた文書分類タスクへの mix-up の適用. 第 27 回言語処理学会年次大会, 2021.

[10] 田村光太郎. 疑似ニュース生成による 固有表現抽出タスクのデータ拡張. 第 19 回 Web インテリジェンスとインタラクション研究会, 2023.

[11] et al. H. Zhang. mixup: Beyond empirical risk minimization. In **ICLR 2018**, 2018.

[12] et al. S. Yun. Cutmix: Regularization strategy to train strong classifiers with localizable features. In **arXiv:1905.04899**.

[13] 田村光太郎, 北内啓, 高山温. 固有表現抽出によるニューステキスト内の企業名抽出. 人工知能学会全国大会論文集, 2023.

[14] H. Shindo H. Takeda I. Yamada, A. Asai and Y. Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention. pp. 6442–6454. Association for Computational Linguistics, 2020.

[15] （2023-01 閲覧）. https://huggingface.co/uzabase/luke-japanese-wordpiece-base.