ニュース記事テキストにおける組織名の抽出

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

ニュース記事は日々多量に配信され，その情報の瞬時の整理・構造化が求められる．我々は，ビジネスや業界固有のトピックを含むニュース記事データを利用し，テキストに現れる組織名とそれに関わる情報を固有表現として抽出するモデルを構築した．モデルは，エンティティ単位での知識も埋め込まれた LUKE を採用し，単語境界をよりとらえられるようにトークナイザーや知識を更新した．公開データセットを用い，提案モデルによる精度検証を行い，一定程度の精度改善が見られることを確認し，また，実データに適用することで，実データでの情報抽出の精度改善に必要な課題を考察した．

1 はじめに

固有表現抽出は，情報の抽出や構造化を目的とした自然言語処理における基礎的なタスクである．主には，固有表現の出現位置が付与された系列ラベルデータを用い，条件付き確率場（Conditional random field，以下，CRF とする）などの手法を使った解法が提案 [1] されていた．昨今では，BERT[2] などの transformer を使った深層学習モデルでのタスクの取扱いが中心となっている [3, 4, 5]．

このような固有表現抽出の適用先の例として，ニューステキストにおける情報の抽出がある [6, 7, 8, 9]．ニュースが日々高頻度に配信され，機械的な情報の整理・構造化が求められていることから，自然言語処理タスクの対象として研究されている．具体的には，業種や記事内容を分類することや，企業名，日付，数量やイベントなどの特徴となる情報を固有表現として抽出し，構造化することが行われる．

ニュース記事情報を固有表現抽出タスクとして情報抽出を行う研究では高橋，田村などがモデルを提唱している．高橋は，BiLSTM-CRF によりニュース情報からキーワードを抽出 [6] することを行っている．一方，BERT-MRC のモデルでは改善が見られず，系列ラベリングを解くタスクでは CRF の有効性が示唆 [8] されている．田村も同様に BERT に CRF 層を付加することでモデルを構築 [9] している．これらのモデルをベースとして，データ拡張 [10, 11] や能動学習 [12] などが議論されている．

本研究では，固有表現としてニュースの主体となる組織名を高精度に抽出することに注目する．田村らの提案手法におけるモデルを改善し，公開データにより精度評価を行いその有用性を評価する．

1 章では，本研究の背景となるニュース記事データの情報抽出に関連する研究を紹介した．2 章と 3 章では，本研究で利用するデータとモデルの説明をそれぞれ行う．4 章では，モデルを評価するための数値実験の設定について，5 章ではその結果を報告する．最終章では，本研究の成果をまとめ，今後の課題を述べる．

2 データ

ニュース情報に関する固有表現抽出モデルを構築し，評価するにあたって，固有表現に関する公開データセットとニュース記事テキストデータセットの 2 つのデータを利用する．双方とも，組織名やそれに関連する情報が固有表現ラベルとして，テキスト中に系列ラベリング（開始位置，終了位置，固有表現タイプ）として付与されている．また，2 つのデータセットは法人名（企業名）や製品名のラベルを共通に持つため，これらを抽出する際の基礎的なデータセットとして利用する．

公開データセット

ストックマーク株式会社が提供する Wikipedia の日本語固有表現抽出データセット [13] である．Wikipedia の記事データから構築された全 5343 件のテキストデータに対して，固有表現ラベルが付与さ

表 1　利用データにおけるラベルの内訳

公開データ		ニュース記事データ	
ラベル	固有表現数	ラベル	固有表現数
人名	2980	-	-
法人名	2485	組織名	1895
政治的組織名	1180	-	-
その他の組織名	1051	-	-
地名	2157	-	-
施設名	1108	-	-
製品名	1215	製品・サービス名	363
イベント名	1009	-	-
固有表現総数	13185	固有表現総数	2258
データ件数 (負例)	5343(484)	データ件数 (負例)	1991(1317)

れており，各データは 1 文で構成されている．各文に出現する固有表現に対し，表 1 にあるような 8 種類のラベルが付与されている．また，484 件 (約 9%) のテキストデータが負例として，ラベルを含まないデータとなっていることが特徴である．

ニュース記事データセット

株式会社ユーザベースが運営する NewsPicks が取得した 2023 年 1 月〜6 月までのニュース記事から，金融・経済分野の記事を中心にサンプルした 1991 件の記事データである．政治・スポーツなどの一般的なジャンルの記事からもサンプルされている．

前述のデータと異なり，1 件のデータは主に 1 記事[1)]に対応し，タイトル・本文とこれらに含まれる固有表現の情報から構成される．固有表現の前後の文脈を含むことや，本文途中での言い換えなどで，略称や別名として組織名が登場するケースも含んでいる．1 件の記事データは，記事本文の長さにばらつきがあり，アノテーションコストや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，タイトルと本文冒頭 10 文までのテキストに対し，アノテーションがなされている．

実データからのサンプルであるため，全 1991 件のデータのうち，負例 (固有表現を含まない記事) が 1317 件 (66%) となっている．公開データにあるラベルに対して，表 1 のような類似するラベルが付与されている．

1)　モデルに入力する最大トークン長を考慮して，長文の記事は分割している．

データ間のラベルの対応

公開データとニュース記事テキストデータに付与されているラベルの内訳は表のとおりである．本研究では組織名の抽出に注目するので，公開データセットにおける「法人名」ラベルとニュース記事データの「組織名」ラベル，「製品名」ラベルと「製品・サービス名」をそれぞれ対応させる．そのため，そのほかの 6 種類のラベルについては，本稿では過去研究との比較のみ利用する．

3　組織名抽出のモデル

本研究で実施するニュース記事における組織名の抽出には，田村らが提案したモデル [9] を参考とする．特に，先行研究でベースとされていた方式の中で，固有表現抽出モデルの部分を利用する．

先行研究では，東北大が公開する BERT[14] を利用しているが，本研究では山田らが提案した LUKE[15] に置き換える．LUKE は，entity-aware self-attention と呼ばれる機構を内部に持ち，Wikipedia から選ばれたエンティティなどを通常の単語とは別に学習することで，エンティティ単位を重視するタスクに対し有効であることが知られている．さらに，LUKE は SentencePiece をトークナイザーとすることで公開されているが，我々は日本語の単語境界をより正確にとらえるために，トークナイザーを MeCab に変更し，Wikipedia の 7 月 1 日時点の日本語記事で事前学習し，知識を更新している [16]．

我々は，この LUKE をベースとして，最終層の埋め込み表現を，CRF 層に接続する構造とした．CRF 層は，入力系列から発生させる出力系列が，教師データの系列と等しくなるように遷移関係を調整す

ることができる．これを，各トークンごとのラベル分類の問題として解くモデルと組み合わせることでの有効性が報告 [3] されているため，ここでは，最終層の各トークンの埋め込み表現を，ラベルの個数に対応した埋め込み表現に変換し，CRF 層に入力する構造としている．

4 数値実験の設定

本研究でのモデル改善による組織名の抽出精度の評価については，2 つのデータセットにおける学習と精度評価を通して行う．データの利用にあたり，それぞれのデータのラベルをニュース記事データに対応させる．つまり，固有表現抽出のための系列ラベルの種類は，組織名と製品・サービス名の 2 種類を扱うこととしている．また，モデルには表 2 のようにサブワードに対して，固有表現の開始部 B，中間部 I，終端部 E とした IOBE 形式で入力する．

表 2 IOBE 形式におけるデータの保持

句	タグ
(株)	B-ORG
ユーザ	I-ORG
ベース	E-ORG
が	O
発表	O
した	O
.	O

数値実験設定として，下記の 3 つのケースで，モデルをファインチューニングし，精度評価を行う．

(I) 公開データセット 5343 件を (訓練: 検証：テスト)=(8:1:1) に分割したデータセット．データに付与された 8 種類の固有表現ラベルを学習する．

(II) ニュース記事データ 1991 件のうち，500 件をテスト用データとする．データに付与された 2 種類の固有表現ラベルを学習する．

(III) ニュース記事データ 1991 件のうち，500 件をテスト用データとする．ニュース記事データ 1491 件の学習データと公開データセット 5343 件のデータを混合させ 6834 件データセットとしたうえで，ラベル対応させた 2 種類の固有表現ラベルを学習する．

ケース (I) は，公開データセットであるため，先行研究と本研究の提案モデルと比較するために設定する．ケース (II)，(III) は，独自の実ニュース記事データセットに対し，モデルをあてはめ，本モデルのニュース記事の情報抽出モデルとしての精度を評価する．また，データの混合により精度改善するかを検証する．

モデルの学習において，IOBE 形式における禁止遷移に関しては，CRF 層の遷移コストを-100 に設定している．また，512 トークンを超えるテキストに関しては，512 トークンを超えないようにトークン数を当分割している．各データセットの分割を 5 セット行い，それぞれで学習を 10epoch 行い，テストデータでの精度検証を行う．

5 結果

モデルがテストデータで推論した固有表現の出現位置とラベルの組が，教師データと一致するかを，Precision, Recall, F1 を評価指標として表 3 として算出した．

本研究による提案モデルにより，過去研究 [8] で示されたものより全体精度の向上が見られている．具体的には，政治的組織名，その他の組織名や施設名以外のラベルにおいて F1 値の改善が見られていることや，全体の F1 値が改善している．このことから，本研究におけるモデルの改良により，固有表現抽出モデルとしての一定程度の改善が見られている．

さらに，同モデルをニュース記事データのみで学習したケース (II) の場合よりも，ニュース記事と公開データを混在させて学習したケース (III) の場合において，精度値の改善が見られている．製品・サービス名の精度値はもともとの水準が低いが，これは，企業名と製品・サービス名が同名であるケースがデータ中に多く含まれ，固有表現として位置情報を正確に抽出しつつも，ラベルのタイプ判断を誤るケースが見られた．一方，公開データは短文で構成されていることや製品名ラベルに関するデータが多いことから，検証データにおける短文や製品名ラベルにおける精度の改善が見られた．しかし，固有表現の出現割合（正例の比率）が公開データセットの方が高いため，検証データセットでの組織名ラベルの Precision が下がる形でのモデルの傾向の変化が見られた．

6 まとめと今後の課題

組織名を抽出するモデルとして，LUKE を改善し CRF 層を付加する固有表現抽出モデルを提案した．

表 3 数値実験における精度指標．数値は各試行で得られた精度指標の平均値である．

	ラベル	Pr	Re	F1
データセット (I) (公開)	人名	97.22	94.66	95.92
	法人名	94.20	91.95	93.06
	政治的組織名	91.26	87.85	89.52
	その他の組織名	79.90	80.12	80.01
	地名	90.91	92.68	91.79
	施設名	85.59	87.96	86.76
	製品名	85.09	85.84	85.46
	イベント名	91.58	91.58	91.58
	Micro	90.95	90.51	90.72
	Macro	88.99	88.88	89.26
データセット (II) (ニュース)	企業名	83.77	80.12	81.90
	製品・サービス名	85.11	43.48	57.55
	Micro	83.89	74.26	78.78
	Macro	84.44	61.80	69.73
データセット (III) (公開＋ニュース)	企業名	82.06	82.40	82.23
	製品・サービス名	80.36	48.91	60.81
	Micro	81.89	77.04	79.39
	Macro	81.21	65.66	71.52

公開データセットとニュースの実データセットでそれぞれ精度評価を行った結果，提案モデルが過去研究における精度を一定程度上回ることが確認され，固有表現抽出モデルとして期待できる水準であることが示唆された．

さらに実データを使った検証により，ニュース記事データにおいても先行研究 [9] と同水準となり，ニュース記事の情報抽出として有用であることが期待できる．ただし，類似のデータセットではあるが，異なるもののため，厳密な比較ではないことは注意が必要である．

一方で，公開データと実データであるニュース記事データを混合して使うことでのモデルの全体精度改善の効果が見られたが，データ間での特徴に乖離があることから，その混合の比率や分布の補正などは検討の余地がある．特に，データ間の特徴的な差異である，文章長や負例の割合は大きくモデルの傾向を変えていて，これらの特徴を加味したうえでデータの混合や拡張を行うことが必要である．

参考文献

[1] A. McCallum J. Lafferty and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In **ICML 2001**, 2001.

[2] K. Lee J. Devlin, M. Chang and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **arXiv:1810.04805**.

[3] R. Nogueira F. Souza and R. Lotufo. Portuguese named entity recognition using bert-crf. In **arXiv:1909.10649**.

[4] L. Cui and Y. Zhang. Hierarchically refined label attention network for sequence labeling. In **EMNLP-IJCNLP**, 2019.

[5] F. Xiao Y. Luo and H. Zhao. Hierarchical contextualized representation for named entity recognition. In **AAAI-2020**, 2020.

[6] 奥田裕樹, 橋寛治. ニュース記事からの企業キーワード抽出. 言語処理学会 第 26 回年次大会 発表論文集, 2020.

[7] 橋寛治, 奥田裕樹. 辞書に基づく組織名抽出における辞書整備の影響. 言語処理学会 第 26 回年次大会 発表論文集, 2020.

[8] 橋本航, 笛木正雄, 黒木裕鷹. 日本語固有表現抽出における bert-mrc の検討. 言語処理学会第 28 回年次大会, 2022.

[9] 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.

[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] 田村光太郎. 固有表現抽出タスクにおける文章のランダム連結によるデータ拡張. In **2023-IFAT-153(3)**, 2023.

[12] 橋本航, 橋寛治. Bert を用いた固有表現抽出におけるバッチ能動学習. 信学技報, vol. 121, no. 82, 2021.

[13] Wikipedia を用いた日本語の固有表現抽出データセット, (2023-01 閲覧).

https://github.com/stockmarkteam/ner-wikipedia-dataset.
[14] （2023-01 閲覧）. https://huggingface.co/cl-tohoku/bert-base-japanese-v3.
[15] 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.
[16] （2023-01 閲覧）. https://huggingface.co/uzabase/luke-japanese-wordpiece-base.