日本語固有表現抽出における BERT-MRC の検討

橋本航 [1]　笛木正雄 [1,2]　黒木裕鷹 [1] 髙橋寛治 [1]
[1]Sansan 株式会社　[2] 東工大工学院経営工学系
{wataru.hashimoto, fueki, kuroki, ka.takahashi}@sansan.com

概要

固有表現抽出は，テキスト中から人名や組織名などの固有表現を抽出する技術である．近年は，BERT を始めとした事前学習モデルをベースとしたアプローチの発展が著しい．中でも BERT を用いた機械読解 (Machine Reading Comprehension; MRC) として固有表現抽出を解くモデルである BERT-MRC は，複数の英語データセットにおいて最高性能を示している．本稿では，日本語データセットを対象にして BERT-MRC の精度検証やエラー分析を行い，有効性を検討した．その結果，BERT-MRC は精度としての有効性は確認されなかったものの，固有表現抽出と機械読解形式それぞれの場合で有効性を示すケースを確認することができた．

1　はじめに

固有表現抽出 (Named Entity Recognition; NER) は，テキスト中から人名や組織名など固有表現を抽出する，自然言語処理において基礎的かつ重要な技術である．固有表現抽出は個人情報のマスキングや文書からの情報抽出などに用いられ，Sansan 株式会社が提供する営業 DX サービス「Sansan」[1)]や名刺アプリ「Eight」[2)]においてもニュース記事中から抽出した企業名をキーとしてニュース配信に用いている．

近年は，事前学習済みの言語モデルである BERT[1] を始めとした様々な言語モデルを固有表現抽出タスク向けにファインチューニングし，優れた性能を示す例が数多く報告されている．その中でも，2019 年に提案された，固有表現抽出を機械読解形式で解くモデルである BERT-MRC[2] は，複数の英語データセットにて当時の最高性能を示したモデルである．BERT-MRC は 1 つのトークンに複数のラベルが割り当てられる場合も考慮する nested-NER にも利用可能など，拡張性の高いモデルである．機械読解形式で固有表現抽出を解く場合，固有表現の各ラベルは質問文と紐付けられ，対象文から質問文に対応する固有表現が抽出される．例えば，抽出対象の文が「Sansan 株式会社は、Eight を提供する。」であり，質問文が「会社名や企業名を含む法人を探せ」の場合，「Sansan 株式会社」が抽出される．機械読解形式で解く利点として，ラベルを構成するトークン列が持つ意味的な情報を活用できる点がある．

しかし，BERT-MRC は英語では優れた性能を示すことがわかっているものの，日本語での有効性は明らかになっていない．本稿では，日本語に対する固有表現抽出タスクにおいて精度検証やエラー分析を行い，日本語での BERT-MRC の有効性を検討する．

2　関連研究

本章では固有表現抽出に関連する研究について述べる．固有表現抽出では，同じトークンでも文脈によってラベルが異なる場合があるため，系列における過去と未来の双方向の依存性を考慮したモデルが主流である．Huang ら [3] は BiLSTM-CRF による固有表現抽出を行っている．BiLSTM は，過去と未来の双方向について，それぞれ隠れ表現を得るモデルである．Akbik ら [4] は，BiLSTM-CRF に入力する表現として，語自身の意味表現だけでなく文脈上の使われ方を考慮した新しい単語埋め込みを入力する手法を提案し，高い性能を示している．

近年は事前学習済み言語モデルの導入が進んでおり，その顕著な例が BERT である．BERT は Transformer[5] をベースとし，Masked Language Model および Next Sentence Prediction による双方向の事前学習を導入した事前学習モデルであり，固有表現抽出をはじめとした様々な自然言語処理タスクに応用可能である．また，BERT に CRF 層を組み合わせた BERT-CRF により，BERT を用いた固有表現抽出モデルと比較して精度向上することが報告されており [6]，日本語における BERT-CRF の有効性も確認されている [7].

1）https://jp.sansan.com/
2）https://8card.net/

さらに，自然言語処理タスクを機械読解形式で解く方法も提案されており，Li ら [8] は機械読解形式で関係抽出を行い，大幅に精度が向上したことを報告している．本稿では，機械読解形式で日本語固有表現抽出タスクを行った場合の有効性を検討する．

3 実験及び評価

3.1 モデル

本稿では，比較対象となる通常の固有表現抽出モデルとして BERT-CRF，機械読解形式で解くモデルとして BERT-MRC を用い実験を行った．

3.1.1 BERT-CRF

BERT-CRF は，BERT を用いた固有表現抽出の出力ラベルの遷移が正しくなるように制約をかけたモデルである．入力系列のトークン数を n，ラベルの数を K，入力系列を $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$，出力ラベルを $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ とすると，遷移を考慮した系列のスコア $s(\mathbf{X}, \mathbf{y})$ は以下のように書ける．

$$s(\mathbf{X}, \mathbf{y}) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{1}$$

ここで $A \in \mathbb{R}^{(K+2)\times(K+2)}$ はラベル間の遷移行列，P_{i,y_i} はトークン i における出力 y_i の確率である．目的関数は式 (1) のスコア関数を用いた対数尤度関数で表わせ，最大化するように学習を行う．

$$\log p(\mathbf{y}|\mathbf{X}) = s(\mathbf{X}, \mathbf{y}) - \log\left(\sum_{\hat{\mathbf{y}} \in \mathbf{Y_X}} e^{s(\mathbf{X}, \hat{\mathbf{y}})}\right) \tag{2}$$

ここで $\mathbf{Y_X}$ は $\mathbf{y}$ がとりうる全系列の集合である．$\mathbf{Y_X}$ の要素数は膨大であるため，式 (2) の最大化にはビタビアルゴリズムを用いる．

3.1.2 BERT-MRC

BERT-MRC は，固有表現抽出を質問に対する回答に該当するスパンを抽出する機械読解形式で解くモデルである．BERT-MRC では，入力を質問文と抽出対象文を結合したトークン列 $\{[\text{CLS}], q_1, q_2, ..., q_m, [\text{SEP}], x_1, x_2, ..., x_n\}$ を BERT に入力し，対象文のトークンのベクトル表現行列 $E \in \mathbb{R}^{n\times d}$ を得る．ここで m は質問文のトークン数，d は BERT の最終層の出力ベクトルの次元である．得られた E を用いて，そのトークン位置がスパンの始点か否かを予測する．

表 1 固有表現抽出のデータ例 (IOB2 フォーマット)

トークン	ラベル
Sansan	B-組織名
株式	I-組織名
会社	I-組織名
は	O
、	O
Eight	B-製品名
を	O
提供	O
する	O
。	O

表 2 機械読解形式のデータ例

対象文	Sansan 株式会社は、Eight を提供する。
対象文 (トークナイズ後)	Sansan/株 式/会 社/は/、/Eight/を/提供/する/。
質問文	会社名や企業名を含む法人を探せ
スパン	0;2
始点 (start)	0
終点 (end)	2

$$P_{\text{start}} = \text{softmax}(E \cdot T_{\text{start}}) \in \mathbb{R}^{n\times 2} \tag{3}$$

ここで $T_{\text{start}} \in \mathbb{R}^{d\times 2}$ は重みパラメータである．スパンの終点に対しても同様の予測を行う．

また，上記の予測結果を用いるだけでは質問文に対応する回答のスパンが一意に定まらない可能性がある．そのため，スパンの始点と終点のマッチングについても予測を行い，これらを統合することで質問文に対するスパンを出力する．

$$P_{i_{\text{start}}, i_{\text{end}}} = \text{sigmoid}(W \cdot \text{concat}(E_{i_{\text{start}}}, E_{i_{\text{end}}})) \tag{4}$$

ここで $W \in \mathbb{R}^{1\times 2d}$ は重みパラメータである．

3.2 実験設定

データセットとして，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 [9] を用いた．学習データ，開発データ，テストデータの比率が 8:1:1 になるようにデータを分割した．また，当該データセットを固有表現抽出および機械読解のタスクへ適用するため，それぞれ表 1，表 2 のように整形を行った．固有表現抽出のラベルのフォーマットとして IOB2 フォーマットを用

表 3　固有表現ラベルに対応する質問文

ラベル名	質問文
人名	人を探せ
法人名	会社名や企業名を含む法人を探せ
政治的組織名	政治的組織名、政党名、政府組織名、行政組織名、軍隊名、国際組織名を探せ
その他の組織名	競技組織名、公演組織名を含む組織名を探せ
地名	都道府県や市区町村、住所、郵便番号を含む地域を探せ
施設名	公園や駅を含む施設を探せ
製品名	芸術作品名、出版物名、規則名、乗り物名、キャラクター名を含む製品を探せ
イベント名	イベントを探せ

表 4　BERT-CRF と BERT-MRC の固有表現抽出における固有表現ラベルごとの結果

	BERT-CRF			BERT-MRC		
	Pre.	Rec.	F1	Pre.	Rec.	F1
人名	93.56	96.17	94.85	83.12	91.43	87.07
法人名	94.17	91.90	93.00	87.34	84.49	85.89
政治的組織名	93.04	91.45	92.24	86.67	88.89	87.76
その他の組織名	84.30	85.00	84.65	82.57	75.00	78.60
地名	91.53	90.58	91.05	87.22	82.20	84.64
施設名	84.91	90.00	87.38	82.29	79.00	80.61
製品名	77.95	88.39	82.85	72.07	74.77	73.39
イベント名	85.60	89.17	87.35	84.48	81.67	83.05
micro avg	89.53	91.26	90.39	83.87	83.67	83.77
macro avg	88.13	90.33	89.17	83.22	82.18	82.63

いた．また，BERT の事前学習済みモデルとして東北大の日本語 BERT モデルを使用した．[3)]

BERT-CRF については，不正な遷移[4)]に対応するパラメータを-100 で初期化した状態で学習を行った．

BERT-MRC については，negative sampling を導入して学習を行った．ここでいう negative sample とは，質問文に対応する固有表現が対象文中に存在しないデータのことであり，negative sample を間引いて学習するのが negative sampling である．固有表現抽出を機械読解形式で解く場合，1 文に対してラベルの個数分の質問文を生成する．そのため，文によっては該当するラベルがない事例が生成されることになるケースが多い．negative sample が多い場合，学習が効率的に進まないため，negative sampling が必要である．本稿では，positive sample (質問文に対応する固有表現が対象文に存在するデータ) に対して negative sample の比率が 1.5 倍になるよう学習データからサンプリングし学習を行った．また，各固有表現ラベルに対応する質問文の一覧を表 3 に示す．質問文は [9] の各ラベルを参考にして作成した．

3.3　評価方法

評価指標として Precision(適合率), Recall(再現率), F1(F-measure) を用いた．トークン列に対して正しいラベルが付与できたか否かを基にして算出を行った．また，固有表現抽出タスクにおいて BERT-MRC と BERT-CRF の性能を同一条件で比較するため，BERT-MRC を固有表現抽出向けの精度評価ができるように出力結果の後処理を行った．具体的には，BERT-MRC では各固有表現ラベルごとに独立に推定が行われるため、flat-NER（1 つのトークンに 1 つのラベルのみを割り当てる固有表現抽出）向けに変換する後処理が必要となる．本稿では，人名 > 法人名 > 政治的組織名 > その他の組織名 > 地名 > 施設名 > 製品名 > イベント名の優先順位をつけて各トークンにラベル付けを行った．

3）　https://github.com/cl-tohoku/bert-japanese
4）　不正な遷移とは，例えば「B-組織名→ I-人名」のような，IOB とラベルが一致せず起こりえない遷移を指す．

3.4 結果

BERT-CRF および BERT-MRC の結果を表 4 に示す．それぞれのラベルにおけるすべての評価指標において，BERT-CRF の方が良い性能を示した．各評価指標のスコアを比較しても差が大きいものも見られることから，日本語の固有表現抽出（flat-NER）においては，機械読解形式のモデルよりも通常の固有表現抽出モデルを用いた方が良いことがわかる．

3.5 エラー分析

本セクションでは，BERT-MRC および BERT-CRF のエラー事例を示し考察を行う．まず，BERT-MRC において正しく抽出されなかったが BERT-CRF では正しく抽出された例の抜粋を表 5 に示す．

表 5 では，BERT-CRF においては鉤括弧内の固有表現を抽出できているが，BERT-MRC では抽出できていないことを示している．鉤括弧内に何らかの固有表現が含まれているのは今回しようしたデータセットにおいて数多く見られ，実際にテストデータの全 534 件において鉤括弧が含まれるものは 137 件存在する．その中でも鉤括弧内がすべて固有表現である 83 件において BERT-MRC は 69 件正解していたのに対し，BERT-CRF は 77 件正解していた．本事例において BERT-MRC が BERT-CRF と比較し正しく抽出できていないことが，日本語固有表現抽出においては固有表現抽出モデルに劣後している理由の 1 つであると考えられる．

次に，BERT-CRF において正しく抽出されなかったが BERT-MRC では正しく抽出された例の抜粋を表 6 に示す．こちらは，ラベル名に対応する質問文「政治的組織名、政党名、政府組織名、行政組織名、軍隊名、国際組織名を探せ」と抽出した固有表現「最高裁」が意味的に近い事例であり，日本語においても BERT-MRC が有効である領域であるといえる．本結果から，対象としたい固有表現とより対応するように質問文を改善すれば，精度が向上する可能性があると考えられる．

4 おわりに

本稿では固有表現抽出を機械読解形式で解くモデルである BERT-MRC を日本語に対する固有表現抽出に適用した．精度としては通常の固有表現抽出として解くモデルには劣後し，その原因の一つとして今回使用した日本語固有表現抽出データセットの文中に鉤括弧が多いためであることが確認された．一方で，日本語においても対象文における固有表現と質問文の意味が近い場合は機械読解形式が機能することも確認された．今後は，nested-NER や zero-shot の問題設定における日本語固有表現抽出の検討が課題である．

表 5 BERT-MRC のエラー例

トークン	正解ラベル	BERT-MRC	BERT-CRF
原作	O	O	O
と	O	O	O
し	O	O	O
た	O	O	O
、	O	O	O
「	O	O	O
うち	B-製品名	O	B-製品名
は	I-製品名	O	I-製品名
サラ	I-製品名	O	I-製品名
ダ	I-製品名	O	I-製品名
編	I-製品名	O	I-製品名
」	O	O	O
39	O	O	O
話	O	O	O
で	O	O	O
は	O	O	O

表 6 BERT-CRF のエラー例

トークン	正解ラベル	BERT-MRC	BERT-CRF
2003	O	O	O
年	O	O	O
の	O	O	O
最高	B-政治的組織名	B-政治的組織名	O
裁	I-政治的組織名	I-政治的組織名	O
判決	O	O	O
を	O	O	O
もっ	O	O	O
て	O	O	O

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Association for Computational Linguistics, 2019.

[2] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Association for Computational Linguistics, 2020.

[3] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. **CoRR**, Vol. abs/1508.01991, , 2015.

[4] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In **Proceedings of the 27th International Conference on Computational Linguistics**. Association for Computational Linguistics, 2018.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017.

[6] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. **arXiv preprint arXiv:1909.10649**, 2020.

[7] 田川裕輝, 西埜徹, 谷口元樹, 谷口友紀, 大熊智子. 生成された読影所見の自動評価に向けた固有表現認識とモダリティ推定. 言語処理学会 第 26 回年次大会, 2020.

[8] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Association for Computational Linguistics, 2019.

[9] 近江崇宏. Wikipedia を用いた日本語の固有表現抽出のデータセットの構築. 言語処理学会 第 27 回年次大会, 2021.