Vol.2019-NL-243 No.5
2019/12/4

固有表現抽出によるブログテキストからの品名・店名抽出

池田 流弥[1,a)]　安藤 一秋[2,b)]

概要：固有表現抽出は，自然言語処理における重要な基礎技術の一つである．近年では，深層学習を用いた固有表現抽出手法が提案されており，CoNLL2003 データセットをはじめとする英語で書かれた整ったテキストに対して高い性能が得られている．我々は，現在，現地でしか購入できない土産に関する情報を Web 上から収集・提供するシステムの構築を進めている．システムを構築するために，日本語で書かれたブログ記事や Q&A サイトなどの User-Generated Content から，土産名と店名などを収集する必要がある．しかし，既存の固有表現抽出手法の多くは，UGC や日本語のテキストに対して評価されていない．本稿では，日本語ブログ記事から構築したデータセットを用いて，Conditional Random Fields（CRF）と深層学習による固有表現抽出手法の性能評価結果について述べる．評価結果より，CRF は学習データ中に出現しない固有表現に対して，深層学習モデルは学習データに含まれている固有表現に対して有効であることを確認した．

Extraction of Food Product and Shop Names from Blog Text using Named Entity Recognition

1. はじめに

固有表現抽出（Named Entity Recognition）は，人名や組織名などの固有名詞や日付表現，時間表現など，特定の実体を指す固有表現（Named Entity）を文中から自動抽出する技術であり，自然言語処理において，古くから研究されてきた．近年，深層学習を用いた固有表現抽出手法が提案され，従来の抽出手法より高い抽出性能が得られようになったと報告されている [4,8,11,13]．それらの手法のほとんどは，CoNLL2003 データセット [22] などの英語で書かれた整ったテキストで評価され，不特定多数の人が書いたレビューサイト上の書き込みや SNS 上のテキストなどの User-Generated Content（UGC）や英語以外で書かれたテキストに対して評価されることは少ない．

我々は，現地でしか購入できない土産に関する情報を Web 上から収集し，ユーザに提供するシステムの開発を進めている [16,27]．本システムを実現するためには，日本語で書かれたブログ記事や Q&A サイトなどの UGC から，土産名と店名，評判などの土産情報を収集する必要がある．土産情報の中でも，土産名が特定できなければ，どの土産に関する情報なのか紐づけできない．そこで，固有表現の一種である土産名と店名に注目し，固有表現抽出により，ブログ記事から食品に属する土産名と店名を自動抽出することを目指す．

現在までに，多数の固有表現抽出手法が提案されているが，UGC や日本語のテキストに対して評価されているものは少ない．また，CoNLL2003 のようなデータセットが対象とする固有表現クラスは，主に，人名，場所，組織などの一般的な固有表現である．一方，食品に属する土産名は，表現の多様性が高く，一般的な固有表現の抽出と比較して，難しいタスクと考えられる．

本稿では，食品に属する土産名と店名に注目し，日本語ブログ記事から構築したデータセットを用いて，Conditional Random Fields（CRF）と深層学習による固有表現抽出手法の性能評価結果について述べる．

2. 関連研究

2.1 固有表現抽出

固有表現抽出は系列ラベリング問題として解くことができ，従来手法として，SVM（Support Vector Machine）や CRF を用いる手法が提案されている [9,12,18,21]．これら

[1] 香川大学大学院 工学研究科
[2] 香川大学 創造工学部
[a)] s18g454@stu.kagawa-u.ac.jp
[b)] ando@eng.kagawa-u.ac.jp

の手法では，単語の表記や品詞などを素性として利用する．

また，近年では，深層学習による固有表現抽出手法が提案されている．Huang らは，Bidirectional-LSTM（BiLSTM）と CRF を組み合わせた Bidirectional-LSTM CRF（BiLSTM-CRF）を提案している [8]．BiLSTM-CRF モデルでは，word embedding を BiLSTM に入力することで得られるベクトルを，素性の代わりに CRF に与えることで固有表現を抽出する．また，Huang らの BiLSTM-CRF モデルに対して，注目単語に含まれる文字情報を加えることで性能を向上させるモデルも提案されている．Ma らは，注目単語に含まれる文字を CNN に入力することで得られた単語ベクトルを word embedding に結合し，BiLSTM-CRF に入力して固有表現タグを予測する BiLSTM-CNN-CRF を提案した [13]．Lample らは，CNN の代わりに BiLSTM に対して注目単語に含まれる文字を入力して単語ベクトルを獲得し，BiLSTM-CRF の入力に加えるモデルを提案した [11]．さらに，文脈情報を加味した分散表現である BERT や Contextual String Embeddings などを利用した手法も提案されており，固有表現抽出タスクで高い性能を報告している [4, 6]．

しかし，これらの手法のほとんどは，整った英語のテキストからなるデータセットで，一般の固有表現を対象に評価されており，日本語の UGC に対する性能は不明である．

2.2 日本語に対する固有表現抽出

日本語に対する固有表現抽出についてもいくつかの手法が提案されている．Sassano らは，構造的な解析などから得られる大域的な情報を素性とした SVM で固有表現を抽出する手法を提案した [21]．また，Iwakura は，アノテーションされていないデータから取得したルールと単語情報を利用した抽出手法を提案した [9]．

近年では，深層学習を用いた日本語に対する固有表現抽出手法が提案されている．Misawa らは，Ma らが提案した BiLSTM-CNN-CRF が日本語の固有表現抽出には有効的でないことを示し，Character-Bidirectional-LSTM-CRF（Char-BiLSTM-CRF）を提案した [15]．図 1 に Char-BiLSTM-CRF モデルを示す．このモデルは文字単位で固有表現タグを推定するモデルである．文字ベクトルと word embedding から得られた単語ベクトルを結合し，BLSTM Layer への入力として与える．その後，BLSTM Layer で得たベクトルを CRF Layer の入力として与え，固有表現タグを推定する機構となっている．このモデルは，毎日新聞コーパスに対する実験において，最高性能を示している．

また，Mai らは，日本語の拡張固有表現に対して深層学習を用いた抽出手法を提案している [14]．このモデルは，Wikipedia から構築した辞書情報と固有表現クラスを利用することで日本語拡張固有表現に対して，抽出性能を向上させている．

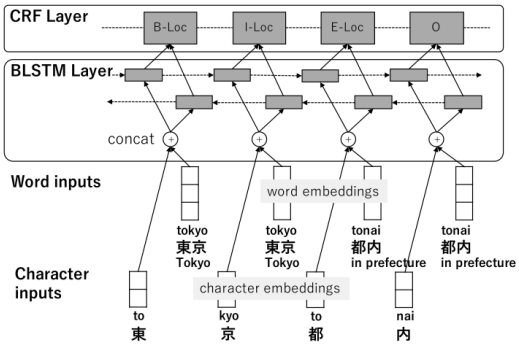

図 1 Misawa らの提案モデル（[15] より引用）
Fig. 1 A neural model proposed by Misawa et al. (sited from [15])

しかし，これらの手法も，日本語で書かれた新聞記事を対象に評価したものがほとんどであり，日本語の UGC に対する性能は不明である．

2.3 UGC に対する固有表現抽出

近年では，表現が多様で表記揺れが頻出する UGC に対する固有表現抽出手法も提案されている．Ritter らは，ツイートに対する固有表現抽出手法を提案している [20]．Aguilar らは，マルチタスク学習を用いた固有表現抽出手法を提案し，ソーシャルメディアから構築された WNUT2017 データセットに対して，高い性能を報告している [3]．

一般的に，UGC はニュース記事と比べて，表現や語彙の多様性，スペルミスが多いことから，固有表現抽出の難易度が高く，抽出性能が低くなる．例えば，Akbik らの提案した固有表現抽出手法 [4] は，CoNLL2003 データセットに対して F 値で 93.09%を達成しているが，Aguilar らの提案したモデルは，WNUT2017 データセットに対して，45.55%という F 値が最高である．

3. 土産の品名・店名抽出手法

我々は，現地でしか購入できない土産に関する情報を Web 上から収集し，ユーザに提供するシステムの開発を進めている．現地でしか購入できない土産の情報をまとめた情報源は存在しないため，システムを構築するためには，Web 上から土産情報を抽出する必要がある．土産名や店名，評判などの土産情報は，ブログ記事や Q&A サイトなどの UGC に多く書かれている．しかし，土産名が特定できなければ，どの土産に関する情報なのか紐づけできない．土産名が得られれば，それをクエリとして，新たな土産情報を検索することもできる．そこで，本研究では，固有表現の一種である土産名と店名に注目し，固有表現抽出により，ブログ記事から食品に属する土産名と店名を自動抽出する．本稿では，最適な固有表現抽出手法を検討するため，CRF(Conditional Random Fields) と深層学習による固有表現抽出手法の性能を比較する．以下，データセットの構

築と各抽出モデルについて述べる．

3.1 データセットの構築方法

固有表現抽出手法の性能評価に利用するデータセットは次の手順で構築する．

(1) 土産の品名を必ず 1 つは含むブログ記事を収集し，1 文ずつ形態素解析する．
(2) 各形態素に対して，以下のルールでタグ付けする．

- 食品名に品名タグ (PRO) を付与
- 食品を販売している店に店名タグ (SHO) を付与
- 「」などの記号を含めて品名，店名タグを付与
- 品名，店名でないものに O タグを付与

食品名に品名タグを振る理由は，食品が土産を包含しているからである．土産でない食品も将来的には土産になる可能性があるため，土産と商品は区別せずにタグ付けする．また，ラベリングには，BIOES タグ形式を用いる [19]．

3.2 CRF モデル

我々の先行研究 [27] により，CRF は単語単位でラベリングするモデルを採用する．素性としては以下を利用する．

- 表記
- 品詞・品詞細分類
- 文字種
- 括弧内の単語にフラグを立てる（inBracket）

文字種には，数字，英小文字，英大文字，ひらがな，カタカナ，漢字，その他の 7 種を用いる．1 単語に複数の文字種が含まれていた場合，単語に含まれるすべての文字種を素性とする．日本語は英語に比べて文字種が多いため，この素性を加えることによる性能向上に期待できる．また，土産の品名は様々な文字種で構成されることが多いが，店名はアルファベットだけや漢字だけなどと，1 つの文字種で構成されることが多いため，固有表現クラスの識別にもこの素性が有効的に働く可能性がある．

また，ブログ記事中で，土産名や店名が「」や () などの括弧中に書かれやすいことに注目し，“inBracket”素性を採用する．

3.3 深層学習モデル

本稿では, 深層学習による固有表現抽出において, 代表的なモデルである BiLSTM-CRF (Huang Model) [8], BiLSTM-CRF（Lample Model）[11]，BiLSTM-CNN-CRF[13] の 3 種と日本語固有表現抽出において最高性能を報告している Char-BiLSTM-CRF[15] を用いる．

4. 評価実験

本研究で構築したデータセットを用いて，CRF と深層学習モデルの性能を比較する．本実験において，形態素解析器には MeCab[10] を，CRF の実装には CRFsuite[17]，深層学習モデルの実装には AllenNLP[7] を用いる．MeCab の辞書には IPADIC を用いる．CRF の window size は 2 とし，その他のハイパーパラメータは CRFsuite のデフォルト値を用いる．

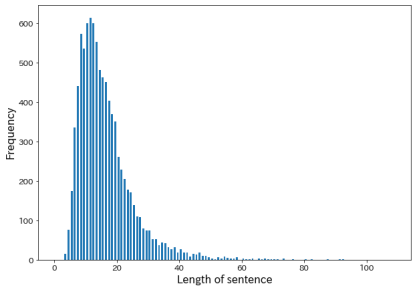

図 2 データセット中の文長の分布
Fig. 2 Distribution of the length of sentences in the experimental data

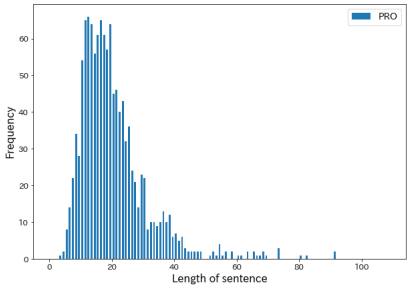

図 3 データセット中の品名を含む文の分布
Fig. 3 Distribution of the length of sentences by product entities in the experimental data

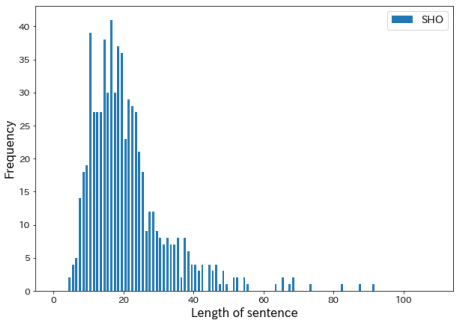

図 4 データセット中の店名を含む文の分布
Fig. 4 Distribution of the length of sentences by shop entities in the experimental data

4.1 データセット

3.1 節で述べた方法でデータセットを構築するため，日本全国の著名な土産がまとめられた Web サイトである OMIYA![1] に 2018 年 4 月 27 日時点で掲載されていた 7,531 件の土産名をクエリとして，Yahoo!ブログ [2] の菓

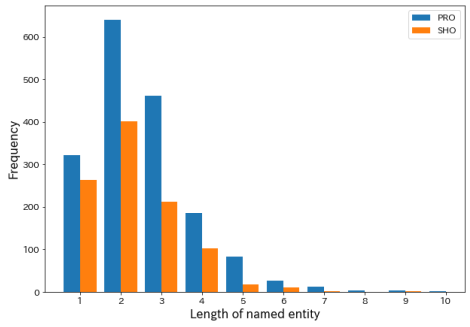

図 5 データセット中の固有表現長の分布
Fig. 5 Distribution of the length of named entities in the experimental data

子・デザートカテゴリ内でヒットしたブログ記事の本文を利用する．得られたブログ記事の中から土産名を含む記事をランダムに 680 エントリを選択し，13,890 文に対して人手で固有表現タグを付与した．また，菓子・デザートカテゴリのブログ記事中には菓子の画像に対するキャプションや箇条書きだけで 1 文が構成されるものが多い．単語のみの文に対して固有表現抽出を適応する場合，固有表現の表層文字列に抽出性能が影響されやすいため，本実験では，以下の条件を満たさない文を取り除く．

- 文中に名詞及び助詞を含む
- 文中に動詞，形容詞，助動詞のいずれかを含む

本実験では，最終的に得られた 9,483 文をデータセットとして用いる．データセット中に，品名は 1,418 件，店名は 750 件含まれており，そのうち品名は 939 種，店名は 476 種であった．

図 2 にデータセット中の文長の分布を示す．文長は文に含まれる形態素の数を用いて計測した．図 2 から，データセット中には文長が 10〜30 の文が多いことを確認できる．図 3 に品名（PRO）を含む文の長さの分布を，図 4 に店名（SHO）を含む文の長さの分布を示す．図 3 と図 4 から，品名を含む文と店名を含む文の間で，文長に大きな違いは見られないことがわかる．図 5 にデータセットの固有表現長の分布を示す．図 5 より，品名と店名の固有表現長の傾向には大きな差はないことがわかる．

4.2 深層学習モデルの設定

表 1 に深層学習モデルで用いたパラメータを示す．これらのパラメータは予備実験により決定した．一般的に，深層学習モデルでは，最適化手法に SGD を用いることで高い性能が得られやすいと報告されている [23] が，予備実験で SGD と Adam を比較した結果，Adam の方が性能が高くなったため採用した．Dropout は，BiLSTM への入力の前に適応した．

表 1 の下部に BiLSTM-CNN-CRF と Lample Model で用いたパラメータを示す．Huang モデル以外の深層学習モデルで用いる文字ベクトルは 640 次元とし，BiLSTM-CNN-CRF と Lample Model の CNN または Char-BiLSTM の隠れ層は 240 次元とした．これらのパラメータは [24] を参考に決定した．文字ベクトルは [13] を参考に，$[-\sqrt{\frac{3}{dim}}, \sqrt{\frac{3}{dim}}]$ の範囲で初期化した．ここで dim は文字ベクトルの次元数であり，本実験では 640 とする．また，日本語の 1 単語あたりの文字数が英語に比べて少ないため，BiLSTM-CNN-CRF で用いられる CNN の window size は 2 とする．

表 1 深層学習モデルのハイパーパラメータ
Table 1 Parameters used for all neural models

Common parameters	
BiLSTM の隠れ層の次元数	128
BiLSTM の層数	1
最大エポックサイズ	50
バッチサイズ	32
学習率	0.001
Dropout rate	0.5
勾配クリッピング	5.0
最適化手法	Adam
Early stopping patience	20
Parameters of BiLSTM-CNN-CRF model	
CNN のフィルタ数	240
CNN の window size	2
Parameters of Lample model	
character BiLSTM の隠れ層の次元数	240
character BiLSTM の層数	1

表 2 事前学習した単語分散表現のハイパーパラメータ
Table 2 Parameters used for pretraining word embeddings

モデル	cbow
次元数	200
Window size	5
ネガティブサンプリング	5
ダウンサンプリング	0.001

単語ベクトルとして用いる分散表現には日本語 Wikipedia の本文全文で事前学習されたものを用意した[*1]．事前学習に用いたパラメータは表 2 に示すものが使われている．単語ベクトルおよび文字ベクトルはモデルの学習とともに値を更新する．

4.3 評価方法

適合率，再現率，F 値を評価指標とし，10 分割交差検証により抽出性能を評価する．交差検証の際，データセットを 90:5:5 の割合で学習，開発，テストデータに分割する．人手でタグ付けした結果とモデルによってラベリングされ

[*1] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

表 3 品名に対する抽出性能
Table 3 Experimental results for PRO

Model	all			unknown			known		
	precision	recall	f1-measure	precision	recall	f1-measure	precision	recall	f1-measure
CRF	**76.51**	61.03	**67.61**	**59.44**	**49.39**	**53.51**	88.51	68.74	77.03
Huang Model	70.45	62.10	65.46	48.15	42.96	43.71	86.59	75.59	80.60
BiLSTM-CNN-CRF	67.82	62.40	64.48	43.31	39.82	40.16	86.25	**78.26**	**81.93**
Lample Model	65.89	**64.88**	65.13	42.58	47.21	44.29	85.36	77.53	80.84
Char-BiLSTM-CRF	71.12	60.48	65.01	46.97	38.59	41.41	**88.59**	75.31	81.20

表 4 店名に対する抽出性能
Table 4 Experimental results for SHO

Model	all			unknown			known		
	precision	recall	f1-measure	precision	recall	f1-measure	precision	recall	f1-measure
CRF	**83.45**	56.13	66.64	**61.10**	**34.57**	**41.84**	**94.49**	69.70	79.84
Huang Model	83.12	62.77	71.11	47.21	19.49	25.43	92.66	89.79	90.91
BiLSTM-CNN-CRF	79.32	64.79	71.03	40.62	23.87	28.62	91.68	90.36	90.84
Lample Model	81.83	66.78	73.20	51.55	27.27	32.82	92.37	**90.40**	**91.16**
Char-BiLSTM-CRF	83.42	**67.34**	**74.35**	57.23	30.43	38.38	94.21	89.03	91.15

た結果を比較し，完全一致した場合のみを正解と判定する．

本研究では，現地でしか購入できない土産の品名と店名が主要な抽出対象であるため，学習データ中に含まれない未知の固有表現に対する性能が重要となる．固有表現抽出では，一般的に未知の固有表現に対する抽出性能が低くなることが知られている [5, 25]．そこで，固有表現の既知／未知（学習データに含まれる／含まれない）に分けて評価する．なお，交差検証による実験において，テストデータ中の未知の品名と店名の割合は平均約 40%である．

4.4 実験結果

表 3 に品名に対する実験結果を，表 4 に店名に対する実験結果を示す．すべてのモデルで最も性能が高いものを太字で示し，深層学習モデルの中で最も性能が高いものに下線を引いている．また，「all」はすべての固有表現に対しての性能，「unknown」は未知固有表現に対してのみの性能，「known」は既知固有表現に対してのみの性能を表す．

既知未知を区別しない場合，表 3 の all より，品名に対しては CRF が全体的に性能が高いことがわかる．また，表 4 の all より，店名に対しては深層学習モデルの性能が高いことがわかる．次に，固有表現の既知未知に注目すると，未知の固有表現に対しては CRF の性能が全体的に高い．特に未知の品名に対して，CRF は深層学習モデルと比べて，F 値で 9 ポイント以上性能が高い．一方，既知の固有表現に対しては，深層学習モデルの方が性能が高く，特に既知の店名に対する再現率は，CRF と比べて 20 ポイント以上高い．これらの結果より，CRF は未知の固有表現に対して有効であり，深層学習モデルは既知の固有表現に対して有効であることがわかった．本研究の主要な抽出対象である現地でしか購入できない土産名やその販売店舗の名前は，学習データ中に含まれないことがほとんどであるため，本研究においては CRF が有効である可能性が高い．

4.5 追加実験

表 3 と表 4 より，深層学習モデルは未知の固有表現に対する性能が低いことを確認した．そこで，以下の情報を深層学習モデルに追加し，未知の固有表現に対する性能変化を確認する．

- 品詞・品詞細分類
- ブログ記事から学習された分散表現

Aguilar らは，提案モデルに品詞情報を追加することで，ソーシャルメディア中のテキストから構築された WNUT2017 データセットに対する抽出性能が向上したと述べている [3]．また，4.4 節での実験においても，CRF の素性として品詞情報を利用することで，CRF の性能が向上した．そこで，深層学習モデルに品詞情報を組み込むことで，抽出性能が変化するかを確認する．

品詞情報を深層学習モデルで利用するため，品詞と品詞細分類でそれぞれランダムに初期化した品詞ベクトルを用意する．次元数は予備実験の結果より，品詞，品詞細分類でそれぞれ 5 とし，文字ベクトルと同様の方法で初期化する．品詞ベクトルは単語ベクトルや文字ベクトルを BiLSTM に入力する際に同時に入力し，モデルの学習とともに値を更新する．

表 5 と表 6 に品詞ベクトルを加えた時の実験結果を示す．表 5 と表 6 より，品詞ベクトルを追加した場合，BiLSTM-CNN-CRF モデルの抽出性能の向上が確認できた．特に，未知の品名に対する F 値が 5 ポイント，未知の店名に対する F 値が 8 ポイント向上した．しかし，他のモデルの性能の向上は見られなかった．

次に，単語分散表現を変更した場合の抽出性能の変化を確認する．ここまでの実験では，Wikipedia の本文を用いて事前学習された分散表現を単語ベクトルとして利用していた．分散表現のドメインをデータセットのドメインと一致させることで，性能がどのように変化するかを確認する．

表 5　品詞ベクトルを加えた時の品名に対する抽出性能
Table 5　Experimental results for PRO by neural models with POS embeddings

Model	all			unknown			known		
	precision	recall	f1-measure	precision	recall	f1-measure	precision	recall	f1-measure
Huang Model	67.37	60.28	63.36	43.85	40.33	41.05	84.90	74.60	79.18
BiLSTM-CNN-CRF	72.08	62.43	66.62	49.64	44.06	45.80	88.98	75.52	81.35
Lample Model	68.64	60.57	63.90	43.69	42.21	42.20	88.76	73.09	79.67
Char-BiLSTM-CRF	68.92	60.58	64.16	44.29	40.81	41.94	87.10	73.64	79.54

表 6　品詞ベクトルを加えた時の店名に対する抽出性能
Table 6　Experimental results for SHO by neural models with POS embeddings

Model	all			unknown			known		
	precision	recall	f1-measure	precision	recall	f1-measure	precision	recall	f1-measure
Huang Model	80.91	63.39	70.75	44.41	22.94	28.90	90.55	88.40	89.29
BiLSTM-CNN-CRF	80.72	67.06	72.81	52.28	30.94	36.41	92.37	88.91	90.51
Lample Model	81.20	64.31	71.33	47.55	24.83	31.46	92.00	88.43	89.98
Char-BiLSTM-CRF	81.97	67.77	73.89	56.46	30.90	37.55	91.78	89.38	90.32

実験のために，以下の 2 種類の分散表現を用いる．

- ブログ記事のみで学習した分散表現
- ブログ記事と日本語 Wikipedia の本文の両方で学習した分散表現

分散表現の学習に用いるブログ記事のテキストには，Yahoo!ブログ [2] に含まれる記事の本文とコメントを利用する．ブログ記事は 2005 年 1 月 1 日〜2014 年 12 月 31 日までに投稿された記事のうち，「土産」をクエリとして得られる記事を，日本語 Wikipedia の本文は 2019 年 5 月 27 日時点で得られるすべての記事に含まれるものを利用する．分散表現の学習に用いるパラメータは，表 2 に示すものを用いる．

表 7 に単語分散表現を変更した場合の実験結果を示す．3 種の分散表現の中で最も抽出性能が高いものを太字で示す．表 7 の all より，既知未知の区別をしない場合，日本語 Wikipedia の本文のみで学習した単語分散表現を採用したモデルが最高性能になることが多いといえる．また，未知の固有表現に対しては，ブログ記事のテキストで学習した分散表現を単語ベクトルに用いたモデルの性能が高くなることを確認できた．このことから，ブログドメインでのみ出現する語彙を分散表現に加えることで，未知の固有表現に対する深層学習モデルの性能が向上することがわかった．

追加実験の結果から，品詞情報の追加やデータセットと分散表現のドメイン統一により，深層学習モデルの性能が若干向上するが，未知固有表現に対する性能においては，CRF と深層学習モデルの間に大きな差があることがわかった．CRF の F 値と深層学習モデルの中で最も高い F 値を比較すると，CRF が未知の品名に対して約 8 ポイント，未知の店名に対して約 3 ポイント高い．

5. 考察

5.1 定量的な分析

図 6 と図 7 に CRF と深層学習モデルがそれぞれ抽出に失敗した品名と店名の長さの分布を示す．なお，深層学習モデルは BiLSTM-CNN-CRF とし，品詞ベクトルを加え，単語ベクトルに日本語 Wikipedia の本文のみで学習した分散表現を利用した時のモデルとする．固有表現の長さは固有表現に含まれる形態素数を用いて計測した．

図 6 より，CRF は深層学習モデルと比べて，長さの短い品名に対するミスが多いことがわかる．また，品名の長

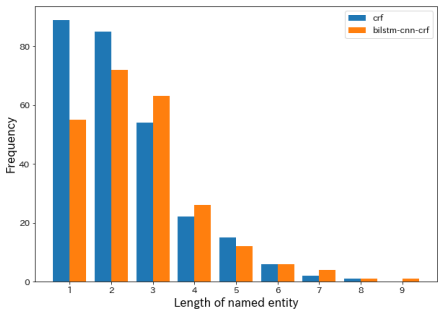

図 6　抽出に失敗した品名の長さの分布
Fig. 6　Distribution of the frequency of extraction errors in product entities by the number of words

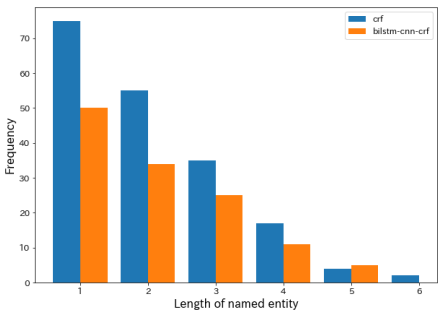

図 7　抽出に失敗した店名の長さの分布
Fig. 7　Distribution of the frequency of extraction errors in shop entities by the number of words

表 7　分散表現を変更した時の深層学習モデルでの実験結果（F 値）
Table 7　Experimental results by neural models changing pretrained word embedding

Model	all						unknown					
	PRO			SHO			PRO			SHO		
	Wiki	Blog	Blog+Wiki	Wiki	Blog	Blog+Wiki	Wiki	Blog	Blog+Wiki	Wiki	Blog	Blog+Wiki
Huang Model	**65.46**	65.12	64.67	**71.11**	70.38	69.65	43.71	**45.71**	43.18	25.43	**27.78**	24.04
BiLSTM-CNN-CRF	64.48	65.81	**65.82**	71.03	70.71	**71.65**	40.16	**45.08**	42.40	28.62	**32.69**	29.52
Lample Model	65.13	**65.44**	64.84	**73.20**	72.27	70.93	44.29	42.68	**44.90**	32.82	**33.45**	29.14
Char-BiLSTM-CRF	**65.01**	64.93	64.78	**74.35**	72.64	74.02	41.41	**44.27**	43.02	**38.38**	35.10	37.51

さが長くなると，CRF のミスの数が深層学習モデルとほとんど変わらなくなることがわかる．この結果から，深層学習モデルは，長さの短い品名の抽出に有効であるといえる．深層学習モデルは，Bidirectional-LSTM を利用することで，文中に存在するすべての単語情報を利用してラベリングできるため，このような結果になったと予想する．

また，図 7 より，店名の長さにかかわらず，深層学習モデルのミスが少ないことがわかる．深層学習モデルは，未知の固有表現に対する抽出性能が低く，既知の固有表現に対する性能が高い．このことから，店名に対する抽出ミスは固有表現の長さより，固有表現が学習データ中に出現するかどうかが大きく関わっているといえる．

5.2　抽出誤りの傾向分析

固有表現抽出誤りの傾向について分析する．固有表現抽出誤りを以下の 4 種に分類し，ミスの傾向を調査する．

- 固有表現に O タグを振る（NE2O）
- O タグに固有表現タグを振る（O2NE）
- ラベリングのクラスは正しいが，範囲を誤る（RangeMiss）
- ラベリングの範囲は正しいが，クラスを誤る（ClassMiss）

図 8 に CRF と深層学習モデルの固有表現抽出誤りの割合を示す．誤りの割合は，交差検証中の全ての抽出ミスで計測した．深層学習モデルは，5.1 の分析と同じ BiLSTM-CNN-CRF モデルとする．

図 8 より，CRF は，文中に固有表現が出現しても O タグを振ってしまうこと（NE2O）が多いことがわかる．一方，深層学習モデルは，CRF と比べて NE2O の割合が少なく，他の抽出ミスが多い．これらの結果より，CRF と深層学習モデルは，抽出誤りの傾向が異なり，それぞれの手法に適した文や固有表現にばらつきが存在すると考える．この傾向を利用し，中山ら [26] のように，複数の情報抽出モデルの抽出結果を用いたアンサンブル学習により，抽出性能の改善が見込まれる．

5.3　エラー分析

深層学習モデルでのみ抽出できた固有表現と CRF モデルでのみ抽出できた固有表現についてそれぞれ考察する．まず，深層学習モデルのみで抽出できた固有表現と固有表現を含む文の一例を表 8 に示す．深層学習モデルを用いることで，固有表現と“買う”や“菓子”などの手がかり語の距離が遠い文から固有表現抽出が可能になる場合を確認した．また，「“店名”（の）“品名”」のような文から固有表現を抽出できる場合も確認できた．

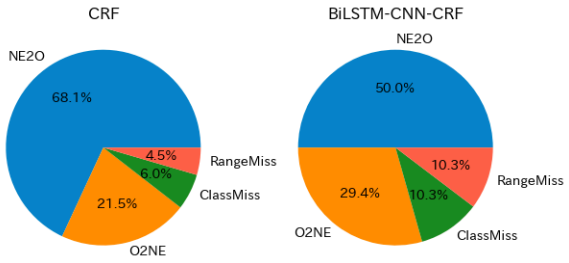

図 8　固有表現抽出誤りの割合
Fig. 8　Proportion of Misses of Named Entity Recognition

次に，CRF モデルでのみ抽出できた固有表現と固有表現を含む文の一例を表 9 に示す．深層学習モデルは長さが短い文や手がかり語が含まれない文から固有表現を抽出できない傾向が確認できた．また，固有表現自体に手がかりが含まれる場合でも抽出に失敗することがあった．例えば，“丸玉製菓”は“製菓”から店名の可能性が高いと識別することができる．CRF モデルは形態素の表記自体を素性に加えているため，文長が短い場合でも表記素性を利用して，固有表現抽出できる場合があった．このように，固有表現の表記に手がかりが含まれる場合であったとしても，深層学習モデルは，固有表現抽出に失敗することが多いことを確認した．

6.　おわりに

本稿では，ブログ記事から土産の品名と店名を抽出する手法を検討するため，CRF と深層学習モデルによる固有表現抽出の性能を比較した．実験結果から，CRF は学習データ中に含まれない未知の固有表現に対して有効であり，深層学習モデルは既知の固有表現に有効であることを確認した．

今後は，固有表現抽出によって抽出した品名や店名を利用して，品名や店名以外の土産情報を抽出する手法や，現地でしか購入できないことを判定する手法について検討する．その後，システムを実装するための土産情報をまとめたデータベースの構築を目指す．

表 8 深層学習モデルでのみ抽出できた固有表現の例
Table 8 Examples of named entities recognized only by the neural models

sentence	CRF	DL
桂月堂さんの「山川」と「若草」と「薄小倉」を購入いたしました。	桂月堂，薄小倉	桂月堂，山川，若草，薄小倉
他にもコーンチョコや六花亭の霜だたみ，雪やこんこも買いました．	六花亭，雪やこんこ	コーンチョコ，六花亭，霜だたみ，雪やこんこ
千葉県マザー牧場「カマンベールチーズフロランタン」チーズ味がしっかりとした焼き菓子	マザー牧場	マザー牧場，カマンベールチーズフロランタン

表 9 CRF でのみ抽出できた固有表現の例
Table 9 Examples of named entities recognized only by the CRF model

sentence	CRF	DL
丸玉製菓のが 1 番美味しいです	丸玉製菓	なし
ルーヴ／讃岐の岐三（きさん） みるくつつみ饅頭	ルーヴ，みるくつつみ饅頭	ルーヴ
バイオリンタウンにある和菓子屋「藤屋」	藤屋	なし

参考文献

[1] OMIYA! https://omiyadata.com/jp/.
[2] Yahoo!ブログ (2019 年 12 月 15 日サービス終了予定). https://blogs.yahoo.co.jp/.
[3] Aguilar, G., López Monroy, A. P., González, F. and Solorio, T.: Modeling Noisiness to Recognize Named Entities using Multitask Neural Networks on Social Media, *Proceedings of the NAACL*, pp. 1401–1412 (2018).
[4] Akbik, A., Blythe, D. and Vollgraf, R.: Contextual String Embeddings for Sequence Labeling, *Proceedings of the COLING*, pp. 1638–1649 (2018).
[5] Augenstein, I., Derczynski, L. and Bontcheva, K.: Generalisation in Named Entity Recognition: A Quantitative Analysis, *Computer Speech & Language*, Vol. 44, pp. 61–83 (2017).
[6] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the NAACL* (2019).
[7] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M. and Zettlemoyer, L.: AllenNLP: A Deep Semantic Natural Language Processing Platform, *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pp. 1–6 (2018).
[8] Huang, Z., Xu, W. and Yu, K.: Bidirectional lstm-crf models for sequence tagging, *arXiv: 1508.01991* (2015).
[9] Iwakura, T.: A Named Entity Recognition Method using Rules Acquired from Unlabeled Data, *Proceedings of the RANLP*, pp. 170–177 (2011).
[10] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of EMNLP*, pp. 230–237 (2004).
[11] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C.: Neural Architectures for Named Entity Recognition, *Proceedings of the NAACL*, pp. 260–270 (2016).
[12] Li, Y., Bontcheva, K. and Cunningham, H.: SVM Based Learning System For Information Extraction, *Proceedings of Sheffield Machine Learning Workshop, Lecture Notes in Computer Science* (2005).
[13] Ma, X. and Hovy, E.: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, *Proceedings of the 54th Annual Meeting of the ACL*, pp. 1064–1074 (2016).
[14] Mai, K., Pham, T.-H., Nguyen, M. T., Nguyen, T. D., Bollegala, D., Sasano, R. and Sekine, S.: An Empirical Study on Fine-Grained Named Entity Recognition, *Proceedings of the COLING*, pp. 711–722 (2018).
[15] Misawa, S., Taniguchi, M., Miura, Y. and Ohkuma, T.: Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition, *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp. 97–102 (2017).
[16] Nagao, N. and Ando, K.: Extraction of Product Names for Constructing a Database of Souvenir Information, *Proceedings of the fifth International Conference on Informatics and Applications*, pp. 88–96 (2016).
[17] Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007).
[18] Passos, A., Kumar, V. and McCallum, A.: Lexicon Infused Phrase Embeddings for Named Entity Resolution, *Proceedings of the CoNLL*, pp. 78–86 (2014).
[19] Ratinov, L. and Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition, *Proceedings of the CoNLL*, Stroudsburg, PA, USA, pp. 147–155 (2009).
[20] Ritter, A., Clark, S., Mausam and Etzioni, O.: Named Entity Recognition in Tweets: An Experimental Study, *Proceedings of the EMNLP*, pp. 1524–1534 (2011).
[21] Sasano, R. and Kurohashi, S.: Japanese Named Entity Recognition Using Structural Natural Language Processing, *Proceedings of the IJCNLP*, pp. 607–612 (2008).
[22] Tjong Kim Sang, E. F. and De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: LanguageIndependent Named Entity Recognition, *Proceedings of the CoNLL*, pp. 142–147 (2003).
[23] Yang, J., Liang, S. and Zhang, Y.: Design Challenges and Misconceptions in Neural Sequence Labeling, *Proceedings of the COLING* (2018).
[24] 新堂安孝，友利　涼，兼村厚範，宮尾祐介，森　信介：日本語の食べ物・飲み物表現の抽出における文字 CNN の効果，言語処理学会第 25 回年次大会発表論文集，pp. 1399–1402 (2019).
[25] 福島健一，鍛治伸裕，喜連川優：日本語固有表現抽出における超大規模ウェブテキストの利用，電子情報通信学会第 19 回データ工学ワークショップ論文集 (2008).
[26] 中山功太，小林暁雄，関根　聡：共有タスクにおける GA 重み付け加重投票を用いた属性値アンサンブル，言語処理学会第 25 回年次大会発表論文集，pp. 1547–1550 (2019).
[27] 池田流弥，安藤一秋：ブログ記事からの土産の品名・店名抽出，人工知能学会第 32 回全国大会論文集, 1E302 (2018).