深層学習によるブログ記事からの土産の品名・店名抽出

池田 流弥　　安藤 一秋
香川大学大学院 工学研究科　香川大学 創造工学部
s18g454@stu.kagawa-u.ac.jp ando@eng.kagawa-u.ac.jp

1 はじめに

旅行の際，9 割以上の人が土産を購入するというアンケート結果 [1] がある．また，土産を選ぶ際，その場所でしか手に入らないものが重視されるという結果も得られている．オンラインショップの普及により，多種多様な商品が手軽に購入できるようになり，現地でしか購入できない土産の需要が高まっている．Web 上には，土産情報を提供する各種の Web サービス [2] が存在するが，「現地でしか購入できない」という情報はほとんど提供されていない．また，既存のサービスで情報が提供されている土産のほとんどはオンラインショップで購入できるものである．そこで，本研究では，現地でしか購入できない土産に関する情報を Web 上から自動で収集・整理し，ユーザに提示するシステムの構築を目的とする．なお，本研究では，菓子類の土産のみを対象とする．

システムを構築するために，現地でしか購入できない土産の情報を収集する必要がある．これらの情報はブログ記事や Q&A サイトなどの Web 上に散在している．しかし，テキスト中の土産名や販売店舗名が特定できなければ，土産情報を活用することができない．そのため我々は，これまでの研究において，CRF（Conditional Random Fields）による固有表現抽出手法を用いて，ブログ記事から土産の品名，販売店舗名を抽出する手法を提案した [3]．本稿では，固有表現抽出で高い性能を報告している深層学習に基づくモデルを利用して，土産の品名，販売店舗名を抽出する手法を提案し，その性能について評価する．

2 CRF による固有表現抽出を用いた土産の品名・店名抽出手法

我々の先行研究では，土産の品名と販売店舗名が固有表現であることに注目し，CRF により，ブログ記事から土産の品名と販売店舗名を抽出する手法を提案した [3]．先行研究の手法では，土産情報を含むブログ記事中の文を形態素解析し，系列ごとに品名 (PRO)，店名 (SHO) のタグを付与することで学習データを作成する．品名タグは食品名に，店名タグは食品を販売している店舗名に付与する．

Yahoo ブログの菓子・デザートカテゴリを対象に土産情報の書かれた記事を 373 件収集し，記事中の 5,170 文に対して人手でタグ付けし，実験を行った．10 分割交差検証により評価を行った結果，F 値で未知の品名に対して 0.605，未知の店名に対して 0.519 という結果を得た．

また，実験結果に対してエラー分析を行った結果，抽出誤りの約 6 割が文中に品名・店名が存在するが，品名・店名のタグが付与できないものであることがわかった．これらの多くは，手がかり語 (買う，貰う，食べる等) が文中に出現しない，手がかり語と固有表現の距離が遠いなどが原因であると考えられる．CRF による固有表現抽出では，windowsize で指定した範囲の形態素のみを参照し，固有表現タグを決定する．そのため，手がかり語と固有表現の距離が遠い場合，手がかり語をラベリングに活用できない．深層学習による固有表現抽出モデルを用いることで，このような文に対応できると考える．

3 深層学習による固有表現抽出を用いた土産の品名・店名抽出手法

3.1 深層学習による固有表現抽出モデル

本稿では，深層学習による固有表現抽出モデルを用いて，ブログ記事から土産の品名・販売店舗名を抽出する手法を提案する．深層学習による固有表現抽出では，Lample らの提案モデル [4] に代表される BLSTM（Bidirectional LSTM）と CRF を組み合わせた BLSTM-CRF が高い性能を報告している．Ma らは文字の分散表現を CNN の入力として与えて得た文字表現を単語分散表現と結合し，BLSTM-CRF の入力とするモデル [5] を提案し，CoNLL-2003 データセットに対して F 値で 0.912 という結果を報告している．Misawa らは Ma らの提案モデルを日本語に適応する場合，次の問題があることを示している [6].

- 日本語には単語境界がないため，単語単位で固有表現抽出する場合，形態素解析を行う必要があり，形態素の一部に固有表現が含まれている場合，固有表現抽出ミスを引き起こす．
- 英語に比べ日本語は，単語長が短い傾向があるため，CNN による文字表現の抽出は適していない．

上記の問題を踏まえて，Misawa らは図 1 のモデルを提案している [6]．Misawa らの提案モデルは，文字

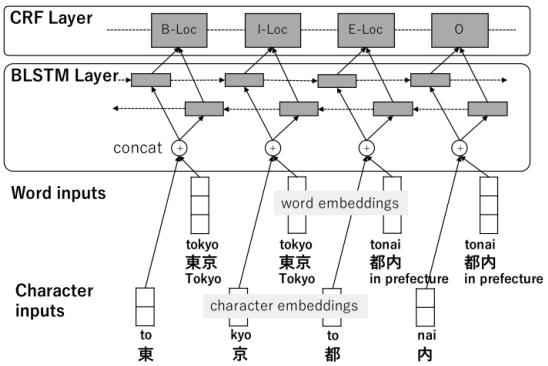

図 1: Misawa らの提案モデル（[6] より引用）

単位で固有表現タグを推定するモデルである．文字分散表現（Character inputs）と単語分散表現（Word inputs）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構成となっている．Misawa らはこのモデルで毎日新聞コーパスに対して実験を行い，既存のモデルを用いた場合と比べて，F 値で 1 ポイントほど高い 0.881 という性能を得たことを報告している．

本稿では，深層学習による固有表現抽出モデルとして，日本語に対する性能が最も高かった Misawa らのモデルを採用する．

3.2 学習データの作成方法

本稿で，固有表現抽出に利用する学習データは次の手順で作成する．

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．タグの形式には IOB2 タグ形式を用いる．
- 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を付与する．

4 評価実験

4.1 評価方法

先行研究 [3] で提案した CRF モデルと深層学習モデルの 2 つで行い，性能を比較する．なお，深層学習モデルでは，word2vec[7] と fastText[8] で学習した分散表現を用いて，その性能も比較する．適合率，再現率，F 値を評価尺度とし，5 分割交差検証で抽出性能を評価する．人手でタグ付けした結果とラベリングされた結果を比較し，完全一致した場合のみを正解と判断する．

本実験では，未知の固有表現（学習データに含まれない固有表現）に対する評価も行う．学習データ中に含まれている固有表現の場合，固有表現の表層文字列を学習するため，性能が高くなる傾向がある [9]．本研究では，現地でしか購入できない土産の品名と販売店舗名が主要な抽出対象であるため，学習データ中に含まれていない未知の固有表現に対する性能が重要になる．特に，現地でしか購入できない土産の品名・店名は，オンラインショップで購入できる土産と比べて出現しにくいため，本稿では再現率を重視する．

4.2 実験データ

本実験において，形態素解析には MeCab[1] を用い，辞書には IPADIC を利用する．実験には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![2] 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!ブログの菓子・デザートカテゴリ内でヒットしたブログ記事の本文を用いた．収集したブログ記事の中でランダムに 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．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，文中に名詞と助詞を含み，動詞，形容詞，助動詞のいずれかを含む 9,488 文を実験データに用いることとした．なお，形態素解析の結果，全角記号に対する品詞が“名詞”と推定される場合が多数確認されたため，全角記号に対する品詞を修正した．実験データ中に品名は 1,939 件，店名は 1,148 件含まれており，そのうち品名は 1,226 種，店名は 625 種であった．

4.3 学習に用いるパラメータ

4.3.1 CRF モデルの素性とパラメータ

先行研究 [3] の結果より，CRF は単語単位に対してラベリングするモデルを採用し，windowsize は 2 とする．CRF の実装には CRFsuite[2] を用い，ハイパーパラメータはデフォルト値を用いる．また，素性としては以下を利用する．

- 表記
- 文字種
- 品詞
- 括弧内の単語にフラグを立てる

[1] http://taku910.github.io/mecab/
[2] http://www.chokkan.org/software/crfsuite/

4.3.2 深層学習モデルの分散表現とパラメータ

深層学習モデルの入力となる分散表現の構築には，12/20 時点での Wikipedia の本文全てを利用する．単語分散表現は word2vec で学習したものと fastText で学習させたものを用意し比較する．word2vec の学習および fastText の学習には公開されているスクリプトを用いる[3] [4]．word2vec の学習に用いたパラメータを表 1，fastText の学習の用いたパラメータを表 2 に示す．

表 1: word2vec 実行時のパラメータ

-cbow	1
-size	50, 300, 500
-window	10
-negative	10
-min-count	5
-alpha	0.001
-iter	10

表 2: fastText 実行時のパラメータ

skipgram	
-dim	300
-ws	10
-neg	10
-minCount	5
-lr	0.001
-epoc	10
-loss	ns

分散表現の次元数については，Melamud らの研究 [10] と Misawa らの研究 [6] を参考に決定した．Melamud らは，単語分散表現を英語の自然言語処理タスクに適応させる際に次元数や windowsize を変更させた時の性能変化について実験し，固有表現抽出では単語分散表現の次元数が 50 次元の時にもっとも性能がよくなることを報告している．その結果を参考に，本稿でも 50 次元の分散表現を用意した．また，Misawa らが実験で用いた 500 次元の分散表現も用意した．文字分散表現は word2vec で学習したものを用意し，次元数は，Misawa らが用いている 50 次元とした．

深層学習モデルの学習に用いるパラメータは，エポック数が 50，バッチサイズが 32，BLSTM の隠れ層の次元数は 300，層の数は 1，dropout を 0.5 とした．最適化手法には Adam を用い，学習率は 0.001，epsilon は 1e-8 とした．

4.4 実験結果

実験結果を表 3 に示す．w50 はモデルの入力に word2vec で学習した 50 次元の単語分散表現を，f300 は fastText で学習した 300 次元の単語分散表現を用いた結果を表す．また，太字はすべてのモデルの中で，下線は深層学習モデルで最も性能が高かったものを示す．

既知未知を区別しない場合および未知の固有表現に対してのみ評価した場合ともに，先行研究 [3] で提案した CRF モデルの性能が高くなることを確認した．ただし，未知の店名に対する再現率は CRF モデルが最も低く，f300 の場合に最も性能が高くなった．また，未知の品名に対する再現率に関しても CRF モデルと深層学習モデルで大きな差はなく，チューニング次第で再現率は深層学習モデルが CRF モデルより優位になる可能性がある．

深層学習モデルの入力に用いる分散表現については，fastText を用いた場合が全体的に性能が高くなることを確認した．subword を活用することが性能に影響している可能性がある．また，word2vec で学習した分散表現を使用した場合では，50 次元を使用する場合が全体的に性能が高かった．

5 考察

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以下に示す．

- コーンチョコ，霜だたみ
  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．
- カマンベールチーズフロランタン
  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
- たまきや
  名張にある赤目四十八滝の入り口付近にお店を構える，たまきやの「へこきまんじゅう」です．
- 松翁軒
  年末に実家に帰省予定だったので，松翁軒のカステラは祖母へ

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．

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次に示す．

- みるくつつみ饅頭
  ルーヴ／讃岐の岐三（きさん） みるくつつみ饅頭
- フルーツな巣ごもりたち
  「フルーツな巣ごもりたち」というアウトテイク
- 丸玉製菓
  丸玉製菓のが 1 番美味しいです

[3] https://code.google.com/archive/p/word2vec/
[4] https://github.com/facebookresearch/fastText

表 3: 実験結果

	区別なし						未知のみ					
	PRO			SHO			PRO			SHO		
	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1
CRF	**0.731**	**0.567**	**0.639**	**0.805**	**0.632**	**0.708**	**0.591**	**0.430**	**0.492**	**0.571**	0.397	**0.468**
w50	0.631	0.513	0.566	0.680	0.572	0.621	0.456	0.429	0.442	0.430	0.426	0.428
w300	0.628	0.490	0.550	0.669	0.588	0.626	0.461	0.399	0.428	0.408	0.428	0.418
w500	0.629	0.510	0.563	0.675	0.560	0.612	0.463	0.429	0.445	0.418	0.402	0.410
f300	0.647	0.514	0.573	0.664	0.584	0.621	0.478	0.413	0.443	0.420	**0.450**	0.434

- 本郷三原堂
  「本郷三原堂」 本郷三丁目駅近くにある和菓子屋さんになります.

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．一方，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固有表現抽出できる場合があった．これらの文に対する性能向上が今後の課題となる．

6 おわりに

本稿では，深層学習による固有表現抽出手法を用いて，ブログ記事から土産の品名・販売店舗名を抽出する手法を提案し，その性能について評価した．性能については，深層学習モデルが CRF モデルを下回ったが，考察により，深層学習を用いることで，固有表現と手がかり語の距離が遠い文からも固有表現抽出できる場合があることを確認した．

今後は，分散表現のチューニングやモデルを拡張し，抽出性能の向上を目指す．本研究の対象としているブログ記事は新聞記事などに比べて文体が崩れやすいため．固有表現抽出することは難しい．Aguilar ら [11] が提案したようなソーシャルメディアに対して固有表現抽出するモデルなどを参考に拡張を行いたい．その後，品名・店名以外の土産情報の抽出手法，現地でしか購入できないことの判定手法について検討し，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．

参考文献

[1] アサヒグループホールディングス. ハピ研: 毎週アンケート 第 641 回. https://www.asahigroup-holdings.com/company/research/hapiken/maian/201707/00641/.

[2] OMIYA! https://omiyadata.com/jp/.

[3] 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.

[4] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proc. of the NAACL*, pp. 260–270, 2016.

[5] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proc. of the 54th Annual Meeting of the ACL*, pp. 1064–1074, 2016.

[6] Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. Character-based bidirectional lstm-crf with words and characters for japanese named entity recognition. In *Proc. of the First Workshop on Subword and Character Level Models in NLP*, pp. 97–102. ACL, 2017.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *arXiv: 1301.3781*, 2013.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv: 1607.04606*, 2016.

[9] 福島健一. 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/第 6 回日本データベース学会年次大会, 2008.

[10] Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. The role of context types and dimensionality in learning word embeddings. In *Proc. of the NAACL*, pp. 1030–1040, 2016.

[11] Gustavo Aguilar, Adrian Pastor López Monroy, Fabio González, and Thamar Solorio. Modeling noisiness to recognize named entities using multitask neural networks on social media. In *Proc. of the NAACL*, pp. 1401–1412, 2018.