E-003

CRF によるブログ記事からの土産の品名・店名抽出法のための素性検討
Feature Engineering for a CRF based Method to Extract Product and Shop Names of Souvenirs from Blog Articles

池田 流弥† 安藤 一秋‡
Ryuya Ikeda Kazuaki Ando

1. はじめに

土産に関するアンケート調査[1]より，旅行の際，9 割以上の人が土産を購入することがわかっている．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ことも確認されている．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，オンラインショップで購入できないような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，上記のサービスで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は，まず，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，テキスト中から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，本研究では，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の検討を進めている．本稿では，先行研究[2]に引き続き，CRF（Conditional Random Fields）を用いて，ブログ記事から土産の品名と販売店舗名を抽出するための素性について検討する．

2. ブログ記事からの土産の品名・店名抽出手法

本研究では，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，固有表現抽出手法により，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[2]．固有表現抽出には，CRF による系列ラベリングを用いる．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），O（その他）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．

先行研究[2]の実験結果に対してエラー分析した結果，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が多いことがわかった．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．本稿では，この問題を改善するため，新しい素性・前処理について検討する．

3. 大域的情報を活用した素性・前処理の提案

本稿では，固有表現の抽出性能を向上させることを目的に，2 つの観点から素性を検討する．3.1 節では，手がかり語が文中に出現しない構造について，3.2 節では，手がかり語が文中に出現するが，固有表現と手がかり語の距離が遠い場合について，それぞれ対応できる素性・前処理を検討する．

3.1 HTML 構造を活用した素性

土産情報が記載されているブログ記事では，土産の品名や販売店舗名のみが書かれている文が度々見られる．そのような文の前後には土産や販売店舗の画像が掲載されていることが多い．つまり，土産の品名や販売店舗名だけで構成される文は，画像のキャプションとしての役割を果たしていると考えられる．土産の品名が画像のキャプションとして活用されている例を図 1 に示す．このような構造に注目し，文の前に画像が出現するかどうか，文の後ろに画像が出現するかどうかの 2 つの素性（beforeImg と afterImg）について有用性を検討する．これらの素性により，画像のキャプションとして書かれる文だけでなく，一般的な文についても「文の前後に画像がある場合，文中に土産の品名，販売店舗名が含まれやすい」という構造を学習できることが期待できる．ただし，これらの素性を利用したとしても，CRF を用いて，画像のキャプションとして書かれている文から固有表現を抽出することは難しい．そのため，CRF 以外の手法を検討する必要がある．

また，本文やキャプションには店舗のホームページへのリンクが張られていることもある．これにも注目し，リンクが貼られているかどうかという素性（URL）についても有用性を検討する．

図 1．画像のキャプションとして文が活用されている例
（https://blogs.yahoo.co.jp/norahirano/36550793.html より）

3.2 文構造に注目した素性・前処理

手がかり語が文中にあるにも関わらず，固有表現抽出ができないものの多くは，手がかり語と固有表現の距離が遠いという傾向がある．CRF による系列ラベリングでは，1 つの系列に対してタグを付与する時，タグを付与する系列の他に，前後 n 系列の情報も活用する．固有表現と手がかり語の距離が遠い場合，参照する n 系列内に手がかり語が含まれないため，手がかり語を活用できていない．n を大

† 香川大学大学院工学研究科 Graduate School of Engineering, Kagawa University
‡ 香川大学創造工学部 Faculty of Engineering and Design, Kagawa University

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

きくすることも可能であるが，本手法では，*n* を大きくすることで抽出性能が低下することを確認した．

距離が遠い系列の情報を活用する手法として，RNN（Recurrent Neural Network）などの深層学習と CRF を組み合わせた手法[3]や係り受け情報を活用した手法[4]が提案されている．本手法では後者を参考に，係り先の文節を利用した素性（dependWord）について有用性を検討する．具体的には，係り先文節の先頭系列の表記を係り元文節中の系列すべてに素性として組み込む．

また，形態素解析では，複合名詞が複数の系列に分けられる場合がある．これらを 1 つの形態素とすることで，手がかり語と固有表現の距離が近くなると考える．そこで，前処理（comNoun）として，名詞が連続で出現した場合，それらを 1 つの系列にまとめ，品詞を「名詞-複合名詞」とすることの有用性を検討する．

4. 評価実験

提案素性の有効性を確認するために，評価実験を行う．

4.1 実験環境・方法

本実験において，形態素解析機には MeCab を，係り受け解析機には CaboCha を用い，辞書には IPADIC を利用する．CRF の実装には CRFsuite を用い，ハイパーパラメータはデフォルト値を用いる．

実験データは，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と画像タグ，URL タグとする．収集した 380 エントリ，7,328 文を形態素解析し，系列ごとに人手で固有表現タグを付与したものを実験データに用いる．タグ付けは IOB2 タグ形式で行う．実験データ内の固有表現数は品名が 2,116，店名が 1,352 となった．

素性として，形態素の表記，文字種，品詞細分類の 3 つを活用したものをベースライン（base）とする．ベースラインに 3 章で示した 4 つの素性と 1 つの前処理を 1 つずつ追加し，抽出性能を評価する．

4.2 評価方法

適合率，再現率，F 値を評価尺度とし，10 分割交差検証で抽出性能を評価する．品名と店名を分けて評価する．

本実験では，未知の固有表現に対する抽出性能のみを評価する．既知の固有表現の場合，固有表現の表層文字列を学習するため，未知の固有表現より性能が高くなる傾向がある．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，学習データに含まれていない固有表現に対する性能を重視する．

4.3 実験結果

品名に対する実験結果を表 1 に，店名に対する実験結果を表 2 に示す．表 1，2 において，赤いセルはベースラインと比べて性能が向上した部分である．

表 1，2 に示すように，いずれの素性・前処理についても，性能向上は僅かであった．全体的に僅かに性能が向上した素性として，beforeImg，afterImg，URL が挙げられる．beforeImg，afterImg を追加することで，品名，店名のいずれについても適合率が向上した．菓子・デザートカテゴリのブログ記事において，画像の前後に出現する文には，土産の品名・販売店舗名が書かれやすいことが性能向上に寄与したと考えられる．また，URL を追加したときは，適合率，再現率ともに上昇していることが確認できる．特に，店名の適合率は，約 1 ポイント向上した．リンクは，品名に比べ，店名に対して張られることが多いため，店名の適合率の向上に寄与したと考えられる．

dependWord を追加することで，品名に対する性能が約 1 ポイント低下し，店名に対する性能が約 2 ポイント向上した．店名を含む文は，文として成立している場合が多いため，係り受け情報の組み込みがうまく働き，結果に影響したと考えられる．今後は，品名の抽出性能を向上させるために，素性に係り受け情報を組み込む方法を検討する必要がある．

ComNoun を追加した場合，品名に対する適合率が若干向上したが，その他の性能は逆に低下した．特に，店名に対する性能が低下した理由として，「堂」や「屋」など，店名の末尾に使われやすい単語が複合名詞化されることで，固有表現の終了を示す特徴として活用できなくなるため，この点が影響していると考えられる．

表 1．未知の品名に対しての実験結果

	precision	recall	f1
base	0.613	0.492	0.549
beforeImg	0.623	0.491	0.549
afterImg	0.621	0.494	0.551
URL	0.614	0.496	0.549
dependWord	0.599	0.480	0.533
comNoun	0.615	0.441	0.514

表 2．未知の店名に対しての実験結果

	precision	recall	f1
base	0.557	0.452	0.499
beforeImg	0.561	0.452	0.501
afterImg	0.562	0.450	0.499
URL	0.569	0.454	0.505
dependWord	0.594	0.467	0.523
comNoun	0.551	0.268	0.361

5. おわりに

本稿では，CRF を用いて，ブログ記事から土産の品名と店名を抽出する手法に活用する 5 つの素性・前処理について検討した．実験の結果，いずれの素性・前処理についても，性能向上は僅かであった．

今後の課題として，素性に係り受け情報を組み込む方法や，これまでに提案した素性の妥当な組み合わせなどについて検討する．そして，土産名と販売店舗名以外の土産情報を抽出する方法を検討し，システムの実装を目指す．

参考文献

[1] @nifty ニュース なんでも調査団 お土産についてのアンケート http://chosa.nifty.com/travel/chosa_report_A20140221/?theme=A20140221&theme=A20140221&theme=A20140221&theme=A20140221

[2] 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(JSAI2018), 1E3-02, (2018).

[3] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C., “Neural architectures for named entity recognition”, Proceedings of the NAACL-HLT 2016, pp.260-290, (2016).

[4] 笹野 遼平, 黒橋 禎夫, “大域的情報を用いた日本語固有表現認識”, 情報処理学会論文誌, Vol.43, No. 1, pp. 44-53, (2008).

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.