The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ブログ記事からの土産の品名・店名抽出

Extraction of Product and Shop Names of Souvenirs from Blog Articles

池田 流弥 *1
Ryuya Ikeda

安藤 一秋 *2
Kazuaki Ando

*1 香川大学大学院工学研究科
Graduate school of Engineering, Kagawa University

*2 香川大学工学部
Faculty of Engineering, Kagawa University

Demand for souvenirs that can be purchased only at the particular location or area is increasing, because anyone can purchase various souvenirs on online shopping sites. Such souvenirs are called local limited souvenirs in this paper. However, there are no Web sites and services that collected information about local limited souvenirs. The purpose of this study is to construct a system that presents useful information about local limited souvenirs, in order to provide support for selecting souvenirs. This paper proposes a method for extracting names of souvenirs and shops from blog articles using CRF (Conditional Random Fields). The effectiveness of the proposed method was confirmed by evaluation experiments.

1. はじめに

土産に関するアンケート調査[アサヒ 17]により，旅行に行った際，9 割以上の人が土産を購入することが確認されている．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．その理由として，オンラインショップの普及や交通網，運送業などが発達することで，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が注目されるようになったと考えられる．しかし，現地でしか購入できない土産に関する情報を一元的に提供している Web サイトやサービスは存在しない．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．

現地でしか購入できない土産の情報は，Web 上に散在している．しかし，全国各地で販売されている土産の品名を網羅したリストは存在しないため，現地でしか購入できない土産に関する情報を抽出する手がかりとして利用できない．そのため，旅行記や口コミなどが書かれたテキストから土産の品名を自動抽出する技術が必要となる．

本稿では，CRF（Conditional Random Fields）を利用して，ブログ記事から土産の品名と店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. 土産情報提示システム

本研究では，現地でしか購入できない土産に関する有用な情報を提示する土産情報提示システム[Nagao 17]の構築を進めている．本システムは，Q&A サイトやブログ記事から土産情報（土産の品名，販売店舗名，金額，評判など）を収集する．「現地でしか購入できない」という情報については，検索エンジンやオンラインショップを利用して土産の購入難易度（レア度）として判定する．ユーザへの情報提示法は，以下の 2 つがある．

① 地名や品名などから検索する方法
② 地図を利用して検索する方法

①では，リスト形式で情報を提示する．ユーザが最も重視する要素で土産情報を並び替えることにより，土産の選定を手助けできる．②では，地図上に土産の販売場所の位置や現在地をプロットすることにより，土産情報を提示する．この機能により，現地でしか購入できない土産の販売場所までの移動を手助けできる．

また，長尾ら[Nagao 17]は，土産情報のうち，Q&A サイトやブログ記事から土産名を収集する手法として，正規表現による土産名候補の抽出と SVM（Support Vector Machine）による品名の妥当性判定を組み合わせた手法を提案している．評価実験により，土産名候補の適合率は 0.05，再現率は 0.52，SVM による品名の妥当性判定では 0.82 という結果が得られている．

しかし，土産名抽出手法の性能は十分であるとはいえない．土産名抽出の性能が低ければ，他の土産情報を収集する際にも影響が出る．そこで，本稿では，土産名の新しい抽出手法を提案する．

3. 土産名抽出の関連研究

土産名を Q&A サイトやブログ記事から収集する関連研究について説明する．

3.1 川野らの研究

川野らの研究[川野 15]では，Q&A サイトから土産名を地域別に収集するツールの構築を目的に，入手地域と土産名の候補になる形態素 n-gram を抽出する手法を提案している．川野らの手法では，初めに，Q&A サイトの土産名が含まれている質問に対するベストアンサーを取得し，形態素解析する．その後，解析結果の形態素列から 1~5 までの形態素 n-gram 群を生成し，土産名になりえない形態素 n-gram をパターンマッチにより除去する．そして，残った形態素 n-gram に対して残差 IDF で重み付けし，スコアが高いものほど土産の可能性が高いと判断する．評価実験では，スコアが高いもののみを土産名として抽出した場合，適合率が約 0.1，再現率が約 0.5 という結果が得られている．

3.2 石野らの研究

石野らの研究[石野 10]では，旅行ブログエントリから自動的に観光情報を収集するための手法を提案している．観光情報としては，土産名と観光名所を対象としている．石野らの手法では，旅行ブログを検出した後，観光情報を抽出する．旅行ブログの判定および観光情報の抽出には，CRF による系列ラベリングを用いている．評価実験により，旅行ブログの検出性能は，適合率 0.87，再現率 0.38，観光情報の抽出性能は，出現頻度に

連絡先：池田流弥，香川大学大学院工学研究科
〒761-0396 香川県高松市林町 2217-20 安藤研究室，
rikeda71@gmail.com

より 100 位までにランク付けされた土産名において精度 0.74, 観光名所において精度 0.71 が得られている.

4. 土産の品名・店名抽出手法

本研究では, 土産の品名が固有表現であること注目し, 固有表現抽出により, Q&A サイトやブログ記事から土産の品名を抽出する手法を提案する. 固有表現抽出には, 石野らの研究を参考に, CRF による系列ラベリングを採用する. また, 土産の販売店舗名も固有表現であることから, 提案手法では, 土産の品名と販売店舗名を抽出対象とする. なお, 抽出対象とする品名は, 菓子類の土産のみとする.

4.1 学習データの作成法

CRF に用いる学習データは次の手順で作成する.

① Q&A サイトやブログから土産について書かれたテキストを収集し, 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.

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

ここで, 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.

IOB2 タグ形式での各タグの意味を表 1 に示す. 固有表現の始まりである系列には B タグを付与する. 系列が固有表現の途中であれば I タグを付与する. 固有表現でない系列には O タグを付与する.

表 1. IOB2 タグ形式でのタグの意味

タグ	意味
I	系列で固有表現が継続している
O	系列が固有表現でないことを示す
B	系列が固有表現の始まりであることを示す

4.2 ラベリング手法

系列ラベリングの手法として, 矢野らの研究[矢野 17]を参考に, 単語ベース手法(word 手法)と文字ベース手法(char 手法)の 2 種類を提案し, 性能を比較する. word 手法のラベリングのイメージを図 1 に, char 手法のイメージを図 2 に示す. 図 1, 2 に示すように, word 手法は形態素単位で, char 手法は文字単位でラベリングする. word 手法では, 品詞情報をラベリングに活用できるといる利点がある. char 手法では, 形態素解析の誤りに影響されないという利点がある.

友達	に	白い	恋人	を	貰い	まし	た
O	O	B-PRO	I-PRO	O	O	O	O

図 1. word 手法でのラベリングイメージ

友	達	に	白	い	恋	人	を	貰	い	ま	し	た
O	O	O	B-PRO	I-PRO	I-PRO	I-PRO	O	O	O	O	O	O

図 2. char 手法でのラベリングイメージ

4.3 ベースライン素性

本稿では, 以下の 3 つを形態素(文字)に対する CRF のベースライン素性とする.

- 表記
- 文字種
- 品詞細分類

文字種は, アラビア数字, 英字小文字, 英字大文字, ひらがな, カタカナ, 漢字, その他の 7 種類とする. 品詞細分類は形態素解析で獲得できる品詞を利用する. word 手法では, 参照している形態素と前後 2 形態素について, 上記 3 つの素性を利用する. char 手法では, 品詞情報は活用できないため, 参照している形態素と前後 2 形態素について, 表記と文字種のみを素性として利用する.

4.4 追加素性・前処理・ラベリングの変更

ベースライン素性に, 以下の(1)〜(3)の素性と(4), (5)の前処理, (6), (7)のラベリングの変更を個別に追加し, 有効性を確認する.

(1) 店名のみをラベリングした後, 店名を含む文にフラグを立て, 品名のみのラベリングの素性に利用(2exec)
(2) チョコ, クッキーなどの一般的な食品名になりえる単語にフラグを立て, 素性に利用(comName)
(3) 括弧内の単語にフラグを立て, 素性に利用(inBracket)
(4) 括弧のタグを O タグに変更する(exBracket)
(5) 形態素解析の品詞間違いを修正する(correct)
(6) 文末からラベリングする(backward)
(7) BIOES タグ形式に変更する(BIOES)

(1)は, 文中で店名が出現したとき, 品名が共起しやすいことを利用した素性である.

(2)は, 一般名と商品名を区別した学習を期待した素性である.

(3)と(4)は, ブログ記事で品名や店名が「」や()などの括弧中に書かれやすいことに注目した素性と前処理である.

(5)の前処理は, 品詞の素性をより正確に利用できると考えて導入する.

(6)の変更は, 文末から文頭に向かってラベリングすることにより, 動詞の情報を活用して学習することを期待して導入する. 例えば, 「食べる」や「貰う」といった動詞が出現する前には, 商品が出現しやすいという傾向を活用できると考える.

(7)の変更は, 先行研究[Ratinov 09]で, BIOES タグ形式の有効性が報告されており, タグ形式による違いを確認するため導入する. BIOES タグ形式でのタグの意味を表 2 に示す. 表中の B, I, O タグは, IOB2 タグ形式と同様の意味を持つ. IOB2 タグ形式との差異は, 固有表現の終わりと単一の系列(1 形態素もしくは 1 文字)で成立する固有表現を区別できる点である.

なお, char 手法では, (2)と(5)を素性として組み込むことができないため, それら以外を用いて評価する.

表 2. BIOES タグ形式でのタグの意味

タグ	意味
B	系列が固有表現の始まりであることを示す
I	系列で固有表現が継続している
O	系列が固有表現でないことを示す
E	系列で固有表現が終わることを示す
S	系列1つで1つの固有表現であることを示す

5. 評価実験

提案手法の性能を評価するために, 評価実験を行う.

5.1 実験環境

本実験において, 形態素解析器には MeCab[*1] を用い, 辞書は IPADIC を利用する. CRF の実装には CRFsuite[*2] を用い,

*1 http://taku910.github.io/mecab/

*2 http://www.chokkan.org/software/crfsuite/

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する．収集した 373 エントリ，6,177 文に対して人手で固有表現タグを付与し，文として成立していないもの（土産の品名や店名のみの文など）を除いた 5,170 文を実験データに用いる．なお，実験データ内の固有表現数は，品名が 1,603，店名が 990 となった．

5.2 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．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，固有表現の既知/未知（学習データに含まれる/含まれない）を区別して評価する．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，性能が高くなる傾向がある[福島 08]．本手法では，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未知の固有表現に対する性能が重要になる．

5.3 ベースライン手法の性能

word 手法でのベースラインの実験結果を表 3 に，char 手法でのベースラインの実験結果を表 4 に示す．表 3, 4 の F 値（F1）より，既知/未知を区別しない場合，word 手法の性能が高いことがわかる．同様に，未知の固有表現に対しても単語ベースの性能が高い．既知の固有表現に対しては，2 つに大きな差はない．以上のことから，char 手法より word 手法の性能が全体的に高いといえる．

char 手法の性能が全体的に低くなった原因として，素性として使える情報が word 手法と比べて少ないことが挙げられる．char 手法で使える情報は表記と文字種のみであるため，文脈を word 手法ほど活用できない．ただし，文字の暗記的な学習は強く働く傾向があるため，既知の品名に対する再現率は char 手法のほうが高いという結果も確認できている．

また，CRF でタグ付けされた結果を確認したところ，char 手法は，文字数が 2, 3 文字の店名に対する性能が高いことがわかった．文字数が少ない店名は漢字で書かれたものが多く，「屋」や「庵」，「堂」といった文字が付くことが多い．これらの文字を手がかりにタグ付けをしているということが予想できる．

表 3. word 手法のベースラインの性能

		precision	recall	F1
区別なし	PRO	0.774	0.621	0.688
	SHO	0.855	0.683	0.759
未知のみ	PRO	0.635	0.526	0.573
	SHO	0.619	0.457	0.524
既知のみ	PRO	0.930	0.724	0.813
	SHO	0.985	0.825	0.897

表 4. char 手法のベースラインの性能

		precision	recall	F1
区別なし	PRO	0.719	0.611	0.659
	SHO	0.830	0.665	0.737
未知のみ	PRO	0.533	0.479	0.510
	SHO	0.547	0.430	0.476
既知のみ	PRO	0.918	0.769	0.836
	SHO	0.975	0.822	0.891

5.4 単一の素性・前処理・変更を追加した実験

word 手法に対して各素性・前処理・変更を 1 つだけ追加した手法（word+one 手法）と，char 手法に対して各素性・前処理・変更を 1 つだけ追加した手法（char＋one 手法）の性能を比較し，有効な素性を確認する．本手法では，未知の固有表現に対する性能が重要になる．そのため，未知の固有表現に対する性能が向上したものを有効な素性・前処理・変更と判断する．

未知の固有表現に対して，word+one 手法，char+one 手法，word 手法の結果を比較したグラフを図 3, 4 に示す．両図において，青い棒グラフが word+one 手法の性能，オレンジ色の棒グラフが char+one 手法の性能，緑色のラインが word 手法の性能である．

図 3 より，word 手法に，(3)inBracket または(7)BIOES を追加したとき，未知の品名に対して性能が向上したことがわかる．また，図 4 より，word 手法に，(5)correct を追加したとき，未知の店名に対して性能が向上したことがわかる．そのため，未知の固有表現に対しては，これら 3 つが有効であるといえる．

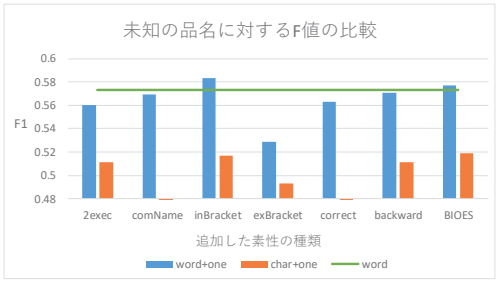

図 3. word+one 手法，char+one 手法，word 手法の未知の品名に対する F 値（F1）の比較

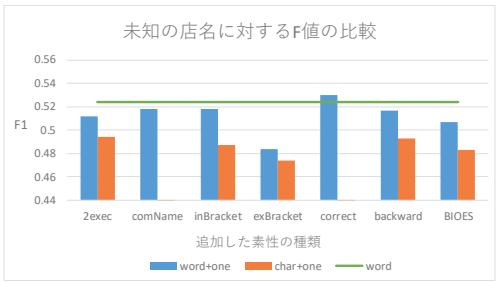

図 4. word+one 手法，char+one 手法，word 手法の未知の店名に対する F 値（F1）の比較

5.5 素性・前処理・変更の組み合わせによる実験

先の実験で未知の固有表現に対する有効性を確認した，(3)inBracket と(5)correct，(7)BIOES を組み合わせた場合の性能を確認する．なお，char 手法にこれらを追加しても，word 手法のベースラインの性能を超えられないため，word 手法にのみ追加し，性能を確認する．

本実験での組み合わせパターンは，以下の 2 つとする．

(a) inBracket + BIOES
(b) inBracket + BIOES + correct

最優先で抽出すべきは未知の品名であるため，未知の品名の抽出性能が向上する(3)inBracket と(7)BIOES の 2 つを両方のパターンに組み込む．また，図 3 から，(5)correct を追加することで，未知の品名の抽出性能が低下することがわかっているため，(a)と(b)により，(5)correct を追加する/しない場合の性能差も確認する．

word 手法に，(a)を追加したときの性能を表 5 に，(b)を追加したときの性能を表 6 に示す．表 5, 6 において，赤いセルは word 手法と比べて性能が向上した部分，青いセルは低下した部分である．表 5, 6 より，word 手法に(a), (b)を追加することで，未知の品名に対するすべての評価指標の値が word 手法より向上することが確認できた．(a)と(b)には大きな性能差はないが，

未知の品名の抽出性能を優先すべきであるため，(b)が優れているといえる．また，(b)の既知/未知の区別をしない場合の品名の F 値（F1）は，本稿での実験において最も高い値となった．

word 手法のベースラインと比べて性能が向上した理由について考察する．(3)inBracket によって，固有表現が“「」”などの記号とともに出現しやすいことが学習でき，(7)BIOES によって，固有表現の終わりを学習できる．“「」”とともに固有表現が出現するとき，必ず固有表現の終わりは“」”となるため，BIOES タグ形式によって“」”が固有表現の終わりを学習できるようになり，性能が向上したのではないかと考える．

表 5. word 手法+(a)の性能

		precision	recall	F1
区別なし	PRO	0.795	0.642	0.709
	SHO	0.858	0.680	0.758
未知のみ	PRO	0.670	0.544	0.597
	SHO	0.631	0.460	0.528
既知のみ	PRO	0.930	0.746	0.827
	SHO	0.986	0.817	0.893

表 6. word 手法+(b)の性能

		precision	recall	F1
区別なし	PRO	0.802	0.643	0.713
	SHO	0.857	0.679	0.757
未知のみ	PRO	0.682	0.548	0.605
	SHO	0.624	0.452	0.519
既知のみ	PRO	0.931	0.745	0.827
	SHO	0.986	0.820	0.895

6. エラー分析

6.1 品名・店名と判定できない系列

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った．以下に，タグ付けできなかった品名・店名系列の傾向を示す．

- 手がかり語（買う，貰う，食べる等）が参照系列と前後 2 形態素の中に存在しない
- 固有表現の文字種が 3 種類以上ある
- 固有表現の文字種がアルファベットのみである
- 固有表現の中や付近に空白が含まれている

1, 2 番目の系列は，参照する系列範囲を広げることでタグ付けできる可能性がある．そこで，参照範囲を広げて実験した結果，逆に性能が低下することを確認した．これまで正しくラベリングできた系列に対応できなくなったことが原因と考える．3 番目の系列は，学習データ内にアルファベットの品名や店名の固有表現が少ないことが理由として考えられる．また，形態素解析した際，単一形態素になること，文字数が多いことなどもアルファベットの固有表現を抽出する難しさとして挙げられる．4 番目の系列は，ブログ記事内での空白の使われ方の多様性が問題として挙げられる．ブログ記事内では，“_うなぎパイ_”のように固有表現の周辺に空白が出現するものや，“マウントバーム_しっかり芽_生タイプ”のように，空白自体も固有表現の一部となるもの等がある．パターンマッチなどによって余分な空白を削除できれば，空白の使われ方を絞り，その構造について学習できる可能性がある．

6.2 誤って品名・店名と推定される構造

品名・店名ではない系列に誤って品名・店名のタグが振られる場合や，品名・店名のタグが途中で途切れる場合のエラー分析を行う．

以下に，エラー分析により確認できた傾向を示す．

- “「○○”のように，固有表現に余分な系列（括弧）を含む
- 固有表現に余分な系列（括弧以外）を含む
- 品名・店名以外が「」などの記号に囲まれている
- 人を指す言葉や人の名前

1 番目の傾向は「白い恋人 12 枚入り」のように，品名の途中に金額や数量などの数字が入っている場合，「直虎_田舎_みそまん」のように，空白の後ろに品名がある場合に見られる．2 番目の傾向は，“花畑牧場シリーズ”の“シリーズ”や“香雲堂本店前”の“前”という文字列ごと，品名・店名と推定されてしまい誤ってしまうものである．3 番目の傾向は，菓子・デザートカテゴリのブログ記事では，品名や店名が“「」”や“（）”などの中に書かれることが多いことが影響している．4 番目の傾向は，“店名+さん”という構造を学習することで誤る例である．人の苗字や名前は漢字で書かれることが多いため，参照系列の範囲からは，店名か人の名前かを判断することは難しい．

これらの傾向のうち，1, 2 番目についてはパターンマッチを用いて文字列の一部を削除することで，正解として扱うことができる．しかし，これらの構造の出現数は少なく，パターンマッチを適応してもほとんど性能は向上しなかった．

7. まとめ

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を提案し，実験により提案手法の有効性を確認した．最終的に，品名に対する F 値が 0.713，店名に対する F 値が 0.757 という結果を得た．エラー分析により，品名・店名を抽出できない場合の系列や固有表現の構造を明らかにした．

今後の課題として，品名・店名手法の性能向上，品名・店名以外の土産情報の抽出手法の提案，レア度の計算方法について検討する．そして，土産情報をまとめたデータベースを構築し，システムの実装を目指す．

参考文献

[アサヒ 17] 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 (http://www.asahigroupholdings.com/company/research/hapiken/maian/201707/00641/)（アクセス日：2018 年 3 月 1 日）

[Nagao 16] Noriyuki Nagao and Kazuaki Ando: Extraction of Product Names for Constructing a Database of Souvenir Information, Proc. of the fifth International Conference on Informatics and Applications, pp.88-16,（2016）.

[川野 15] 川野 覚，溝渕 昭二: Q&A サイトを対象にした地域別土産物情報収集ツール，FIT2015 講演論文集，No.2，pp.221-222,（2015）.

[石野 10] 石野 亜耶，難波 英嗣，竹澤 寿幸: 旅行ブログエントリからの観光情報の自動抽出，知能と情報（日本知能情報ファジィ学会誌），Vol22，No.6，pp.667-679,（2010）.

[矢野 17] 矢野 憲，若宮 翔子，荒牧 英治: 医療テキスト解析のための事実性判定と融合した病名表現認識器，言語処理学会第 23 回年次大会 発表論文集，pp.242-245,（2017）.

[Ratinov 09] Ratinov, L. and Roth, D.: Design challenges and misconceptions in named entity recognition, Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL'09, pp.147-155,（2009）.

[福島 08] 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/第 6 回日本データベース学会年次大会（DEWS2008），（2008）.