5L-03

CRFを用いたブログ記事からの品名・店名抽出

池田　流弥†　　長尾　哲志‡　　安藤　一秋†

†香川大学工学部　‡香川大学大学院工学研究科

1.はじめに

近年，オンラインショップの普及により，現地に行かなくても多種多様な商品が購入できるようになった．そのため，旅行時に土産を選定する際には，現地でしか購入できない商品が好まれるようになった[1]．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リストは存在しないため，土産情報を抽出するための手がかりとして利用できない．そこで，テキスト中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，CRFs（Conditional Random Fields）を利用して，ブログ記事から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.ブログ記事からの品名・店名抽出手法

現地でしか購入できない土産情報を一元的に管理・提供している情報源は存在しない．また，現地でしか購入できない土産は，オンラインショップでも取り扱われていないため，このようなサイトから品名を抽出することもできない．そこで，本研究では，ブログや QA サイト等のテキスト群から土産の品名を自動抽出する手法を提案する．なお，本研究では，食品の土産のみを抽出対象とする．

ブログ記事から土産の品名を抽出する手法としては，形態素 N-gram と残差 IDF を組み合わせた手法[2]や正規表現と SVM（Support Vector Machine）による品名の妥当性判定を組み合わせた手法[3]などがある．[2]の手法は品名を網羅的に抽出するため再現率は高いが，適合率に課題が残る．[3]の手法は再現率が低く，情報の取りこぼしが多い．

土産の品名は単なる複合名詞ではなく，店名や略称と組み合わせられるなど，多様な形態で表現される．土産の品名とその構造の例を表 1 に示す．“じゃがポックル”のように略称同士の繋がり（“じゃがいも”＋“コロポックル”）で構成されているもの，“名物かまど”のように店名自体が品名となっているもの，“白い恋人”や“おたべ”のように品名の一部や全体が名詞以外で構成されるものなどがある．したがって，土産の多様な構造をすべて網羅する規則を定義することは困難であると考える．

そこで本研究では，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，固有表現抽出により，土産情報を抽出する手法を提案する．固有表現抽出には CRFs による系列ラベリングを採用する．

表 1. 土産の品名構造の例

名前	構造	品詞
白い恋人	「白い」＋「恋人」	形容詞+名詞
じゃがポックル	略称+略称	名詞
名物かまど	「名物」+店名	名詞
おたべ	「おたべやす」の略称	動詞
花畑牧場 生キャラメル	店名+品名	名詞+名詞

CRFs で使用する学習データは以下の手順で作成する．

(1) ブログや QA サイトから土産について書かれた文章を収集し，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を付与する．なお，タグの形式は IOB2 を利用する．
- 食品名に品名のタグを付与
- 食品を販売している店に店名のタグを付与
- 「」などの記号ごと品名，店名のタグを付与
- 品名，店名でないものに O タグを付与

食品名に土産タグを振る理由としては，土産と商品に土産⊆商品という関係が成立するからである．土産でない商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抽出する．

ラベリングは，単語ベースと文字ベースの 2 種類で行い，性能を比較する．単語ベースでは形態素単位に，文字ベースでは，1 文字単位にラベリングする．

単語ベースでは，参照している形態素の前後 2 形態素について，以下の 3 つの素性を利用する．

(1) 表記
(2) 品詞細分類
(3) 文字種

文字ベースでは[4]を参考に，参照している文字の前後 2 文字について以下の 2 つの 1-gram，2-gram を作成し，素性として使用する．

(1) 表記
(2) 文字種

また，上記の素性の組み合わせをベースラインとし，ベースラインに以下の工夫や素性を追加することで，これらの有用性を確認する．

(1) 文章に店名があるならフラグを立てる（2exec）
(2) チョコ，クッキーなどの食品名になり得る単語にフラグを立てる（common name）
(3) 文末から固有表現抽出をする（back）
(4) タグ付けに IOB2 タグ形式ではなく，BIOES タグ形式を使用する（BIOES）
(5) 「」などの中の単語にフラグを立てる（in key）
(6) 「」などのタグを O タグにする（key break）
(7) 形態素解析での品詞間違いの修正（correct）

(1)と(2)は，品名，店名が文中に出現する場合と出現しない場合の文構造を区別することを期待した素性である．(3)は，後ろから文を見ることにより，動詞の情報を有効活用できることを期待した素性である．(4)は BIOES が先行研究で高い性能が報告[5]されているため候補とした．

Extraction of Product and Shop Names from Blog Entries Using Conditional Random Fields
† Faculty of Engineering, Kagawa University
‡ Graduate school of Engineering, Kagawa University

(5)と(6)はブログ記事で品名や店名が「」や（）など括弧中に書かれやすいことに注目したものである．(7)は，品詞の素性をより正確に使用できると考えて導入した．

なお，文字ベースでは，(2)と(7)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．

3.実験

ベースラインと比較することで，提案手法の性能を評価する．形態素解析器には MeCab を用い，辞書は IPADIC を利用する．CRFs 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．

評価データは，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 5,170 文に対して，人手で固有表現タグを付与した．タグを付与した固有表現数は，品名が 1,603，店名が 990 となった．

適合率，再現率，F 値を評価尺度とし，10 分割交差検証で評価する．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，完全一致した場合を正解と判断する．

また，固有表現の既知/未知（学習データに含まれる/含まれない）を区別した評価も行う．学習データに含まれる既知の固有表現の場合，固有表現の表層的な文字列自体を学習してしまうため，性能が高くなる傾向がある[6]．本手法では，オンラインショップで販売していない土産の品名と販売店舗名を抽出することを目的としているため，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特に重要になる．そのため，既知/未知の観点でも性能を評価する．

単語ベース，文字ベースでのベースラインの実験結果を表 2，表 3 に示す．

表 2. 単語ベースでのベースラインの性能

		適合率	再現率	F値
区別なし	PRO	0.774	0.621	0.688
	SHO	0.855	0.683	0.759
未知のみ	PRO	0.62	0.514	0.56
	SHO	0.611	0.452	0.517
既知のみ	PRO	0.947	0.737	0.828
	SHO	0.988	0.829	0.901

表 3. 文字ベースでのベースラインの性能

		適合率	再現率	F値
区別なし	PRO	0.719	0.611	0.659
	SHO	0.83	0.665	0.737
未知のみ	PRO	0.536	0.471	0.497
	SHO	0.549	0.401	0.462
既知のみ	PRO	0.923	0.765	0.836
	SHO	0.979	0.832	0.898

表 2 と表 3 を比較することで，全体的に単語ベースの性能が高いことがわかった．また，文字ベースでは，単語ベースと比べて，既知固有表現の再現率が高いことが確認できた．

次に，ベースラインに対し，検討した素性・工夫を追加し，その性能を確認した．単語ベースのベースラインと比較して，既知/未知を区別しない場合に性能が向上したものは，単語ベースに(6)key break を追加した場合のみであった．品名に対する性能が向上し，店名に対する性能が低下したものは，単語ベースに(3)back，(4)BIOES，(5)in key を追加した場合の 3 通り，品名に対する性能が低下し，店名に対する性能が向上したものは文字ベースに(6)key break を追加した場合の 1 通りであった．このうち，品名の未知固有表現に対して性能が向上したものは，単語ベースに(4)BIOES，(5)in key を追加した場合の 2 通りであった．最優先で抽出したい品名は未知の品名であるため，これら 2 つの工夫は本手法で有効であると考える．これら 2 通りの実験結果を表 4 と表 5 に示す．赤で示した部分はベースラインと比較して向上した部分，青で示した部分は低下した部分である．表 4 と表 5 より，固有表現ラベルごとの性能の向上と低下には，トレードオフの関係があることが確認できた．

表 4.単語ベース+BIOES の性能

		適合率	再現率	F値
区別なし	PRO	0.793	0.628	0.7
	SHO	0.857	0.673	0.753
未知のみ	PRO	0.642	0.508	0.565
	SHO	0.611	0.435	0.503
既知のみ	PRO	0.952	0.756	0.842
	SHO	0.988	0.822	0.897

表 5.単語ベース+in key の性能

		適合率	再現率	F値
区別なし	PRO	0.775	0.63	0.694
	SHO	0.847	0.679	0.753
未知のみ	PRO	0.631	0.528	0.572
	SHO	0.594	0.452	0.512
既知のみ	PRO	0.94	0.741	0.828
	SHO	0.991	0.82	0.896

4.まとめ

本稿では，ブログ記事から品名・店名を自動抽出する手法を提案し，実験により提案手法の性能と有効な素性を確認した．今後は，素性の組み合わせの有効性を確認する．また，本手法を用いて土産の品名と店名を抽出し，それらを用いて，現地でしか購入できない土産情報を収集する手法を検討する．

参考文献

[1] お土産についてのアンケート・ランキング http://chosa.nifty.com/travel/chosa_report_A20140221/?theme=A20140221&theme=A20140221 (参照 2018-01-08).

[2] 川野他，“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, pp.221-222, (2015).

[3] 長尾他，“土産情報 DB 構築に向けた品名候補の抽出”, IPSJ2017 講演論文集, pp.461-462, (2017).

[4] 矢野他，“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”，NLP2017 発表論文集, pp.242-245, (2017)

[5] Lev, et al, “Design Challenges and Misconceptions in Named Entity Recognition”, Proceedings of the Thirteenth Conference on Conmputational Natural Language Learning, CoNLL ’09, pp.147-155, (2009).

[6] 福島他，“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”, DEWS2008, (2008).