意外性のある手土産抽出手法の提案

11471061　瀧川　亮（灘本研究室）

あらまし：旅行をする際，多くの旅行者はお土産専門店でお土産を買う．この時他の旅行者とお土産が重複するなどの問題がある．そこで本研究では旅行先ならではの土産として「店名+商品名」という表記のものを手土産とし，中でも旅行者がお土産として買わないものを意外性のある手土産とする．そして，深層学習を用いて意外性のある手土産を抽出する手法を提案する．

1. はじめに

旅行をする際，多くの旅行者は空港や駅，観光スポット周辺のお土産専門店でお土産を購入することが多い．お土産を購入する場所が限定されることにより，以前と同じお土産を購入してしまう事や，他の旅行者とお土産が重複してしまう場合がある．また，地元の人たちが推薦するお土産を購入しようと思っても，どのように探したら良いか分からなかったり，現地の土産情報に詳しくなかったりする事により，このような土産品を探し出すことは困難である．

一方で，友人や親戚などを訪問する際，訪問先へ持っていく土産のことを一般的に手土産という．手土産は地元の品を持っていく事が一般的で，その土地をよく知っていることから，旅行者が選ぶお土産とは違ったものを持っていくと考えられる．また手土産の中にはほかの地域では知られていない物も多く存在しており，意外性という観点からお土産としても適していると考えられる．しかしインターネット上ではお土産と手土産の情報が混在しており，その中からそのような手土産を見つけることは困難である．一般に手土産は「富士屋のとうまん」のように「店名+商品名」からなる場合が多い．本研究では，この手土産の特徴に着目し，「店名+商品名」の表記のものを手土産と定義する．さらに，手土産は一般の旅行者が買わないものとする．つまりは，本研究では手土産は，「店名+商品名」の表記のものから，旅行者が購入しているお土産を引いた残りを手土産と定義し，この手土産を自動で抽出する手法を提案する．

以下に手土産の抽出手順を示す．

① ユーザの入力した地名と”手土産”をクエリとして，その地名に関する手土産情報を含む Web ページ群を取得する．
② 取得した Web ページ群の本文を形態素解析し，「○○の○○」という表現を抽出する．これを手土産候補キーワードと呼ぶ．
③ 手土産候補キーワードをクエリとし，検索を行う．
⑤ 検索結果のスニペットを用いて深層学習を行い，その結果から手土産のものかを判定し，手土産候補を決定する．
⑥ 手土産候補の中から，予め我々が用意したお土産名リストと比較して意外性のある手土産を決定する．

2. 関連研究

白数[1]らは，ジオタグツイートを用いて群衆の認知特性を抽出し，任意の場所において国民性に合わせてその地域の郷土品を提示可能な推薦システムを提案している．長尾[2]らは，オンラインショップ，ブログ，QA サイトを用いて，オンラインショップでは購入できない土産情報をオンラインショップで購入できる土産と共にユーザに提示するシステムを提案している．それに対して，本研究では手土産に着目した土産品を抽出する点と，Web 上から手土産に関するスニペットを抽出し，深層学習を用いて手土産の候補を決定する点が異なる．

3. 手土産抽出手法

3. 1. Web ページ群の抽出

ユーザの入力した地名と「手土産」をクエリとして，地名に関する手土産の情報を含む Web ページ群を取得する．

3. 2.「○○の○○」表現の抽出

手土産の情報は「ルタオのチーズケーキ」や「横浜文明堂の極上金カステラ」のように，商品名に加えて店名も合わせて表記されることが多い．本研究ではこのような「○○の○○」という表現を手土産候補キーワードとして抽出する．具体的には，3.1 にて取得した Web ページ群を形態素解析器 Juman を用いて形態素解析を行い，「○○の○○」を取得する．ここで，○○は名詞を対象とする．そして，手土産候補キーワードを用いて Web 検索を行い，検索結果よりスニペットを取得する．

3. 3. 手土産候補の決定

手土産候補キーワードの中には「熊の生息地」や「こどもの日」のように手土産に関係のないキ

ーワードも含まれているため，それらを取り除く必要がある．本研究では深層学習を用い，手土産候補キーワードを手土産とそうでない物の２つに分類をする．深層学習には CNN を用いる．

CNN のネットワークには Yoon Kim[3] の Convolutional Neural Networks for Sentence Classification を使用し，入力には Wikipedia の記事を基に学習した Word2Vec による分散表現を用いる．パラメータはエポック数を 30 とし，それ以外はデフォルトの設定を使用する．学習データには人手で作成した手土産に関係のある単語をクエリとして得られたスニペット 12866 件と手土産に関係のない単語をクエリとして得られたスニペット 10064 件を用いる．

手土産候補キーワードより得られたスニペットを CNN で学習し，手土産であると分類されたものを本研究では手土産候補とする．

3．4．お土産の除去

手土産候補の中には既によく知られたお土産も含まれる．本研究ではそのようなお土産は意外性の観点から取り除く必要がある．そこで，お土産はじゃらん net に掲載されているお土産 1073 件とする．手土産候補の中にお土産がある場合取り除き，残った手土産候補を意外性のある手土産とする．

4．評価実験

4．1．実験条件

提案手法を用いて，意外性のある手土産が正しく抽出出来ているかどうかを評価するため，実験を行った．実験は予め用意した地名 100 件と「手土産」をクエリとして取得した Web ページ群より抽出した手土産候補キーワード 1000 語を用いた．そして，ベースラインと比較を行い，提案手法の有用性を示した．ベースラインは，3.2 節で提案した手法を用いる．このとき，「○○の○○」の後ろの名詞が”カテゴリ:人工物-食べ物”と解析された手土産候補キーワードをベースラインによる手土産候補とする．提案手法ではまず，手土産候補キーワードの中から手土産に関係のある単語を 50 語，手土産に関係のない単語 50 語，合計 100 語を無作為に選び，各単語をクエリとした検索結果のスニペットを取得した．スニペットの件数は不要なスニペットが混ざるのを防ぐため，上位 10 件を対象とした．そのスニペットを深層学習で手土産に関係のあるスニペットかどうかを判定し，6 件以上手土産に関係のあると判定されたスニペットを持つ手土産候補キーワードを提案手法による手土産候補とした．そして，ベースラインによる手土産候補と提案手法による手土産候補を手動で用意したお土産名リストと比較し，お土産名リストに載っていないものを正解データとして，それぞれの再現率, 適合率, F 値を求めた．

4．2．実験結果と考察

実験結果を表 1 に示す．ベースラインでは適合率，再現率共に低い値となった．これは「寶月堂の瀬戸の波」や「くらづくりの喜多のかけ橋」のように手土産には名前に食べ物が入っていないものが多く存在していた．そのため，解析結果が食べものであるか否かを重視するベースラインでは，うまく手土産を取得できなかったと考えられる．これに対し，提案手法はベースラインよりも高い値を示している．これは CNN を用いてスニペットを学習することで，手土産の名称に左右されずに取得できたためであると考えられる．しかしながら適合率 0.56 であり，良いとは言いがたい．これはスニペットを学習に用いているため，手土産ではないが似たような表記のスニペットを持つ単語も手土産であると分類されてしまっているためと考えられる．以上より提案手法のほうがベースラインよりも手土産抽出には適していると考えられる．

手法	適合率	再現率	F 値
ベースライン	0.153	0.162	0.157
提案手法	0.560	0.940	0.694

表 1：実験結果

5．まとめと今後の課題

本研究では，地名と「手土産」をクエリとして Web ページ群を取得し，その中から手土産候補キーワードを抽出した．その後深層学習を用いることにより手土産候補を決定し，お土産を除去することにより意外性のある手土産を推薦する手法の提案を行った．今後の課題としては深層学習の分類精度の向上と抽出した手土産の人気度を用いて，よりお意外性のある手土産の抽出を行っていきたい．

参考文献

[1]白数紘之，王元元，河合由起子，アダムヤトフト，“ジオタグツイート分析に基づく群衆の認知特性抽出および郷土品推薦システム” DEIM Forum 2017，-P2-2.

[2]長尾哲志，安藤一秋，“オンラインショップで購入できない土産を提示するシステムの構築”FIT2015(第 14 回情報科学技術フォーラム).

[3] Yoon Kim，“Convolutional Neural Networks for Sentence Classification” EMNLP 2014，pp.1746-1751.

[4] JUMAN: http://nlp.ist.i.kyoto-u.ac.jp/