K-053

つぶやきマイニングによるお出かけ情報生成法とその評価

A Tweet Mining Method for Generating Going Out Information and Its Evaluation

深谷 昭宏† 浦 正広‡ 遠藤 守† 山田 雅之† 宮崎 慎也† 安田 孝美‡
Akihiro Fukaya Masahiro Ura Mamoru Endo Masashi Yamada Shinya Miyazaki Takami Yasuda

1. はじめに

出かけ先を選定する際の情報収集の方法として，観光サイトや口コミサイトなどが挙げられるが，それらには情報が不足していたり，内容に偏りが生じていたりするといった問題点がそれぞれ挙げられる．一方，その時に思ったことを気軽に発言できる，Twitter に代表される「つぶやきメディア」が登場し，人々の日常生活の様々な内容がウェブ上に集約されてきている．

これらを背景として，筆者らはこれまでに，位置情報が付加されたつぶやきのマイニングによりスポット情報や地図を推測・生成し，それら不確定なお出かけ情報を掲載する「あいまいマップ」を提案している[1]．これは，つぶやきを一様にマイニングすることで，ガイドブックに掲載されないようなお出かけスポット情報も抽出できるため，出かけ先での発見や体験をユーザ間で広く共有できるようになることが期待できる．しかしながら，つぶやきメディア上でなされる発言には口語調のものが多いため，正しい文章を対象とした形態素・構文解析では抽出できる情報に限りがあるなど，情報抽出の精度をいかにして高めるかが，より多くのお出かけスポット情報を掲載する上での課題となる．

そこで本研究では，より多くの情報が抽出できるように既研究の抽出手法を再考し，提案手法を用いてお出かけ情報の生成が行えるか，実際のつぶやきを対象に実験を行うことで，その評価を行う．

2. あいまいマップ

例えば出かけ先の選定プロセスにおいて，発案の段階では候補地の情報や位置は曖昧な記憶に基づくものであることが多い．しかしながら，各々が情報を持ち寄ることで，それらが組み合わさって集合知的な効果が発生し，1 つの地図が全体で共有される．あいまいマップは，この概念をモデルとする．図 1 で示すように，つぶやきに付加された位置情報を用いて推測された地形と道，つぶやきの時間・位置情報・内容から推測されたスポット情報により構成される．新しいつぶやきを取り込んでいくことで，掲載するお出かけ情報の随時更新が可能となる．

お出かけスポット情報は，各つぶやきに Mecab[2]を用いた形態素解析を行い，その結果のうち「名詞＋助詞＋（副詞＋）形容詞」「名詞＋助詞＋（副詞＋）動詞」のように主語と述語の関係になっているのの，「（副詞＋）形容詞＋名詞」のように主語のみのもの，「（副詞＋）形容詞」のように述語のみの品詞パターンのもののみを抽出し，つぶやきに付加された位置情報が一定の距離以内のものを 1 つにまとめることで生成する．

†中京大学, Chukyo University

‡名古屋大学, Nagoya University

図 1　生成された地図とスポット情報

3. つぶやきマイニングによるお出かけ情報生成

Twitter のつぶやきには，口語表現やスラングが用いられているものも多い．ウェブ上にあるそのような文章を対象とした形態素解析法が提案されているが[2]，本研究では，スポット名やその特徴といった，お出かけスポット情報の抽出を主目的とする．そこで，前章で述べた既研究における抽出手法を基に，Mecab を用いた形態素解析において，主語と述語の関係が明確になるような処理を考える．まず，Twitter のスラング的な表現やハッシュタグなどの固有表現を除去する．つぎに，主語や述語として 1 つに括ることのできる品詞を纏めて，それらを，過去のつぶやきを参考に設定した品詞の出現パターンとマッチングさせることで抽出を行う．

3.1 ノイズの除去

Twitter には，「なう，うぃる，わず」のように，その行為をいつ行ったり，その場所にいつ行ったかを示すスラング的な表現が存在したり，「#○○」のような共通の話題を示すハッシュタグ，「@○○」のような特定ユーザに対しての返信，「RT ○○」や「QT ○○」のような他者のつぶやきのリツイートなど，独自の表現がつぶやき中に用いられるケースや，外部 URL のリンクが数多く見受けられる．

そこで，形態素解析を行う前に正規表現などを用いてこれらのノイズと考えられる情報を除去することで，抽出精度の向上を図る．まず，Twitter の機能に基づく固有表現を除去する．「#○○」や「@○○」のような識別子的なものについては，該当文字列を除去する．「RT ○○」や「QT ○○」のような他者のつぶやきを含むものについては，つぶやきに付加された位置情報と，他者のつぶやきとの間に関連性がないケースも多いことが考えられるため，以降の文字列も併せて除去する．つぎに，外部 URL のリンクについては，リンク先には位置情報に関連する写真や動画が掲載されている可能性が高いことから，文字列は除去し，リンク先のメディアはスポットに関する視覚的な情報として活用する．最後に，文末にある「なう，うぃる，わず」といった文法に依存しないスラングを除去することで，テキストを整形する．

3.2 品詞のパターン

文章が「主語＋述語，主語のみ，述語のみ」の形になるように処理を行う．まず，文章を一文ごとに処理するために，句点や「!?，?!」などの記号で分割する．つぎに，纏められる品詞を結合する．名詞が連続した場合には複合名詞として 1 つの名詞として処理し，並立助詞や連帯化の助詞で接続された名詞についても 1 つの名詞として処理する．動詞については，動詞と動詞を助詞で接続しているもを一つの動詞として処理する．動詞の直後の助動詞もその動詞に含める．形容詞についても，直後の助動詞はその形容詞に含める．接頭詞は，その直後に出現する名詞と形容詞に含める．

以上の処理により整理された品詞の組み合わせを対象に，表 1 に示すパターンを抽出のマッチングに用いる．このとき，マッチングは表に示す優先順位で実施する．また，口語表現においては，助詞が省略されているなどの構文が不完全なものがある．このため，各パターンにおいて係助詞がないものについても，マッチングの対象とする．なお，表において[]と{}はそれぞれ主語，述語的な役割を示している．

表 1　品詞の組み合わせ

優先順位	組み合わせ
1	[名詞] + 係助詞 + {(副詞 +) 形容詞}
2	[名詞] + 係助詞 + {(形容詞 +) 動詞}
3	[名詞] + 係助詞 + {名詞}
4	{(副詞 +) 形容詞}
5	[名詞]

4. 評価実験

実際に Twitter でつぶやかれたテキスト 200 件を対象に，前章で示した提案手法を適用することで，その有効性を確認する．

4.1 実験

実験の対象とするつぶやきは，2010 年 4 月 18 日に Twitter の streaming API によって取得したもののうち，日本においてつぶやかれた日本語のつぶやき計 200 件とする．これらつぶやきを，2 章で示した既研究における品詞パターンと，3 章で示した提案手法によりそれぞれマイニングすることで，抽出状況を比較する．

抽出結果を表 2 に示す．表は，既研究と提案手法の双方において，抜け落ちが無く文意に沿った正しい抽出が行えたもの，文意には沿っているが抽出が文章の一部分のみのもの，正しい抽出が行えなかったもの，パターンに該当せずに抽出自体が行えなかったものの件数をそれぞれ示している．

表 2　抽出結果

抽出結果	既研究	提案手法
正しい抽出	30	71
一部のみ抽出	24	67
正しくない抽出	25	50
できなかった	121	12
合計	200	200

表より，既研究において抽出の対象となったつぶやきが 39.5%にとどまっていることに対し，提案手法では 94%のつぶやきが抽出の対象となった．また，対象としたつぶやきのうち正しく抽出できた割合は，既研究が 15%に対して提案手法は 35.5%であり，抽出できたつぶやきのうち正しく抽出できていたものの割合も，既研究手法が約 38%に対し提案手法も約 38%と，同等の精度を維持している．なお，提案手法において正しく抽出できた 71 件のうち，スポット名やそのスポットにある商品などの名前，また，そこで行われる行為など，実際のお出かけ情報に結び付くものについて言及していたものは 26 件であった．一方，一部のみ抽出できたものや正しく抽出できなかったもの，抽出自体ができなかったもの 129 件のうちで，お出かけ情報に結び付くつぶやきは 35 件であった．また，一定距離以内に複数のつぶやきが存在しなかったため，複数の特徴を持つスポットは，一つのつぶやき中に複数の文を持ち，それぞれが特徴として抽出できた 2 件に限定された．

4.2 考察

正しく抽出できなかったもののうちでお出かけ情報を含んでいたものは 35 件であったが，そのうちで「○○ (@○○) http://4sq.com/○○」，「○○ on @foursquare! http://4sq.com/○○」，「ｲﾏｺｺ L:○○」，「I'm at ○○（○○）」，「○○ L: ○○付近。 http://bit.ly/○○」などの API を用いてつぶやかれた，位置情報を含む定型文によるものは 30 件であり，割合にして 86%であった．また，不動産の紹介や地震の震源地を示すものもあった．このため，これらを抽出のテンプレートとして考慮することで，抽出率の向上が期待できる．

5. おわりに

本研究では，つぶやきマイニングによるお出かけ情報の生成手法を提案した．また，実際のつぶやきを対象とした提案手法の評価実験を行い，その有効性を確認した．

今回の実験において対象とした件数では，1 つのスポットについて言及する複数のつぶやきが出現しなかったため，今後は，つぶやきの収集を行い，複数のつぶやきからスポット情報を形成する手法の有効性についても確認したい．

謝辞

本研究の一部は文部科学省科学研究費補助金，財団法人人工知能研究振興財団研究助成，財団法人 JKA 補助金による．

参考文献

[1] 深谷昭宏，浦正広，遠藤守，山田雅之，宮崎慎也，安田孝美：あいまいマップ　～時空間情報が付加された「つぶやき」からの地図の生成～，第 26 回 NICOGRAPH 論文コンテスト論文集，IV-1（2010.09）．

[2] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto Applying Conditional Random Fields to Japanese Morphological Analysis, Proc.of EMNLP-2004, pp.230-237 (2004).

[3] 持橋大地，鈴木潤，藤野昭典：条件付確率場とベイズ階層言語モデルの統合による半教師あり形態素解析，言語処理学会第 17 回年次大会予稿集，B5-2（2011.03）．