大量のつぶやきから日本酒の美味しい店を発掘する
知識源としてのマイクロブログ活用の試み

那須川 哲哉　吉田 一星　西山 莉紗　吉川 克正
伊川 洋平　大野 正樹　金山 博　鈴木 祥子　村上 明子
日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

ビッグデータ活用への関心が高まる中、Twitterをはじめとするマイクロブログデータのテキストマイニング[1]に対する根強い期待が感じられる。多様な人々の日常的なつぶやきを大量に収集して分析することにより、例えば、災害情報を特定[2]し、災害の状況を把握[3]することが実際に可能になってきており、マイクロブログの分析は自然言語処理およびテキストマイニングの有望なアプリケーションとして発展していくことが期待できる。

但し、人が全ての考えをマイクロブログで発信するわけではなく、自社商品に対する人々の反応を分析したいという企業の期待に対し、応えられないことも多い。マイクロブログの分析を自然言語処理およびテキストマイニングの有望なアプリケーションとして発展させるためには、実データの調査を通じて有用なタスクを見極め、その有用性を高める手法を開発していく必要がある。

マイクロブログで発信された情報を活用する新しいタスクを検討したいという観点から、筆者らはマイクロブログへの書き込み（ツイート）から日本酒の美味しい店を探すことを試みた。このタスクを選んだ背景として

- データ量の優位性：飲食に関する情報はマイクロブログ上で比較的発信されやすく、データが揃いやすい可能性
- 記述内容の特色：レビューサイトへの書き込みよりも気楽に発信されるマイクロブログでは、例えば本音が出やすいなど、レビューサイトとは異なる情報が得られる可能性
- ノイズ排除の可能性：レビューサイトで危惧されるサクラ的な書き込みが同じユーザIDの他の書き込みを見ることで排除できる可能性

への期待がある。対象データがこういった要件を満たすタスクであれば、有用な結果につながる可能性が高いのではないかと考えた。さらに、Twitterのプロフィール欄に「日本酒」という表現を含むユーザ数はツイプロ[1]によると19,232に及ぶ[2]ことから、日本酒に関心のあるTwitterユーザが相当な数存在すると考えられ、ツイートに日本酒の美味しい店に関する情報が存在する可能性は高そうである。

本稿では、4百万件を越すツイートから日本酒の美味しい店を発掘し、実地調査によりその有効性を評価した一連の取組みと、そこで得られた知見を示す。

2. 分析対象データの収集とデータの特徴

まず、Twitter社が提供しているSearch API[3]を利用して、2012年12月21日から2013年6月12日までの半年程度の期間において「日本酒」もしくは「ビール」を含むツイートを収集した。「日本酒」だけでなく「ビール」も含めたのは、飲酒に関するツイートを少し広めに取ることで、「日本酒」という表現が出ていなくても日本酒の銘柄の表現が出ているデータを検出したり、ビールに関するツイートと日本酒に関するツイートの比較を行ったりできるようにと考えたためである。

タスクの目的に合わないデータを減らすという観点から、下記に該当するデータは対象外として削除するようにした。

- 重複：RTで始まるデータ
- 自動発信：ユーザ名に「_bot」「_matome」「_news」などの文字列を含む botの可能性の高いデータ
- 宣伝：業者用語の可能性が高い表現（「ご来店」など）を含むデータ

期間内で収集できなかった日もあるため、網羅的ではなく、自動発信や宣伝の可能性の高いデータをすべて排除できていないという点も含め、多様なノイズが含まれているが、最終的に4,123,950件のデータが分析対象となった。この約412万件のうち、「日本酒」を含むものは約77万件であった。

収集したデータをIBM ® Watson Content Analytics Version 3.5（以降、WCAと略記）[4]に投入し、日本酒の美味しい店を探すという観点から、まずは下記内容に関する調査を行った。

- 日本酒の銘柄の言及
- 店名の言及
- 好不評を示す評価表現

[1] http://twpro.jp/
[2] 2015年1月16日現在
[3] https://dev.twitter.com/rest/public/search

2.1. 日本酒の銘柄の言及の調査

日本酒の銘柄を Wikipedia の『日本酒の銘柄一覧』から抽出し、得られた 1,618 語を WCA に辞書登録することで、日本酒の銘柄に言及している書き込みを特定することを試みた。その結果、大量のデータが日本酒の銘柄の表現を含むものとして抽出された。「日本酒」という表現を含まずに日本酒の銘柄を含むツイートが見出された反面、抽出された銘柄表現には、「おめでとう」「おやじ」「まごころ」「ん」といった、日本酒の銘柄以外の意味でも使われるものが多数存在し、こういった表現が実際に日本酒の銘柄として用いられていることは稀であった。すなわち、単純に辞書登録し、形態素解析レベルの処理でマッチさせるだけでは、日本酒の銘柄として記述されている表現を特定することが困難なケースが多いことが分かった。今回対象としているデータは、140 文字以内という短い記述の中に「ビール」か「日本酒」を含んでいるデータである。お酒に関するトピックを含む可能性が高いこのデータにおいてさえ日本酒の銘柄以外の意味で用いられることの多い表現（「おめでとう」「おやじ」「まごころ」「ん」など）が、特定の文脈において日本酒の銘柄の意味で使われていることを判断する処理は難易度が高く、高度な語義解消の仕組みが必要となると考えられる。また、日本酒の美味しい店を探す上では、銘柄の言及の特定が必須ではないと考えられることから、日本酒の銘柄の情報を用いるアプローチは取らないことにした。

2.2. 店名の言及の調査

本タスクにおいて店名の特定は必須であるため、まずは店名の言及がどの程度存在するかを調査し、さらに店名をどのように抽出するかを検討した。

「日本酒」をプロフィール欄に含むユーザの「日本酒」を含むツイート1,000件において、飲食店名らしき文字列を含むツイートを調査したところ、わずか6件であった。店名に言及しているデータはあまり多くないことから、店名を含むデータを効率良く集めるためには、何らかの工夫が必要になる。

店名は地名と併記される可能性が高いと予想し、「ビール」もしくは「日本酒」と地名を含むデータ200件の中で飲食店名らしき文字列を含むツイートを調査したところ72件（うち「日本酒」を含むデータでは31件）であった。したがって、ツイートのみから店名を抽出するには地名の活用が役立ちそうであるが、そのためには地名を認識する処理に加え、地名を含むツイートの中から店名を特定する処理が必要となる。

店名を特定するためには、例えば「○○という店」「○○で飲んだ」といった表現から○○の表現を抽出する方法が考えられるが、こういった表現の出現頻度は低く、今回対象とした約412万件のデータにおいて、「という店」を含むデータはわずか175件であった。「で飲んだ」は6,721件存在したが、「3秒で飲んだ」「ストローで飲んだ」「新潟で飲んだ」など店名抽出にはつながらないデータが大半であった。

以上の通り、ツイートデータそのものから店名をリストアップするのは容易でないため、Foursquare API[4]を利用して、ツイート以外のデータから店名の情報を取り出すことにした。このアプローチの大きなメリットとして、Foursquareのvenueデータに店名と共に含まれる位置情報を利用できるようになる。良い店を探しても遠方では訪問が困難なため、位置情報を活用し、例えば東京駅周辺といった、特定の地域に存在する店の名称をリストアップすることで、実用性の高い店探しが可能になる。

2.3. 評価表現の調査

美味しい店を特定する上では、美味しいことを示す評価表現の言及を認識することが有効ではないかと考え、本データの中で評価表現の自動抽出[5]を試みた。

ツイートのデータは、商品レビューのデータのように評価に用いることが主目的ではない上、140文字という制約があるため、評価表現の自動抽出の対象としては不適切なデータではないかと予想されたが、実際に自動抽出処理を適用したところ、「日本酒」を含むツイートにおいて好評な表現として、「取り揃える」「日本酒にあう」「日本酒がすすむ」「日本酒が揃う」「くせがない」「利き酒ができる」といった表現が抽出され、データ量が多ければ、ツイートのデータからも評価表現の自動抽出が可能であるという感触が得られた。また、「ビール」を含むツイートにおいては、好評な表現として「食欲を増進する」や、不評な表現として「体が冷える」「トイレが近い」など、「日本酒」のツイートとは異なる表現も抽出されることが確認できた。

但し、こうして得られた各評価表現の出現頻度は数百万件規模のデータにおいて、わずか数件から数十件にとどまる。そのため、評価表現に依存して美味しい店を特定するアプローチは取らないことにした。

3. 美味しい店の特定手法

前節の調査結果を踏まえ、まずは店名に言及しているデータを特定することに注力し、下記のステップで、言及されている店を特定した上で、美味しい可能性が高そうな店を選択することにした。

1.エリア内の店名の取得
2.店名の検索リストの作成
3.検索結果からの店の選択

初期入力する情報は駅名などの入力地点名と、その地点から近い順に抽出する店名の件数：N である。

3.1. エリア内の店名の取得

Foursquare の venue search API を用い、入力地点名の地点から近い順にN件のvenueから店名を抽出した。店名のカテゴリは「和食店」「寿司屋」「居酒屋」「もぐり酒

[4] https://developer.foursquare.com/

場」を対象とした。

抽出結果には重複を含めた様々なノイズが含まれており、表現の多様性が大きいため、下記のようなフィルタリングを行った。

1. 鍵括弧(「」、『』)内は店名とする
   例: 和風創作料理「菜」 SAI
2. 括弧(半角、全角)および括弧内の文字列は店名ではないものとする
   例: ティータ(居酒屋)
3. スペース(半角、全角)で区切る
   3-1. 最後の要素が「〜店」の場合は支店名として除去しておく
   例: すみやき屋 串軍 西青梅店 → 「西青梅店」は支店名
   3-2. 残った要素のうち末尾の要素のみを店名として抽出する
   例: すみやき屋 串軍 → 店名は「串軍」

3.2. 店名の検索リストの作成

日本酒の銘柄と同様、店名を示す表現には店名以外の意味で使われるものが多数存在する。また、チェーン店に限らず、複数の地域に同名の店が存在することも多い。そこで、下記のケースに関しては、駅名などの地点名が同じデータ(ツイート)中に存在する場合のみマッチするように AND 検索条件を付加するようにした。

- venue に駅名を含み、かつ、店名に駅名を含まないもの
- 全角文字一文字の店名
- ひらがな語・カタカナ語のみの店名
- 分類語彙表に載っている語の店名

すなわち、検索条件は基本的に

"<店名>"

もしくは

("<店名>" AND "<地名>")

となる。

<店名>と<地名>には具体的な名称が入るが、例えば東京駅の場合は、<地名>部分を

("東京" OR "八重洲" OR "丸の内")

のように設定した。

3.3. 検索結果からの店の選択

前ステップで作成した検索条件用いて、候補店数N件の検索を行い、その結果をチェックした。

例えば、東京駅周辺の500件の検索条件の場合、

("<店名>" AND ("東京" OR "八重洲" OR "丸の内"))

の形式になったのは114件である。500件の検索条件のうち、約半数の241件(その中でAND条件付が69件)は該当するデータが存在しなかった。該当データが存在した259件中、データ数が10件以上あるものが110件存在した(その中でAND条件付が13件)。当初は、このデータ数が多いものに関して、評価表現の有無などでランキングを行えば良いのではないかと考えていたが、実データを見たところ、むしろ、データ数が多いものにはノイズが多いことが分かった。AND条件付で地名と共起している店名であっても、例えば、「車」や「響」といった表現は、店名以外の意味で使われていることが多い。したがって、良い店を選択する作業を効率的に進めるためには、ノイズの多い店名を避けて、店名が言及されている可能性が比較的高い低頻度の店名のデータに絞った方が良い。

結果的に東京駅周辺に位置する131店の店名の表現を含むデータ373件に対象を絞り込み、店名がマッチした文字列をハイライトさせた状態で元データ(ツイート)に目を通すと、

- 店名ではないもの
- 店名ではあるが他地域の店に関するもの
- 宣伝的なもの

が多く、実際に東京駅周辺の店に関して、美味しい店であることを示唆しているデータは10件程度に過ぎなかった。こうして浮かび上がった店に関して、レビューサイトを含む外部情報を参照した上で、実地調査対象とする店を決めるという手法をとった。

4. 評価

前節のアプローチによって、東京駅周辺だけでなく、赤坂、渋谷など 10 以上の地域において日本酒の美味しそうな店の情報をツイートから取得し、店を選んで、2 名から十数名の有志による実地調査を行った。前節で示した東京駅周辺の例では調査対象店数を 500 としたが、通常は 200 から 300 程度であり、最終的に目を通したツイートの件数は数十から数百件、店の選択にかかった時間は概ね 20 分から 30 分程度であった。

実地調査を行った結果として、有効性を評価するため、参加者に満足度と再訪希望度を回答してもらった。満足度は 5 段階(5.とても満足, 4.まあ満足, 3.普通, 2.いまひとつ, 1.残念)で、再訪希望度は 4 段階(5.是非また行きたい, 4.また行ってもかまわない, 2.あまり行きたくない, 1.二度と行きたくない)で評価してもらった。

集計できた 10 店に関して、満足度 3 以下、再訪希望度 2 以下が付いたことはなく、4 店では満足度と再訪希望度の両方に 5 を回答者全員(最大 7 名)が付けた。

本手法で選択して実地調査した店は 12 店であるが、そのうち少なくとも 4 店は参加者が実際に再訪しており、満足度の高い店を選ぶことができたと考えられる。

5. 結論及び今後の課題

大量のツイートから日本酒の美味しい店を発掘するというタスクを設定した当初は、「数百万件のデータを対象にすれば美味しい店のランキングリストができるのではないか」という期待もあったが、試行錯誤の結果、宣伝的なツイートを除くと、一つの店に対するツイートの件

数は数件程度と少ないことが多く、最終的には、店名で検索したツイートを人手でチェックするというアプローチで店の発掘を行った。結果的には、日本酒の美味しい店を選ぶきっかけとなる記述が実際にマイクロブログのデータに存在することを確認でき、それを頼りにして、実際に満足度の高い店を発掘することができた。

当初予想したよりも店名の言及が少なく、その中から良い店が発掘できたということは、わざわざ言及する価値のある店名がツイートに含まれているという解釈もできそうである。日本酒の美味しい店を発掘するための知識源としてのマイクロブログの有効性が認められたため、このタスクを進展させる価値があると考えられる。

技術的な課題として、自然言語処理の観点からは、語義解消の重要性を強く感じた。前述の通り、店名や酒の銘柄には「彩」や「南」など汎用性の高い表現が使われることが多く、今回の取組みでは、これら汎用性の高い表現を対象外としてしまった。自然言語処理およびテキストマイニングのアプリケーションとして、このタスクの高度化を目指すにあたっては、正面から取り組むべき課題である。

今回は、「日本酒」もしくは「ビール」という表現をベースにツイートを収集したが、日本酒に関してつぶやいているユーザのツイートを網羅的に収集して分析対象とすることで、より良い情報が得られる可能性がある。日本酒という観点では、その味わいの表現に関する研究[6,7]の知見なども取り込むことで、ユーザの嗜好性や精通性に踏み込んだ分析につながる可能性が考えられる。

本タスクを行うにあたっては、マイクロブログ以外のデータも参照した。店名と店の位置の情報をFoursquareから取得しただけでなく、実地調査の店を選択するにあたっては、マイクロブログ以外のレビューサイトの情報を参照し、メニューや価格帯を確認することで、リスクを抑える必要があった。価格や店の連絡先などはマイクロブログのデータに含まれていることが期待し難く、もしもそういった情報を含んでいる場合は宣伝的な意図をもって発信されている可能性が高い。その反面、マイクロブログのデータがなければ選択肢に入らなかったと思われる店も存在したことで、良い店を選ぶきっかけとなる情報を得るうえでのマイクロブログの有用性を実感できた。特に、小規模の店はレビューサイトへの書き込み件数が少なく、その大半が評価の高い書き込みであっても、サクラの危険性を考慮しているためか、レビューサイトの評価点が低くなる傾向にあるように感じられる。今回の取組みでは、レビューサイトでのランキングが 100 位よりも低いのに実地調査で高い評価を得られた店を発掘することができた。マイクロブログのデータの特長としてユーザ名の単位でデータを分析できることから、同一ユーザの他のツイートを見てサクラを排除できる可能性が高く、プライベートなクチコミに近い情報を活用できるメリットを感じた。タスクの目的に応じて適切なデータの情報を組み合わせることが重要である。

6. おわりに

本研究の取組みの背景には、レビューサイトのサクラを使ってでも集客したいという店の対極にある、集客よりも店を良くすることに注力し、内輪のクチコミだけで成り立っているような良心的な店を発掘し、光を当てる仕組みの実現につなげたいという願いもあった。

実はこの取組みで発掘し、筆者が何度も通うようになった店の一つが 2014 年末に閉店してしまった。夫婦二人で切り盛りしている小さな店であり、レビューサイトでのランキングは非常に低い（100 位にも入らないレベル）ながらも、一緒に行った人から良い店だと喜ばれ、その人が別の人を連れて再訪するという、まさにクチコミで客が増える典型的な店であった。「閉店してしまうので、その前に一度来店して欲しい」という連絡を店主から受け、予約して行ってみたところ、店の閉店を惜しむ客で盛況であった。店主が「常にこれだけのお客さんが入っていたら、店を閉めずに済んだのに」と話していたのが残念であった。このような店がもっと注目され、繁盛するような仕組みを実現する一端になれば幸いである。

謝辞

本稿のアプローチによる店の発掘と実地調査を進めるにあたっては、IBM TEC-J SIG Text Mining のメンバーをはじめとする多くの方々のご協力をいただきました。ここに記して深謝いたします。

参考文献

[1] 那須川哲哉．テキストマイニングを使う技術/作る技術―基礎技術と適用事例から導く本質と活用法．東京電機大学出版局, 2006

[2] 斎藤翔太，伊川洋平，鈴木秀幸，村上明子．Twitter を用いた災害情報の早期発見，電子情報通信学会技術研究報告，信学技報 NLC2014-2, pp.7-12. 2014

[3] Akiko Murakami and Tetsuya Nasukawa. Tweeting about the Tsunami?: Mining Twitter for Information on the Tohoku Earthquake and Tsunami. WWW 2012 SWDM'12 Workshop, In WWW '12 Companion Proceedings, pp. 709-710. 2012.

[4] Wei-Dong Zhu, et al. IBM Watson Content Analytics: Discovering Actionable Insight from Your Content. An IBM Redbooks publication. ISBN-10:0738439428. 2014.

[5] Hiroshi Kanayama and Tetsuya Nasukawa. Unsupervised lexicon induction for clause-level detection of evaluations. Natural Language Engineering, Volume 18, Issue 01, pp 83-107. 2012.

[6] 大塚裕子，日本酒を味わう表現についての分析，人工知能学会第 2 種研究会ことば工学研究会資料 14, pp.31-36, 2003.

[7] 福島宙輝，味わいの言語化を支援する「日本酒味わい図式」の提案，人工知能学会第 2 種研究会ことば工学研究会資料 44, pp.1-4, 2013.