F-031

マイクロブログにおける発話シミュレーターに向けた基礎研究
Utterance Simulator in Microblog

新井雅也† 川村秀憲† 鈴木恵二†
Masaya Arai Hidenori Kawamura Keiji Suzuki

１．概要

本研究では、マイクロブログにおける発話シミュレーターの開発に向けて、Twitter から得られる発言を基に、ユーザのプロファイル情報を取得することを目指す。取りかかりとして、比較的粒度の粗いユーザのプロファイル情報が推定可能であるか検討する。また実験を通して、ユーザ居住地域を札幌・東京・大阪の 3 地域の中から推定することで、Twitter からプロファイル情報の取得可能性について示す。

２．はじめに

ソーシャルネットワーク技術の発達により、お互いが知り合いではない人同士が関係も持つことや情報をリアルタイムに取得することが容易になってきている。とりわけ、エジプトで行われた反政府運動の呼びかけや東日本大震災の際の情報取得にマイクロブログサービスの一つである Twitter[1] が多用されたという事実は比較的新しい。このような背景を受けて、現在ではマイクロブログを対象とした研究が盛んに行われている。

Twitter はニュースやスポーツ情報、イベント、映画、雑談といった普段のライフスタイルの状況に密着しての利用が多く見られる[2]。そのため、広告を始めとした情報配信サービスやユーザとのコミュニケーションを通したアンケート調査への応用が期待される。ここで、Twitter を情報伝搬の媒体として精度良く利用するには、ユーザのプロファイルを把握することで伝えるべきターゲットを明確にし、パーソナライズ化された情報を適切なコミュニケーションを通して振る舞うことができるかどうかが鍵となる。

本研究では、マイクロブログサービスである Twitter 上にて、ユーザに対して適切にコミュニケーションをとることが可能な発話シミュレーターの開発を目的とする。本稿では、始めの手掛かりとして、コミュニケーションに必要となるユーザのプロファイルをどのように取得するかという点について検討を行う。加えて本稿では、そもそも Twitter 上にてプロファイル情報が取得可能であるのかという点についても考察する。

本研究は、テキストマイニングを通して断片的な情報から人のプロファイル情報をリアルタイムに推定を行う点や、条件にマッチしたユーザをどのように見つけることができるかといったデータマイニングの点で学術的に知見を与えるものと期待できる。本研究を行うに先立ち、次章にて関連研究と比較した本稿の位置付けを行う。

† 北海道大学 大学院情報科学研究科

3. Twitter に関する関連研究について

Twitter は 2007 年に開始されたマイクロブログサービスである。世界中における 1 億 9000 万人以上ものユーザがサービスを利用しており、毎日 6500 万件以上の発言が世界中を飛び交っている[6]。一回の発言は 140 文字以内に制限されていることに加え、話題やニュースなどの情報の伝搬速度は非常に速く[2]、また見知らぬ人とフォロー関係を結ぶことで、多用な意見や情報が得られやすいという点から、本来のブログサービスとは異なる性質を有する。

現在では、フォロー関係が及ぼす発言の影響力や話題の広がり方などのネットワーク分析を対象とした研究[2][3]や, エージェントベースアプローチからのネットワーク分析[4], ユーザが興味を持つ内容を含む URL の推薦[5]といった研究がなされている。

ユーザの発言内容自体を分析することで、ユーザの興味分野や発言の傾向などを得る研究[6]も行われているが、ユーザに対して適切にコミュニケーションをとるための仕組みやユーザ自身がどのようなプロファイルを有しているかといった研究はあまり類を見ない。

そこで本稿では 3 章にして示す実験を通して、基礎的なユーザのプロファイル推定方法について検討を行う。

4. ユーザのプロファイル推定の実験と環境設定

本研究において取得を目的とするプロファイル情報は「30 代の主婦」、「自動車を所有している 20 代男性」、「北海道に住む大学生」といった広い意味でのユーザそのものに関する情報である。今回は基礎的なプロファイル情報である居住地域を対象として推定可能であるか実験を通して検証を試みる。実験に先立ち、ユーザの居住地域はプロフィール欄に記載されているものが正しいということを前提において実験を行う。本実験の概要を図 1 に示す。

本実験において対象とする地域は札幌近辺・東京近辺・大阪近辺の 3 つ地域である。

まず始めに、3 つの地域内に居住している任意のユーザ 20 名ずつ選出し、Twitter 上の発言を取得する。本実験では、ユーザ 1 人当たり 500 件以上、3200 件以下の発言数を取得した。次に、地域ごとのユーザの発言に対して形態素解析を行い、地域名詞のみを抽出する。本稿では形態素解析のツールとして、京都大学情報科学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所が共同で開発した MeCab を用いた。ここで、本稿における地域名詞とは MeCab にて地域名詞と分類されるものを指す。

地域名詞を抽出した後、3 つの地域別に地域名リストを作成する。一つの地域名リストを作成するに当たり、約

50000 件の発言を対象に形態素解析を行った。ある地域名リスト内においては、明らかに該当する地域と異なる地域名が含まれているため、代表的な地域名は手動で取り除いた。また、地域名リスト同士を比較し、重複する地名となる名詞は全て削除した。

次にプロファイルを推定する対象となるユーザの発言を取得し、MeCab を用いて形態素解析を行った後、地域名詞のみを取り出す。ユーザの発言から得られた地域名詞が含まれる地域名リストの地域を数え、全体の数に対する割合を算出する。

最後に、Twitter 上においてユーザの個人情報欄に記載されている住所と発言から算出された地域名詞の割合を比較することで、プロファイル情報が推定されているか否かを判断する。

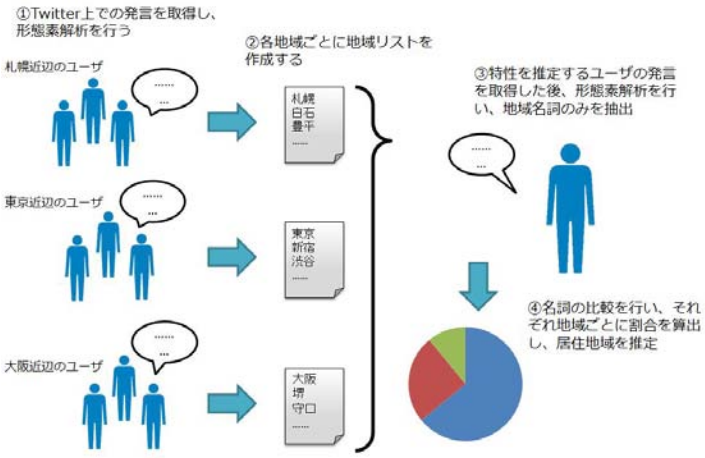

図 1. 実験の概要図

5. 実験結果及び考察

前章にて示した実験において、それぞれの地域に住む任意のユーザを 3 名ずつ選び、得られた結果を図 3 に示す。

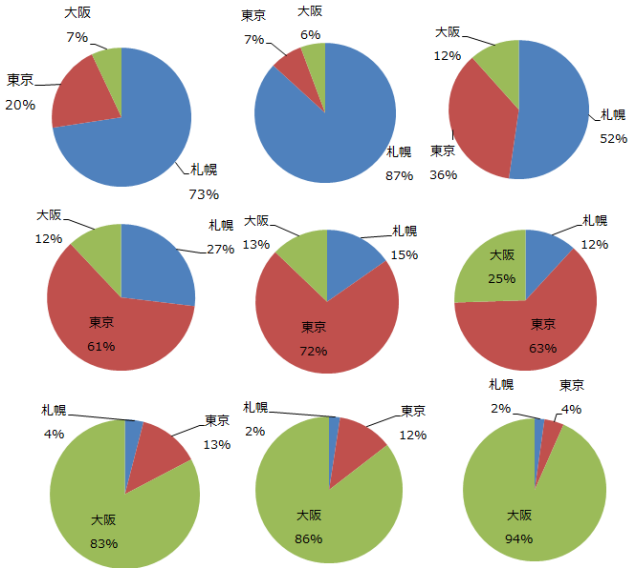

図 2. プロファイル推定の実験において得られた結果を示すグラフ

図 2 における 9 つの円グラフにおいて、上部 3 つのグラフは札幌に住んでいるユーザ 3 名、中央 3 つのグラフは東京に住んでいるユーザ 3 名、下部 3 のグラフは、大阪に住んでいるユーザ 3 名における発言内の地域名詞の割合である。今回の実験でプロファイル推定の対象としたユーザの分析対象となる発言数はそれぞれ約 3000 件とし、bot などのユーザに替わって自動的に発言するアカウントは対象としていない。

グラフ全体を見てみると、全てのユーザにおいて、発言した内容から得られる地域名詞が自分の住んでいる地域名リストと一致していることがわかる。しかし、右上の円グラフにて示される札幌に住んでいるユーザにおいては、他のユーザの割合と比較し居住地域に関する地域名詞の発件数が少ない。ユーザの居住地域とは関係なく、出張や旅行で行った地域に言及する発言やニュース記事を引用した発言などにより、品詞のみでは的確に抽出することが困難であることが原因と予想される。そのため、キーグラフなどの共起分析を用い、品詞間の関係を考慮することで、より詳細なプロファイル情報推定の精度向上に寄与するものと考えられる。

また分析対象としたユーザが発言した地域名詞を見てみると、札幌や東京など市や都道府県単位での件数が多い一方、具体的な住所を含んだ発言やリツイートをしている例も少なくない。この点に関しては、tf-idf 法を用いて特徴的な単語を重み付けすることでより精度が増すものと予想される。

6. まとめと今後の展望

本稿では発話シミュレーター開発に先立ち、ユーザプロファイル情報を取得するための検討を行った。実験として、ユーザの居住地域を大きく 3 ヶ所に分け、推定を試みた。実験の結果、任意のユーザにおいてプロファイル情報を取得できる可能性を得ることができた。

今回は比較的粒度が粗い情報の取得を試みる実験を行った。今後の研究の方向性として、文脈・構文解析、共起分析などの手法を利用し、世代や職業などある程度粒度の細かいプロファイル情報の取得を目指す。

[1] http://twitter.com
[2] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In World Wide Web Conference. ACM Press, (2010) 591-600
[3] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In Proc. of the third ACM international conference on Web search and data mining. ACM (2010) 261-270
[4] N. Nakamura and H. Deguchi, Cognitive-Costed Agent Model of the Microblogging Network, Agent-Based Approaches in Economic and Social Complex Systems VI, Agent-Based Social Systems (2011) Vol.8, Part III, 75-84, Springer
[5] Chen, J., Nairn, R., Nelson, L., et al. Short and Tweet: Experiments on Recommending Content from Information Streams. Proc. CHI '10, ACM Press (2010)
[6] K. Tao, F.Abel, Q. Gao, G.-J. Houben. TUMS: Twitter-based User Modeling Service UWeb Workshop Proc. (2011) 60-75