一人称所有格を用いたプロフィール推定

那須川 哲哉　　西山 莉紗　金山 博　　吉田 一星　　大野 正樹
日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

ソーシャルメディアの発達に伴い、従来はアンケートやフォーカス・グループ・インタビューといった手法を用いてきた行動や嗜好の調査に、ソーシャルメディアで公開されている情報を活用できるのではないかという期待が高まっている。ソーシャルメディア上の膨大なデータを対象にすることで、より豊富な情報を、より低いコストで、より詳細に分析できると考えられる。

しかし、現時点ではソーシャルメディアの分析がアンケートやフォーカス・グループ・インタビューと置き換わるまでに至っていない。その主な原因の一つがプロフィール情報の不足にある。例えば、特定商品のマーケティングを目的とした調査においては、どういった世代のどういった立場の人々がその商品や競合商品に対してどう考えているかを捉えることが重要である。ソーシャルメディア上のデータに、特定商品や競合商品に対する記述があっても、その発信者の人となりが分からなければ、その情報の活用は難しい。

ソーシャルメディアでは、詳細なプロフィール情報の入力を必要としないシステムが多い。システムがプロフィールとして公開している情報に依存せず、システム上で記述された通常のテキストメッセージからプロフィールを推定するのが本研究の目的である。

本稿では、「私の」「うちの」「僕の」といった一人称所有格表現に着目し、その表現が修飾する名詞句からプロフィールを推定する取り組みについて示す。例えば、「私の夫」という表現を含むメッセージの発信者は配偶者を持つ女性であり、「うちの孫」という表現を含むメッセージの発信者は孫を持つ祖父母の世代であると推定する。この単純なアプローチによって推定できるプロフィール情報の種類や精度など、その有効性と、その結果を用いた分析例を日本語と英語のデータで示す。

2. プロフィール推定研究における位置付け

テキストからのプロフィール情報の推定の取り組みは数多い。発信者の性別[2,3,5,10]や年代[1,4,10]、支持政党[6,8,10]などがテキスト情報から推定される。しかしいずれの取り組みも基本的には分類問題を解くアプローチを取っており、機械学習を用いて正解付きデータから分類器を作成するものが多い。

それに対し、我々の取り組みでは、情報抽出の形でプロフィール推定を実現する。学習データを用意する必要が無い上、特定の表現の存在によってプロフィールを推定できることから、情報検索のように任意のプロフィールに属するユーザーを探し出すことが可能になる。例えば、犬を飼っているユーザーを探したい場合、ソーシャルメディアのデータから「うちの犬」「私の犬」といった表現を含むメッセージを検索し、検索されたメッセージに紐付いたユーザーIDのリストを取得すればよい。

3. 一人称所有格の修飾する名詞句

まず、日本語のTwitterデータ(Tweet)を対象として、一人称所有格がどのような名詞句を修飾しているか調査した。

一般公開されている Twitter API[1]を利用して二百万件弱の Tweet を取得したデータを対象に、IBM® Content Analytics (ICA) Version 2.2 [11]のパタン抽出機能を利用し、形態素解析結果の形態素の並びにおいて「私」「自分」「俺」「うち」「僕」「わたし」「あたし」「オレ」の直後に「の」を介して続いている名詞句を抽出した。その結果、調査対象とした 1,827,104 件の Tweet の 1.6%に相当する 29,126 件から名詞句が抽出された。抽出された名詞句の一部を表 1 に示す。「こと」「中」など抽象的な表現が含まれており、抽出された名詞句の全てがプロフィール情報に使えるわけではないと分かる。

表 1:　2011 年 1 月 1 日から 2012 年 4 月 10 日までの期間にランダムに抽出した 1,827,104 件の Tweet データにおいて一人称所有格に続く名詞句とそれを含む Tweet の件数

(各名詞句を含む Tweet の件数が多い順にソートした結果の上位 15 表現)

一人称所有格に続く名詞句	Tweet 件数
こと	1212
中	523
会社	396
事	376
方	336
子	298
説明書	287
せい	283
場合	282
嫁	278
前	265
名前	232
頭	208
もの	205
後輩	203

[1] https://dev.twitter.com/docs/api

表 1 に示された高頻度の(Tweet 件数が多い)表現の中には Twitter の bot 機能を使って自動作成されたメッセージから抽出された名詞句が含まれている。例えば、「私の説明書」「僕の説明書」という表現を伴い bot 機能の設定内容へのリンクを示す Tweet が数多く存在する結果として「説明書」が表 1 に含まれている。一方、表 1 には「会社」「子」「嫁」のようにプロフィール情報を示唆する表現も含まれており、適切な表現を選択することで、プロフィール情報として利用できると考えられる。

表 1 のベースとなった二百万件弱の Tweet からは、一人称所有格の直後に続く名詞句として 7,129 の異なり表現が抽出された。そのうち高頻度の表現に関してプロフィール情報としての有効性を検討した結果、下記の 3 タイプのプロフィール情報は推定できる可能性が高いという知見が得られた。

- 家族構成
  「嫁」「旦那」「娘」「息子」「姉」「妹」「兄」「弟」など
- 所有物
  「犬」「猫」「iPhone」「パソコン」「車」など
- 所属
  「会社」「店」「大学」「学校」「病院」など

4. 一人称所有格を利用したプロフィール情報の有効性

一人称所有格に修飾される名詞句が実際に発信者の所有対象であるか、例えば「私の車」という表現はその発信者が実際に車を所有していること示すかを、その表現を含む Tweet を人手で読み判断する実験を行なった。また、一人称所有格を伴わずに単に「車」という表現を含む Tweet が、発信者の車の所有状況を示唆していないかも、同様に人手で読み判断した。読み手の主観に基づく判断のため、一人の判断結果をさらに二人で確認した。その結果を表 2 に示す。

例えば「俺の iPhone 調子悪い！」「うちの犬かわいいー」という Tweet の場合は、各々発信者が iPhone や犬を所有していると判断した。それに対し、「自分の車じゃないよ～」「自分の車ほしいー」「うちの犬にしたい」といった表現の場合は、発信者が車や犬を所有しているとは判断できないものとした。

表 2: 「車」「iPhone」「犬」「猫」が出現する Tweet において、それを発信者が実際所有していると判断可能かの評価結果[2]

	車	iPhone	犬	猫
一人称所有格＋単語で言及している際に所有しているとみなす場合	97.00% (97/100)	97.00% (97/100)	79.00% (79/100)	87.00% (87/100)
単語のみの言及であっても所有しているとみなす場合	29.10% (16/55)	40.00% (40/100)	21.40% (21/98)	17.20% (17/99)

[2] ()内は(所有していると判断したツイートの件数 / 人手で調査したツイートの総件数)

表 3: Twitter のプロフィール欄を参照した車の所有者の検索と一人称所有格を用いた車の所有者の検出の比較

	プロフィール欄に「車」が含まれるユーザーを検索した場合	プロフィール欄に「ドライブ」が含まれるユーザーを検索した場合	全日本語ユーザーのツイートに本手法を適用した場合
精度[3]	35.7% (45/126)	89.1% (106/119)	97.0% (97/100)
獲得ユーザー数	206,809	47,051	568,000

表 2 で「犬」と「猫」の精度が低いのは、これらの表現がペット以外に奴隷や手下などの意味で使われている例が多かったためである。「車」「iPhone」のように語義的曖昧性が無い物体の場合は、90%を超える高い割合で所有格が実際に所有を示す可能性が高いという結果が得られた。また、所有格を伴わずに出現した表現を所有していると判断できる割合は低いことが分かった。

次に、本手法を使う場合と Twitter が備えているプロフィールページの情報を使う場合との比較を試みた。

「車」の所有者を対象として、本手法では、Tweet 中に「私の車」「自分の車」「うちの車」といったメッセージを記述したユーザーを対象として車の所有者を検出した。Twitter のプロフィール欄は自由記述形式のため、検索サービス[4] を用い、プロフィール欄に「車」もしくは「ドライブ」という表現を含むユーザーを対象として車の所有者を検出した。その結果を表 3 に示す。

プロフィール欄に「車」や「ドライブ」という文字列が入っていても「車、バイク、パーツの買取専門店の bot です」「ネットの中をドライブしていきます。」のように、必ずしも車やドライブが趣味とは限らないため、車の所有者の推定という観点からは、本手法を用いた方が精度が高いという結果が得られた。また、どれだけの数のユーザーを特定できるかという観点からは、プロフィール欄に「車」を含むユーザー数が約 20 万人であり、「ドライブ」を含むユーザー数が 5 万人弱であった。それに対し、ランダムに収集した 1,000 人の日本語ユーザーに対して本手法を適用した結果、1.9%(19 人の)ユーザーが車を所有していると推定されたことから、Twitter 上の全日本語ユーザーは 2,990 万人という試算[7]を基にして概

[3] 各 100 件程度のサンプル調査を行ない、その中で車やドライブが趣味もしくは実際に車を所有していると判断されたケースの割合を精度として算出。括弧内は(所有していると判断したツイートの件数 / 人手で調査したツイートの総件数)

[4] http://twpro.jp

算したユーザー数は 50 万人規模となった。精度との兼ね合いから考えても、車の所有という特定プロフィール情報のユーザーを集め、その嗜好や動向を調査するという観点からは、本手法の有効性が高いと考えられる。

5. 英語データにおける一人称所有格を利用したプロフィール情報の推定

一人称所有格という概念は言語に依存しないことから、本手法を英語のデータに適用する実験を行なった[5]。その結果、一人称所有格を示す my 及び our は日本語における「私の」「うちの」などの表現よりも出現頻度が高いことが分かった。調査対象とした 500 万件の Tweet の 8.6%（日本語では 1.6%）にあたる 430,811 件のデータにおいて、my もしくは our が名詞句に先行している表現が抽出された。その結果の一部を表 4 に示す。

表 4 に示されている高頻度の（Tweet 件数が多い）表現は必ずしもプロフィール情報として有効ではないものの、日本語データからの抽出結果（表 1）と比較すると、抽象名詞が少なく、プロフィール情報としてより有効性の高い表現が抽出されている傾向が見受けられる。

5.1. 特定プロフィールに属するユーザーID の収集

特定プロフィールに属するユーザーをどの程度のペースで収集できるかを英語データで試みた。特定プロフィールとしては下記 3 種類を対象とした。

- 車の所有者
  “my car”を含む Tweet の発信者
- ギターの所有者
  “my guitar”を含む Tweet の発信者
- 祖父母（孫を持つ）
  “my grandson”, “my granddaughter”, “my grandchild”, “my grandkid”, “our grandson”, “our granddaughter”, “our grandchild”, “our grandkid” のいずれかを含む Tweet の発信者

Twitter API の制限を超えないよう、基本的に 3 分で 1500 件以内の Tweet を収集するクローラを流し続け、収集された Tweet のユーザーID の累積異なり数を 10 時間ごとに記録した結果を図 1 及び図 2 に示す。

図 1 に示されている通り、60 時間で 14 万人以上の車の所有者のユーザーID を集めることができた。しかも 60 時間経っても増加傾向が鈍化しなかった。

また、ギターの所有者や祖父母は、車の所有者に比べると明らかに少ないため、図 2 の通り、ユーザーID の収集ペースが遅いが、それでも 60 時間で数千人規模の ID を集めることができ、さらに時間をかければ、より多くの ID を集めることができる見通しが得られた。

[5] 日本語同様 ICA V2.2 を用いており、各表現の複数形や大文字小文字のバリエーションも対象となっている。

表 4: 2012 年 9 月 30 日から 2012 年 10 月 10 日までの期間にランダムに抽出した 5,000,000 件の Tweet データにおいて一人称所有格に続く名詞句とそれを含む Tweet の件数

（各名詞句を含む Tweet の件数が多い順にソートした結果の上位 15 表現）

一人称所有格に続く名詞句	Tweet 件数
mom	11017
phone	8458
friend	6209
hair	5966
dad	5147
heart	4736
day	4734
mind	4577
good friend	3818
sister	3794
house	3685
brother	3648
birthday	3562
room	3521
eye	3423

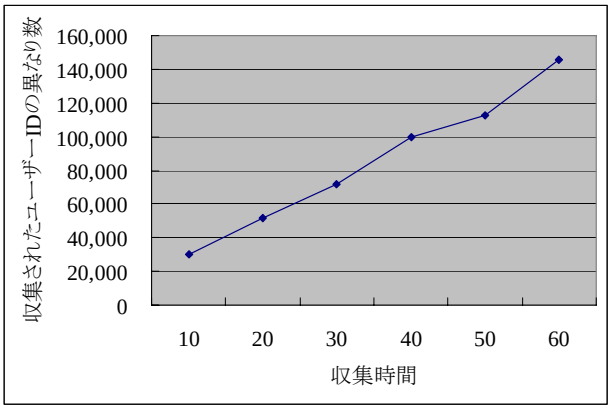

図 1: 車の所有者のユーザーID の収集結果

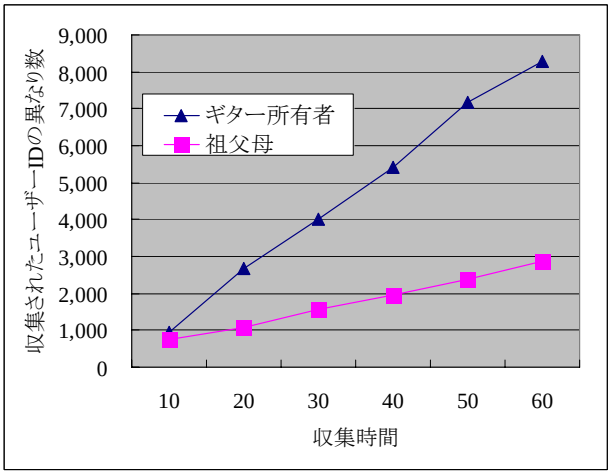

図 2: ギターの所有者及び祖父母のユーザーID の収集結果

表 5: 一人称所有格に基づきプロフィール推定したユーザー及び各ユーザーが発信した Tweet の収集数

プロフィール	ユーザー数	Tweet 数
ギター所有者	30	30,000
祖父母	30	29,577
息子のいる母親	30	30,000
娘のいる母親	30	30,000
息子のいる父親	30	30,000
娘のいる父親	30	30,000
合計	180	179,577

5.2. 特定プロフィールに属するユーザーの特徴の分析

本手法を用いて特定プロフィールに属するユーザーID を収集した結果を活用し、そのユーザー群の Tweet を分析することで実際に行動や嗜好の特徴が検出できるかを調査した。表 5 に示す 180 名のユーザーに関して各人ほぼ千件ずつの Tweet を集めて分析を行なった結果、主な傾向として、以下のような知見が得られた。

- ギターの所有者は学生の割合が高い
  その 37%（30 人中 11 名）のユーザーの Tweet 中に “my school” もしくは “our school”が出現しており、他の 150 ユーザーにおける割合（6%）を大きく上回る
- 祖父母は photo への関心が高い
  その 67%の Tweet に“photo"が出現しているのに対し、他の 150 ユーザーでは 51%
- 母親は pregnancy への関心が高い
  その 37%（60 名中 22 名）に“pregnancy”が出現しているのに対し、他ユーザーは 8%（120 名中 9 名）

また、Tweet を発信している時間帯という観点からは、息子のいる母親と娘のいる母親、及び息子のいる父親と娘のいる父親の Tweet 発信時間帯の類似性が高く、生活時間帯は、息子を持つか娘を持つかよりも、父親か母親かにより影響される可能性が高いと考えられる。

6. おわりに

日本語では「私の」「うちの」「僕の」、英語では my もしくは our といった一人称所有格表現に着目し、この所有格表現が修飾する名詞句からプロフィールを推定する手法とその有効性、及び分析への活用例を示した。

従来のプロフィール推定が基本的に分類問題を解くアプローチだったのに対し、情報抽出の形でプロフィール推定を実現できる意義は大きい。膨大なソーシャルメディアのデータを対象とし、工夫次第で多様なプロフィールのユーザーを特定できる。 例えば、“my XXX”の XXX に高額商品名を指定し、その所有者や利用者を特定したり、 “My Nth birthday”の N を指定することで各年齢のユーザーを特定したりすることが可能になる。

他にプロフィール推定につながる表現としては、日本語の場合、「医者仲間」「テニス仲間」における「仲間」などを挙げることができる。

謝辞

本研究を進めるにあたり、各種実験をはじめとして、青山学院大学の八木春麻氏に多大なご支援をいただきました。ここに記して感謝いたします。

参考文献

[1] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, Jonathan Schler. (2009). Automatically profiling the author of an anonymous text. In *Communications of the ACM* 52(2): pp. 119-123.

[2] John D. Burger, John Henderson, George Kim, and Guido Zarrella. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1301-1309.

[3] Nikesh Garera and David Yarowsky. (2009). Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*: Volume 2 - Volume 2 (ACL '09), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 710-718.

[4] Dong Nguyen, Noah A. Smith, and Carolyn P. Rose. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 115-123.

[5] Jahna Otterbacher. (2010). Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. ACM, New York, NY, USA, pp. 369-378.

[6] Pennacchiotti, Marco and Popescu, Ana-Maria. (2011). Democrats, Republicans and Starbucks Aficionados: User Classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 430-438.

[7] Semiocast. Geolocation analysis of Twitter accounts by Semiocast. http://semiocast.com/publications/2012_01_31_Brazil_becomes_2nd_country_on_Twitter_superseds_Japan. 2012. Accessed on 2012-11-23.

[8] Matt Thomas, Bo Pang, and Lillian Lee. (2006). Get out the vote: determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 327-335.

[9] W. Wu, B. Zhang, and M. Ostendorf, (2010). Automatic generation of personalized annotation tags for twitter users. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT'10*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 689-692.

[10] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. (2012). Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors, In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pp. 387-390

[11] Wei-Dong Zhu, Asako Iwai, Todd Leyba, Josemina Magdalen, Kristin McNeil, Tetsuya Nasukawa, Nitaben Patel, and Kei Sugano (2011), IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content, IBM Redbooks publication. http://www.redbooks.ibm.com/abstracts/sg247877.html