5S-5

ウェブ上での言及と実体の同一性判定

竹澤 友博† 松尾 豊† 石塚 満†
†東京大学大学院情報理工学系研究科

1 概要

人間は，文字をはじめとするシンボルを介し，コミュニケーションを成立させているが，シンボルに託された様々な物理的な意味を共有することで意思疎通が図られ，その際に使われているシンボルは「グラウンドしている」と表現される．これに対し，計算機システムで扱われているシンボルでは，その物理的な意味を明確に表現することが困難でグラウンドしておらず，これをいかにグラウンドさせるかがシンボル・グラウンディング問題である．一方，近年，急速に広がるウェブ上のソーシャルメディアにおいては，ユーザアカウントに代表される実体と，それに対するさまざまなテキスト上での言及が存在する．これを一致させることができれば，限定した範囲内ではあるが，シンボルグラウンディング問題に対する一つの解を提示することになると考えられる．本研究では，ソーシャルメディアにおけるシンボル・グラウンディング問題，特に，固有表現に特化し，言及としてのシンボルと，それが指す内容を特定する方法を構築し，検証する．

2 背景

人間は，文字をはじめとするシンボルを介し，コミュニケーションを成立させている．例えば，音声による対話，ジェスチャー，標識，本やウェブ上のテキストなど，さまざまなシンボルが使われている．この場合，シンボルに託された物理的な意味を共有することで，意思疎通が図られる．このとき，使われているシンボルは「グラウンドしている」と表現される．人工知能のコミュニティでは，従来から，シンボル・グラウンディング問題が指摘されてきた [1].

計算機システムで扱われているシンボルでは，その物理的な意味を明確に表現することが困難であり,「グラウンドしていない」と表現される．これをいかにグラウンドさせるかがシンボル・グラウンディング問題である．システムが環境と相互作用することで，システムが物理的意味を理解できる可能性がある．

一方で，ウェブではさまざまなソーシャルメディアが存在している．例えば，ブログ，ソーシャルネットワーキングサービス（SNS）などである．

A Study on Identification of Mentions and Entities on the Web
†Graduate School of Information Science and Technology, The University of Tokyo

これらのソーシャルメディアでは，ユーザはハンドルネームを使うことで匿名的な活動をしている場合がある．例えば，***などではこの傾向が顕著である．一方，実名と密接に結びついたサイトもある．海外では LinkedIn というビジネス指向の SNS があり，そこでは主に実名が用いられる．国内でも，twitter や facebook は，ユーザは実名を用いることが多い．これは，サイトの性質によるもので，twitter は多くの人に向けたブロードキャスト，facebook は実世界の友人関係を反映したコミュニケーションを指向しているため，実名が推奨，もしくは利便性のために実名がよく用いられる．

【実名の例】例えば，twitter 上で堀江貴文氏のアカウントは@takapon_jp であり，名前は「堀江貴文 (Takafumi Horie)」, プロフィールとしては

> 六本木で働いていた元社長です．巷ではホリエモンともよばれています． あ，メルマガは http://goo.gl/IMkx で登録できます．お問い合わせは，takapon@mag2.com へ．12/22-26 までクリスマスキャロルってミュージカルにでます

と記入されている．

このように，ソーシャルメディア上では，ユーザがアカウントを持っており，シンボルグラウンディングという観点からは，アカウントで一意に定められるユーザという実体が存在することに相当すると考えられる．

一方，この各ユーザを指し示す表現である,「ホリエモン」や「堀江貴文」という文字列は，この実体に対する言及であると捉えることができる．

つまり，ウェブ上の言及表現（テキスト）とその実体（アカウント）をひもづける問題は，シンボルグラウンディング問題の一種と考えることができる．本研究では，ウェブという文脈の中で，特に，固有表現に特化し，言及としてのシンボルと，それが指す内容の関係性について扱う．

3 本研究で解くべき問題

人名として言及されているシンボルを，どのユーザアカウントと紐付ける（グラウンドする）べきか，というのが本研究のタスクである．本研究で提案する手法を用いた応用として，種々なテキストとソーシャルメディアのアグリゲーションを行うアプリケーション

や，誹謗中傷のモニタリングなど，社会的必要性のあるものが考えられる．

まず，シンボル・グラウンディングの枠組みの上での本研究のアイデアとして，ある人名を表すシンボルと紐付けられるアカウントは，そのシンボルとして，社会的に認識されているべきであるといえる．例えば，そのアカウントとしてのメールアドレスが，往々にして使われているのならば，周囲はそのアカウントを認知しており，何という人（シンボル）のものであるかわかる（グラウンドする）であろう．ここで，本人のアカウントであったとしても，認知されていないものは，グラウンドするに足らないと考えられる．

4 提案手法

本研究では，ひとつの実体とそれを指し示す表現の同定を行う．これを行うことができれば，次のようなメリットがあると考えられる．

- 複数のソーシャルメディアの情報をアグリゲートすることができる．例えば，人に関する情報を収集し，その人にお知らせするなど，さまざまなサービスの構築につながる．
- ウェブ上の言及からどのくらい実体が特定可能かということが明らかになれば，プライバシー上の問題の範囲が明確になる．
- 言及から実体を特定する方法の有効性を検証することで，シンボルグラウンディング問題に対する解決策の手がかりとすることができる．

ここで本研究の目的を，次のように定式化する．

あるシンボル x が，ユーザアカウント y についての言及であるかどうかの判定を行う．アプローチとしては，紐付けられる事が，社会的認知となっているかどうかという事である．

前節までをふまえて，提案するアルゴリズムは以下のようになる．

1. シンボル x と，それに対応するユーザアカウントの候補 y が与えられる．
2. ユーザ y が，ソーシャルメディア上でシンボル x を ID として名乗っているか．(問題 1)
3. シンボル x とユーザアカウント y がどの程度，社会的に認知されているか．(問題 2)
4. 認知度が十分に高ければ紐付ける．

4.1 実験 問題 1 名称の一致

有名人の Twitter アカウント 500 個を用い，特徴量として，名前の編集距離と，名前をふりがな，ローマ字化した編集距離を用い，負例として，正例の数と同じ数になるように，ランダムな組み合わせを生成して用い，分類器には Support Vector Machine を使用した．

4.1.1 結果

素性ごとの効果を確認するために，評価実験を，素性 1:表現の類似度，素性 2:ふりがなの類似度，素性 3:ローマ字の類似度，そして素性すべてを用いた場合で行なった．

交差検定を行なった結果，素性 1 だけを用いた場合:89.8%，素性 2 の場合:89.9%，素性 3 の場合:91.0%，すべてを用いた場合は，91.6%となった．

4.2 実験 問題 2 社会的認知度

特徴量として，素性 1:つぶやきの数，素性 2:フォロワーの数，素性 3:登録されているリストの数を用い，認知度のないユーザとして有名人以外のアカウント 500 個を無作為に選んだ．

4.2.1 結果

交差検定を行なった結果，素性 1 だけを用いた場合:89.6%，素性 2 の場合:95.6%，素性 3 の場合:97.8%となった．

5 考察

実験 1 の素性としては，ローマ字化したものを用いるのが一番効果が大きかったが，ほとんどは素性 1 の表現だけで類推可能，つまり，今回用いた著名人のデータセットは多くが，そのまま名前を用いているといえる．

全体としても，著名人については，類推しやすい名前を付けている事が多いであろうと予想でき，これは本研究の手法による可用性をもたらすと考えられる．

また，有名人のアカウントはフォロワーやリストに登録されやすく，これを元に認知されているかどうか，みなせる可能性が高いといえる．

今後は，ウェブ上での言及と Twitter 上での活動の文脈が一致しているかどうか，についての調査が必要である．

参考文献

[1] S HARNAD. The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3):335-346, June 1990.