1T-02

LIWCを用いたユーザ推定手法の検討とSNSデータへの応用

冨平 準喜† 山下 晃弘† 松林勝志†
東京工業高等専門学校 情報工学科†

1. まえがき

近年，TwitterやFacebook，Instagramをはじめとする Social Networking Service(以後 SNS)のユーザが爆発的に伸びている．Twitter社の2017年第三四半期の報告によると世界での約3億3000万人のユーザがアクティブである［1]とされ，ユーザ公開型のSNSで世界最大である．こういったSNS上でのマーケティングが大いに期待され，SNSデータを用いたアカウントの持ち主の年齢や性別，職業といった属性を推定する研究が盛んに行われている．

関連研究として，Schwartzらは Facebookの投稿に含まれる単語とトピックの使用頻度から辞書を作成し，性格，性別，年齢を判別できることを示した［2]．また日本語を対象としたものでは，IBMの那須川らは語彙を抽象化してカテゴリ化するためのツール Linguistic Inquiry and Word Count(以後 LIWC) [3]を独自で日本語に翻訳し，それを用いてTwitterのツイートからユーザの性格を推定できることを示している［4]．

本稿では，これらの先行研究を参考に，英語で作成されたオリジナルのLIWCを独自の方法で半自動的に日本語に翻訳し，得られた特徴量からTwitterユーザの職業推定を行った．

2. LIWCの日本語への翻訳

LIWCとは語彙を抽象化してカテゴリ化するツールであり，選挙運動の分析や著者の年齢推定、性格推定などの様々なタスクの素性として使用されている．LIWCはそれぞれのカテゴリごとの単語を内包した辞書を持ち，カテゴリ内の単語の現れる回数から特徴量を算出する．例えば，“Sad”のカテゴリであれば”cry”や”alone”，“lost”などが定義されている．

翻訳の対象はLIWC2015の全73カテゴリのうち，日本語には存在しないカテゴリを除く66カテゴリとする．基本的にword2vec［5]を用いて単語の分散表現を算出することで，翻訳結果がカテゴリに属しているかを後述するカテゴリの代表語の分散表現とのcos類似度の値から判別する．

2.1 カテゴリの代表語の決定

翻訳結果がカテゴリ内に属しているかを判断するために，元の辞書のカテゴリの中で最も平均的な単語を定め，カテゴリ代表語とする．

カテゴリcat_iの平均ベクトルv_{cat_i}を(1)式より求める．

$$v_{cat_i} = \frac{\sum_{j=1}^{n_{cat_i}} v_{w_j}}{n_{cat_i}} \quad (1)$$

ここでn_{cat_i}は元の辞書に含まれるカテゴリcat_iの持つ単語数，v_{w_i}は単語w_jの分散表現である．

カテゴリcat_iの平均ベクトルv_{cat_i}と単語w_jとのcos類似度を(2)式より求める．

$$cos\left(v_{cat_i}, v_{w_j}\right) = \frac{v_{cat_i} \cdot v_{w_j}}{\|v_{cat_i}\| \cdot \|v_{w_j}\|} \quad (2)$$

cos類似度が最も大きかった単語をカテゴリ代表語と定める．これを全てのカテゴリに対して行う．

2.2 翻訳とカテゴリとの類似度の算出

使用するカテゴリ群の全ての単語に機械翻訳を行う．単語w_iを日本語に翻訳したものを単語t_i，単語t_iを英語に再翻訳した単語をr_iとすると，分散表現v_{rev_i}の平均とv_{rep_i}のcos類似度を求めることで，元のカテゴリとの類似度を算出することができる．

分散表現v_{rev_i}の平均を(3)式より求める

$$v_{rev_i} = \frac{\sum_{j=1}^{n_{rev_i}} v_{r_j}}{n_{rev_i}} \quad (3)$$

ここでn_{rev_i}は日本語から英語に再翻訳を行った後のr_iの単語数，v_{r_j}は単語r_jの分散表現である．また単語t_jと元のカテゴリの類似度を(4)式より求める．

$$cos\left(v_{rep_i}, v_{rev_j}\right) = \frac{v_{rep_i} \cdot v_{rev_j}}{\|v_{rep_i}\| \cdot \|v_{rev_j}\|} \quad (4)$$

2.3 辞書の追加

2.2でカテゴリと単語のcos類似値を求めたが，カテゴリごとにcos類似値が高い順に並べて，元のカテゴリと同じ数だけ取り出して辞書に追加する．

Estimation of user attributes using LIWC and application to SNS
†Department of Computer Science, National Institute of Technology ,Tokyo College

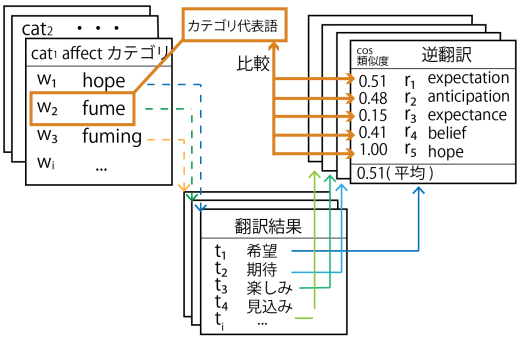

図 1 カテゴリと単語の cos 類似値の算出例

3. 評価実験

日本語に翻訳された辞書の評価のため，カテゴリ毎の単語の出現数を特徴量として，ユーザのツイート内容を機械学習し職業推定を行う．

3.1 ユーザデータの収集

VALU という株式会社 VALU によって運営されているフィンテックサービス [6]では，登録されたユーザのプロフィールに職業情報と SNS のアカウントが登録されている．VALU から twitter のアカウントが登録されているユーザのプロフィール情報を収集した．143 職種ある中から，今回はユーザ数が多かった表 1 の職業で推定を行う．またユーザ 1 人につき，最大で 3000 ツイートの収集した．

表 1 使用する職業のユーザとツイート数

職業	ユーザー数	ツイート数
美容師	153	247406
カメラマン	321	547961
アーティスト	388	715176
学生	875	1218769
ミュージシャン	370	742767
編集者	232	433122
ソフトウェアエンジニア	311	648461

3.2 機械学習の手法

ユーザ 1 人のツイートに対して，辞書内 66 カテゴリに含まれる単語の出現数より 66 次元の特徴量を入力として機械学習を行う．使用したアルゴリズムは RBF カーネルを用いた Support Vector Machine(SVM)，Random Forest，Logistic Regulation の 3 つである．3-fold cross-validation の検証方法で各アルゴリズムのパラメータを最適化した．

3.5 結果

各ユーザがその職業に属するか否かの二値分類を職業別に行なった(図 2)．職業別に分類精度の偏りはあるが，どのアルゴリズムにおいても平均して 75%程度の分類精度となった．また美容師は精度の開きが大きいが，ソフトウェアエンジニアは開きが小さいのはユーザ数の違いによる．なお，8 種類の職業を 8 カテゴリとして分類した場合の精度は約 62%であった．一定の分類精度が得られたことから，日本語に翻訳された LIWC の辞書は多少ノイズを残しつつも，属性推定に有効であることが確かであると言える．

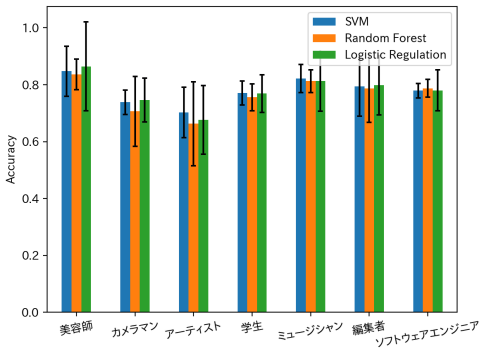

図 2 各アルゴリズムでの職業分類の精度

4. まとめ

本研究では LIWC の辞書を日本語に翻訳することで，テキストからの属性推定の有効性が確認できた．しかし分類精度は平均 75%程度で高いとは言えないため，翻訳された 66 カテゴリのうち，職業別の単語出現数に差がなかったカテゴリにおいて単語の追加と削除を行うことでノイズを減らし，分類精度の向上を期待できる．今後は日本語に合ったカテゴリをオリジナルで追加し，分類精度を高め，職業推定以外の属性推定の課題に取り組む予定である．

5. 謝辞

LIWC の辞書を提供して頂いたテキサス大学の J. Pennebaker 氏，R. Boyd 氏，Pennebaker Conglomerates, Inc. また，名古屋大学の笹原氏に感謝の意を表します．本研究は JSPS 科研費 JP15K16092 の助成を受けたものです．

引用文献

[1] “Investor Relations,” https://investor.twitterinc.com/.
[2] H. A. Schwartz, J. C. Eichstaedt , M. L. Kern, “Personality, gen- der, and age in the language of social media,” The open- vocabulary approach, 2013.
[3] “LIWC,” https://liwc.wpengine.com/.
[4] 眞. 山本 , 哲. 那須川, “LIWC2001 手作業翻訳の方針と 半自動翻訳手法の提案,” 言語処理学会 第 22 回年次大会, 2016.
[5] T. Mikolov, K. Chen, G. Corrado , J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv, 2013.
[6] “VALU,” https://valu.is/.