G-111

エゴグラムを用いた性格と SNS 上で使用する言葉の関係性の調査

Relationship between Character using Egogram and Words on SNS

桝本 智絵†　山岡 拓生††　大井翔‡　佐野睦夫†
Chie Masumoto　Hiroki Yamaoka　Sho Ooi　Mutsuo Sano

1. はじめに

エージェントに性格を付与することで様々な対話シミュレーションなどに応用ができる。例えば，エージェントを生成したときに，一からその性格に合った会話を生成する必要がなくなるだけでなく，多様な対話が可能となる．

エージェントの性格形成の自動化のために，①性格推定システムの作成，②性格の自動分類による学習データ自動取得，③学習データから対話生成の手順が考えられる．

本研究は性格推定の前段階であるその人が使用する単語とエゴグラム診断による性格の関係性を調査する．

初期段階として，Twitter などのマイクロブログはそのユーザの趣味嗜好や行動傾向を探ることができると考え，ユーザの言語情報から性格を推定する．具体的には，Twitter ユーザのツイートのテキストデータを取得し，エゴグラム診断により性格の数値化を行い，グルーピングした結果に基づき解析を行う．そして，Twitter で用いる言葉と性格の関係性を調査する．

2. 関連研究

岡本らの研究[1]では，Twitter から得られるテキストデータからテキストマイニングすることで性格推定している．Twitter ユーザが投稿したテキストデータを形態素解析し，ナイーブベイズ法を用いて分類を行う．性格はエゴグラムに基づく性格を採用しており，人の性格を 243 に分けている．エゴグラム診断を行い，その結果を伴ってツイートしたアカウントを対象に実験を行っている．結果として平均で 54%の推定精度となっている．

山岡らの研究[2]では，マーケティングにおけるペルソナの詳細な項目を推定するため Twitter のテキストデータを解析し，ユーザの趣味を推定している．推定にはナイーブベイズ法を用いて分類器を作成している．結果として平均で 93%の正解率であった．

上記の研究では，性格を細かくわけているためそれぞれの性格のデータが少なく過学習がみられる．そこで，本研究では性格の 5 つの要素それぞれの高低で性格を分けることでデータ量を増やした．このとき，Twitter ユーザが使用する言葉と性格に関連性があるかを調べる．

3. 性格と言葉との関係抽出

3.1. エゴグラム

エゴグラム[3]とはエリック・バーンの交流分析に基づいてジョン・M・デュセイが考案した性格診断法の一つで，人の心を 5 つに分類して分析するものである．以下は 5 つの自我状態である．

・CP（支配性）
・NP（寛容性）
・A（論理性）
・FC（奔放性）
・AC（順応性）

エゴグラム診断ではこれらの状態をそれぞれ点数化し、そのバランスから性格を推定する．本研究では，各自我状態を高いグループと低いグループの 2 段階で取り扱うものとする．

本研究では，エゴグラムをもとに作成している性格診断のサイトを利用している人を対象とした．

3.2. 形態素解析

収集したデータを Janome[4]を用いて形容詞のみを抽出する．Janome とは辞書内包の形態素解析器である．形容詞を抽出したらその単語の頻出回数を調べる．各自我状態の高い人と低い人のグループがそれぞれどのような形容詞をよく使用しているかを調べるために，単純にツイートに出てくる形容詞をカウントした．

3.3. 性格と言語の関係性

性格診断の結果を Twitter に投稿している人のツイートを収集し，自我状態の状態が高い人と低い人分け，使用単語を形態素解析により取り出し，頻出度を出す．ここでは使用単語は形容詞のみに着目する．その結果から自我状態に高低で使う単語に差がないかを調べる．本来性格はそれぞれの自我状態のバランスから決まるものであるが，本研究ではそれぞれの自我状態を別々にして考える．

4. 実験

エゴグラムに基づいて作成されている性格診断の結果を Twitter に投稿しているユーザ 100 人の最新ツイート 200 件に対して Twitter API を用いて取得する．また，この中からリツイートの投稿を除いた．リツイートとは他人のツイートを自分のタイムラインに流す機能である．除外した理由として，他人のツイートそのままであるため自分の表現ではないためである．5 つの自我状態のそれぞれについて高い人と低い人の 2 つに分けた結果を表 1 に示す．

本研究では，形容詞の頻出を調べる際，それぞれの自我状態の高い人と低い人の人数を合わせるため，多い人数をランダムで選択し，少ない人数と同じ人数に合わせる．例えば，CP は高い人の人数の方が少ないため低い人 62 人からランダムで 38 人を抽出し，高い人と低い人がともに 38 人にしてからデータを取得する．

†大阪工業大学，Osaka Institute of Technology
††大阪工業大学院，Graduate School of Osaka Institute of Technology
‡ 立命館大学，Ritsumeikan University Technology

表 1　それぞれの自我状態の人数

	高い人	低い人
CP（支配性）	38	62
NP（寛容性）	77	23
A（論理性）	58	42
FC（奔放性）	67	33
AC（順応性）	43	57

5. 結果

各自我状態の高い人と低い人のグループがそれぞれどのような形容詞をよく使用しているかを調べるために，単純にツイートに出てくる形容詞をカウントしていった．NP と FC，AC の結果を図 1～3 に示す．

結果より，自我状態が高い人と低い人がそれぞれどのような単語をよく使用するかをまとめた．基準としては出現単語の回数の比が 1.5 を超え，かつ出現回数が 5 を超えるものとする．例えば FC の「悪い」という単語は FC の高い人が 19 回，低い人が 45 回使用していた．この時高い人に対する低い人の比は 1.5 を超えるため FC の低い人が「悪い」という単語をよく使用すると判定する．結果を表 2 に示す．CP は条件を満たす単語がなかった．

6. 考察

CP は CP が高い人と低い人の使用形容詞に明瞭な差がないことがわかる．そのため形容詞以外の品詞や別の特徴を使う必要がある．反対に A は A の高い人と低い人で差がでる単語が計 12 個あった．それぞれの自我状態によって差が出やすいものとそうでないものがあることが分かる．NP では NP の低い人は「怖い」，「悪い」などマイナスな単語を使う傾向にある．しかし NP の高い人は「うれしい」，「高い」，「やすい」，「懐かしい」，「新しい」，「っぽい」という単語をよく使う傾向があることが分かったが，これらの単語に共通点はみられない．他も同じように A，FC，AC の高い人と低い人で使用単語に差がでたもののその単語群に共通点は見られなかった．

今回使われる言葉に顕著な差が出なかったのは，使われている言葉を単純にカウントしたからだと考えられる．使われている言葉を単純にカウントするだけの場合，すべてのデータの中で同じ人のデータが占める割合が大きくなるため顕著に結果が現れなくなる．例えば，ある人が性格とは関係ない言葉を癖などで多く使用しているとそのグループ全体の使用単語に影響する．これを改善するために言葉に重みを付けることが挙げられる．同じ人が多く使う言葉はその言葉の重みを小さくすることでこれを改善する．

7. まとめ

今回 Twitter に投稿している言葉と性格の関係性を明らかにした．

今後は，ツイートの文章だけを使うのではなく，行動傾向例えば写真を載せているか，誰かのツイートに返信をしているか，またツイートの頻度などを考慮すればよいのではないかと考えている．さらに単語をカテゴリ化して分類することや，単純に単語だけを見るのではなく文章の構造解析もしていきたい．

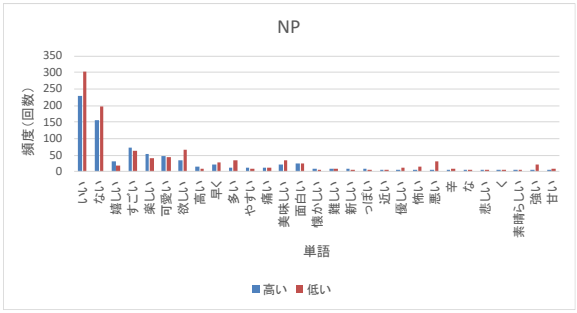

図 1　NP 内の形容詞の頻度

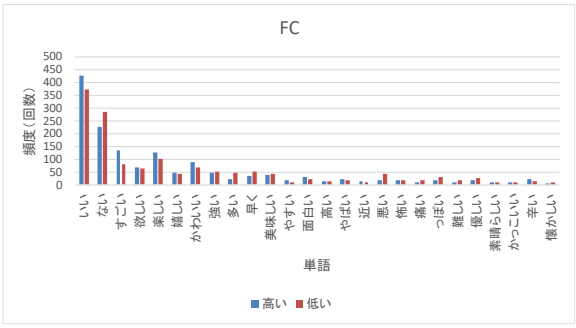

図 2　FC 内の形容詞の頻度

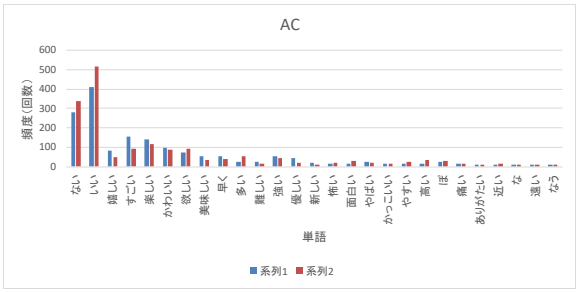

図 3　AC 内の形容詞の頻度

表 2　出現回数の多い形容詞

	高い	低い
CP	―――	―――
NP	うれしい，高い， やすい，懐かしい 新しい，っぽい	多い，怖い 悪い，強い 甘い
A	多い，美味しい 強い，やすい 素晴らしい，近い ええ	かわいい，高い やばい，怖い っぽい
FC	すごい，やすい 難しい，辛い	多い，早く 悪い，痛い っぽい，優しい 懐かしい
AC	うれしい，すごい 優しい，新しい	多い，面白い やすい，高い

参考文献

[1] 岡本 拓馬，松本 和幸， 吉田稔， 北研二，ナイーブベイズ法を用いた Twitter による性格推定，言語処理学会第 20 回年次大会予稿集, pp.1123-1125, 2014.
[2] 山岡拓生，佐野睦夫，ナイーブベイズ法に基づく SNS を利用したペルソナ推定，人工知能学会第 32 回全国大会，2018
[3] ジョン・M. デュセイ, エゴグラム―ひと目でわかる性格 の自己診断, 創元社，2000
[4] Janome，https://mocobeta.github.io/janome/,(アクセス日 2019/7/20)