

Twitterにおける会話量による言葉遣いの 文体的特徴変化の分析手法の提案

江口明利 武藤敦子 森山甲一 横越梓 吉田江依子 犬塚信博
名古屋工業大学

1 はじめに

社会言語学の分野では、アコモデーション理論 [1] に基づき、人は会話における言葉遣いを変化させることで他者との社会的距離を操作したり、コミュニケーションの効率を高めることが示されてきた。また、大規模な人間の行動観察を可能とする Twitter 等の SNS を用いた研究が社会言語学においてトレンドとなっており、アコモデーション理論に基づく研究も行われている。しかしこれらの研究は、対話相手の属性に着目したものが多く、個人間の関係性そのものに着目したものは少ない。個人間の関係性を測る指標として、社会心理学の単純接触効果 [2] がある。これは短期間で接触した回数が多い対象ほど、印象が向上するというものであり、Twitter においては、リプライ数、つまり会話量を接触とみなすことができる。本研究では、Twitter における会話量の増加に伴う言葉遣いの変化の文体的特徴の分析手法を提案し、言語使用の個人間の関係性における変化を調査する。

2 関連研究

林ら [3] は、リプライツイートデータからネットワークを構築し、ユーザをコミュニティに分割することで、リプライをコミュニティの内と外のユーザへ向けたものの2種類のツイート群に分類した。それらと既存の日本語辞書から抽出した複数の特徴量を用いて作成した特徴量行列に、非負値行列因子分解 (Non-Negative Matrix Factorization, NMF) [4] を用いて、文体的特徴の分析を行うことで、コミュニティの内外に向けた言葉遣いの変化を発見した。この手法では (1) 内ツイートと外ツイートの分類, (2) テキスト特徴量の獲得と特徴量行列の作成, (3) NMF による特徴量行列の因子分解, (4) ツイート群の特徴を表す基底の選択, (5) 基底の評価による重要特徴量の抽出というフローを構成した。

A Proposal for a Method to Analyze Changes in Stylistic Features Caused by the Amount of Conversation on Twitter

Akito Eguchi, Atsuko Mutoh, Koichi Moriyama, Azusa Yokogoshi, Eiko Yoshida, and Nobuhiro Inuzuka
Nagoya Institute of Technology

3 提案手法

本研究は、双方向的なコミュニケーションと考えられるリプライツイートデータを会話量ごとにグループに分け、特徴量行列を作成し、NMF を用いて会話量の増加に伴うテキスト特徴量の変化を定量的に評価することを目的とする。提案手法は、林らの手法 [3] のフロー (1),(4) を変更し、(1) リプライツイートのグループ化, (4) 変化した特徴を示す基底の選択、とし、それぞれ 3.1 と 3.2 で説明する。

3.1 リプライツイートのグループ化

相互にリプライのやり取りがあるユーザ a, b において、そのリプライツイート群を「会話」とする。この時、どちらのユーザからもリプライが1度ずつ発生した場合、そのリプライを1回のコミュニケーションとし、コミュニケーションではないリプライ数を「偏り」とする。

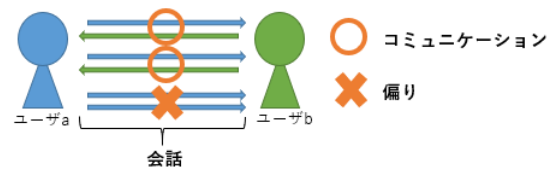


図 1: ユーザ同士のやり取りの例

ユーザ a, b の会話に式 (1) でコミュニケーションと定めたリプライ数に対する偏りの比率 T_{ab} を与える。

$$T_{ab} = \frac{|R_a - R_b|}{S_{ab} - |R_a - R_b|} \quad (1)$$

式 (1) は、a から b へのリプライ数を R_a 、逆を R_b とし、会話量を $S_{ab}(R_a + R_b)$ としている。収集した会話からコミュニケーションが多いものを選択していくため、 T が閾値以上のものは分析対象から外し、また単純接触効果は回数が増えすぎると効果が少なくなること考を慮し、 S も閾値以上のものは分析対象から外す。残った会話にグループ番号 I を与え、 I が同一の会話の集合を会話群 I とする。 I は $S \div V$ (V は任意の定数) の商によって得られる。[3] の手法に基づき、これ

らの会話群とテキスト特徴量から、サイズ $N \times K$ (会話群の数が $i = 1, \dots, N$, 特徴量の数が $j = 1, \dots, K$) の特徴量行列 Y を作成, NMF により $Y \approx HU$ に因子分解する. 基底数 M はエルボー法により定め, 基底行列 H の要素を $h_{i,m}$ とし, 因子行列 U の要素を $u_{m,j}$ とする.

3.2 変化した特徴を示す基底の選択

基底行列 H のそれぞれの基底に対して, 会話群 i と会話群 $i+1$ の成分の差をとる. 全ての会話群に対して行い, 会話量の増加による特徴量への寄与率の変化量を示す基底行列 $(N-1) \times M$ を作成する. この変化量の平均値が最も高い基底に対応する係数行列 U の基底が, 会話量の上昇に伴い普遍的に増加した特徴量を示すと解釈できる. 逆に平均値が最も低い基底は, 普遍的に減少した特徴量を示すと解釈できる. また基底ごとのスケールの違いを考慮し, 成分 $h_{i,m}$ に対応する基底 m の因子行列 U における総和を重みとして掛けることで, 値の補正を行う. 基底行列 H において, 寄与率が最も増加した基底 m_{up} は式 (2) で算出される.

$$m_{up} = \arg \max_m \frac{\sum_i (h_{i,m} - h_{i+1,m})}{N-1} * \sum_j u_{m,j} \quad (2)$$

対して, この平均値が最も小さいときの基底 m_{down} は, 寄与率が最も減少した基底である. その後 [3] の手法に基づき, 基底 m_{up} と m_{down} に対応する因子行列 U の寄与率の差を, 各テキスト特徴量ごとに比較することで, 増加・減少したテキスト特徴量を評価する.

4 実験と考察

4.1 実験環境

実験で用いるツイートデータは Twitter 社が公開する WebAPI を通じて収集した. 対象としたリプライツイートは日本時間で 2021/5/1 から同年 6/30 に渡って発信されたもので, 約 400 万件を収集した. NMF はライブラリ scikit-learn を用いて実装を行った. また, テキストの解析には形態素解析エンジン MeCab を用いた. 辞書は, IPADic と mecab-ipadic-NEologd を併用した. 収集したデータに対して, 提案手法を用いた分析を行った. T の閾値は 0.5, S の閾値は 200 に設定し, $V = 10$ に設定して 20 個の会話群にグループ分けした. 分析に用いたツイートは各会話群のツイートから 5,000 件をランダムにサンプリングした. 対象のツイートからテキスト特徴量を獲得したところ 6,196 種類が確認された. したがって, 特徴量行列の次元は $20 \times 6,196$ である. NMF の基底数 M は, $M = 20$ とした.

4.2 結果と考察

増加・減少した特徴量上位 8 つを, 表 1 に示す.

表 1: 重要特徴量上位 8

	増加した特徴量	減少した特徴量
1	機能表現 (稀) 使用が一般的な表現	絵文字の使用頻度
2	機能表現 (レベル 6) とりたて詞の挿入なし	感情極性対応表 実数値の和
3	語種 和語の割合	MVR 動詞の使用率
4	機能表現 (レベル 8) です/ますを含まない語	名詞の使用率
5	感情極性対応表 ネガティブ単語の使用割合	文末表現 「ます」
6	機能表現 (核) 核の語の位置が先頭から 1 番目	文末表現 「ございます」
7	機能表現 (レベル 5) 音韻的变化がない	顔文字の使用頻度
8	機能表現 (難易度) A1 (最も易しい)	文末表現 「です」

5,6,8 番目に減少した特徴量の丁寧語の文末表現と, 4 番目に増加した特徴量の機能表現の「です/ますを含まない語」から, 会話量の増加に伴い丁寧表現がなくなっていくことがわかる. また 1,7,8 番目に増加した特徴量の機能表現と, 1,7 番目に減少した特徴量の「絵文字, 顔文字の使用頻度」から, 会話量の増加に伴い伝わりやすい端的な文章で会話を行うことがわかる.

5 まとめ

本研究では, 言葉遣いの個人間の関係性による変化を調査するため, Twitter を対象として会話量の増加に伴う文体的特徴の変化を観察した. 結果, 単純接触効果に基づく対象への印象の向上によって, 丁寧表現が減少し, 端的な文章になっていくことが分かった. これは, 個人間の関係性の変化に伴いアコモデーション理論における社会的距離の操作と, コミュニケーションの効率化が行われたと結論付けられる.

参考文献

- [1] Giles, H., Taylor, D., Bourhis, R. (1977): Dimensions of Welsh Identity, European Journal of Social Psychology 7(2), pp. 165-174.
- [2] 川上 直秋 (2015): 単純接触効果と無意識へわれわれの好意はどこから来るのかへ, エモーション・スタディーズ, 第 1 巻, 第 1 号, pp. 81-86.
- [3] 林大知, 武藤敦子, 森山甲一, 横越梓, 犬塚信博 (2020): Twitter における言葉遣いのコミュニティ内外での違いに関する文体的特徴に基づく分析, 信学技報, vol.119, no.469, AI2019-62, pp. 49-54.
- [4] Lee, D. D., Seung, H. S. (2001): Algorithms for nonnegative matrix factorization. In Advances in neural information processing systems, pp. 556-562.