5V-06

Twitterデータにおける初対面会話からの経過に伴う言葉遣いの文体的特徴変化

高橋昂希　武藤敦子　島孔介　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，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，初対面会話におけるポライトネス表示に関する研究として，三牧は，初対面会話データ 15 分間を全て文字化し，待遇レベル管理によってポライトネスがどのように表示されているかについて考察していた [2].

一方で，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いる．江口らは，2 か月間の会話の頻度によって文体的特徴がどう変化したかを分析したが，初対面会話からの会話の追跡は行っていない [4]．本研究では，Twitter 上で出会ったとされるユーザ同士の会話データを抽出し，会話の経過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，潜在的な言語的特徴の変化を分析する．

2　関連研究

林ら [3] は，Twitter のリプライツイートデータからネットワークを構築し，リプライをコミュニティの内と外のユーザ向けの 2 種類に分類した．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特徴量行列を作成し，非負値行列因子分解（Non-Negative Matrix Factorization，NMF）[5] 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．江口ら [4] は，リプライツイートから双方向的なコミュニケーションを抽出し，2 か月間の会話の頻度で，段階的にツイート群のグループ化を行った．この手法では林らの手法を基に (1) 対象データの取得 (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) 特徴を示す基底の選択 (5) 選択された基底に対応する特徴量の寄与率の上位を重要特徴量として抽出というフローを構成した．

しかしながら，江口らの手法は一定期間の会話の頻度の違いに着目していることや，同じペアの会話を追跡していない点で，本研究においてそのまま用いることはできない．また，江口らの (4)「変化した特徴を示す基底の選択」では基底行列 H のそれぞれの基底に対して，式（1）により，寄与率が最も増加した基底 m_{up} を求める．但し，N は会話群の数とし，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．

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

この基底の選択方法では基底行列 H 全ての行を参照せず，1 行目と N 行目の差によって重要基底を決めている．

3　提案手法

本研究では，以下の手順で初対面からの会話の経過に伴うテキスト特徴量の変化を評価する．
(1) 初対面からのリプライデータの取得
(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) 変化した特徴を示す基底の選択
(5) 基底の評価による重要特徴量の抽出

提案手法は江口らのフローを基にしているが，江口らは会話の頻度によってグループ分けをしているのに対し，本研究では初対面からの会話の経過によってグループ分けをしているため (1)(2) が異なる．また (4) の変化した特徴を示す基底の選択方法にも課題があったため変更を加えている．(1)(2)(4) の内容について以下に述べる．

3.1　初対面からのリプライデータの取得

「@アカウント名　はじめまして」から始まる Twitter 上で初めて知り合ったとされる相互にリプライのやり取りがあるペアにおいて，各ペアのリプライを 2 か月間取得し，そのリプライ数が R を超える組を観察対象とする．

Changes in stylistic features of language from first-time conversations in Twitter data
Kouki Takahashi, Atsuko Mutoh, Koichi Moriyama, Azusa Yokogoshi, Eiko Yoshida, and Nobuhiro Inuzuka
Nagoya Institute of Technology

3.2 テキスト特徴量の獲得と特徴量行列の作成

3.1で取得したリプライツイートを各ユーザの1リツイート目を会話群1として，会話群Nまでグループ分けする．但しNは会話群の数を表し，偶数とする．江口らの手法[4]に基づき，これらの会話群とテキスト特徴量から，サイズ$N \times F$(会話群の数が$i = 1, ..., N$，特徴量の数が$j = 1, ..., F$)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Kはエルボー法により定める．

3.3 変化した特徴を示す基底の選択

基底行列Hの各基底の前半と後半の和の差を取る．差が最も大きな基底に対応する係数行列Uの行が，初対面からの経過に伴い増加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．寄与率が最も増加した基底m_{up}は式(2)で算出される．

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

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を通じて取得した．ツイートデータは2022年9月1日に「はじめまして」とツイートしたペアを2022年11月1日まで取得した．NMFはライブラリscikit-learnを用いて，テキスト解析には形態素解析エンジンMeCabを用いた．テキスト特徴量は2785種類が確認された．観察対象とするユーザの組はリプライ数が20以上のものとし（R=20），会話群の数はN=10，NMFの基底数はK=10とした．比較として手順（4）のみ江口らの手法を用いた実験も同時に行い，従来手法として示す．

4.2 実験結果と考察

増加・減少した特徴量の上位3個を表1に示す．提

表 1: 重要特徴量上位 3

	提案手法		従来手法	
	増加	減少	増加	減少
1	です/ますを含まない語	改行数	文体：口語	改行数
2	文体：常態	絵文字使用率	相の類 抽象的な関係を意味する語彙	文字数
3	とりたて詞の挿入のない語	文字数	難易度：A2	絵文字使用率

案手法にて1，2番目に増加した「です/ますを含まない語」「文体：常態」から会話を繰り返すことで丁寧語の使用が減っていくことがわかる．丁寧語を使わないことで社会的距離を縮めようとしていることが考えられる．「改行数」「文字数」「絵文字使用率」が減少していることから，会話の経過につれコミュニケーションの効率化が図られていると考えられる．一般的に絵文字は感情や語調を相手に誤解なく認知させたり印象付けたり配慮したりするために使用されるが，そのような配慮をしなくなるものと考えられる．

4.3 従来手法との比較

増加特徴量となった3つの特徴量行列の値を図1に示す．

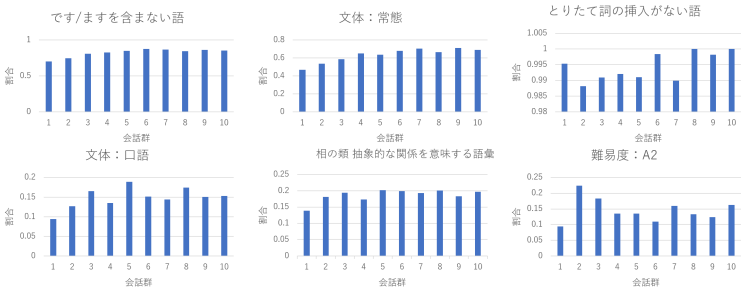

図 1: 会話群毎の特徴量行列の値

提案手法では前半と後半の差を取っているため全体的な増加の傾向をみ取れているが，従来手法では会話群1と10の差のみを取っているため，「難易度：A2」のように必ずしも増加傾向でない特徴量が上位に表れている．

5 まとめ

本研究では初対面からの会話の経過に伴って変化していく文体的特徴を分析する手法を提案しTwitterを対象として会話を観察した．実験の結果，初対面から会話回数が増えていくと丁寧表現が減少し，端的な文章になることが分かった．このようなSNS上の会話の経過は，アコモデーション理論やポライトネス理論に基づけば，相手との心理的・社会的距離を近づけるためのストラテジーが用いられていることを示している．

謝辞 本研究はJSPS科研費JP22K00528の助成を受けたものです．

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] 三牧陽子 (2002)：待遇レベル管理から見た日本語母語話者間のポライトネス表示，社会言語科学，第5巻，第1号，pp. 56-74.

[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] 江口明利，武藤敦子，森山甲一，横越　梓，吉田江依子，犬塚信博，“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”， 情報処理学会第84回全国大会，2022

[5] Lee, D. D., Seung, H. S. (2001)： Algorithms for nonnegative matrix factorization. In Advances in neural information processing systems，pp. 556-562.