逐次的親密度推定に向けた発話対の親疎判別法の提案

木下 泰輝[1]　楠 和馬[2]　蒲原 智也[3]　波多野 賢治[4]
[1,3,4] 同志社大学文化情報学部 〒610-0394 京都府京田辺市多々羅都谷 1-3
[2] 同志社大学大学院文化情報学研究科 〒610-0394 京都府京田辺市多々羅都谷 1-3
[1]bio0098@mail4.doshisha.ac.jp, [2,3,4]{kusu, kambara, hatano}@ilab.doshisha.ac.jp

1 はじめに

近年，エンタテインメントの分野で，非タスク指向型対話システムが注目を集めている．対話システムにおいては，ユーザに長期間利用したいと感じさせるデザインが求められている [1]．このようなデザインの実現には，対話システムとユーザの仲の良さ（親密度）を定義し，それに合わせて発話内容を変化させることが有効であるとされている [2]．

対話システムで親密度を扱うためには，ユーザの発話やシステムの発話候補文から，その親密度を推定する手法が必要となる．対話システムにおける対話の方法には，音声ベースと，テキストベースが存在する．このため，双方の対話システムに搭載するためには，テキストデータから親密度推定を行う手法が必要である．また，親密度推定の既存手法では，テキスト化された対話を事後的に推定するが，親密度の変化を捉えるためには，発話ごとに親密度を推定し直し，その値を更新する方法（逐次的親密度推定）が必要となる．

そこで，本稿では逐次的親密度推定の実現に向けて，一つの発話とそれに対する応答（発話対）の親疎判別を実現する手法を提案する．ここで，親疎判別とは，仲の良さを親密と疎遠の 2 値で判別することを意味するが，発話対の親密度推定に対する客観的評価が困難であることを考慮し，親密度推定の前段階として親疎判別が必要であるとしている．

2 関連研究

本節では，既存の親密度推定手法と，その問題点を説明する．次に，親密度と関係する言語要素を網羅的に選出するため，心理学における代表的なテキスト分析手法を述べる．最後に，本稿で利用する doc2vec とその基礎技術である word2vec について説明する．

2.1 テキストデータからの親密度推定手法

Matsumoto らは，感情表出の頻度と発話と応答の比率を説明変数として，テキストデータから親密度を求める手法が存在する [3] ．この手法では，感情のみを説明変数とした場合は，推定手法における予測値とアンケート調査より得られた値の相関が 0.86 と述べられている．

文献 [3] では，発話と応答の比率が説明変数として使用されている．しかし，対話システムがユーザとの対話において，発話と応答の比率を制御するためには，ユーザの意図に反して発話を行う処理や，ユーザの発話を無視する処理が必要となる．したがって，対話システムがこの変数を制御することは困難だと考えられる．また，上述した二つの推定手法は，テキスト化された対話を事後的に分析しているが，対話システムにおいては，ユーザと対話システムの対話を逐次的に分析する必要がある．

2.2 心理学におけるテキスト分析手法

心理学におけるテキスト分析手法で利用される言語要素は，ユーザ定義辞書，表層的特徴，トピック分析の三つに分類されている [4]．

ユーザ定義辞書とは，心理学的な観点から語彙を分類した辞書を利用して分析を行う手法である．日本語で利用可能な辞書には，日本語感情表現辞書（JIWC）[5] がある．JIWC は，「驚き」「怒り」などの 7 カテゴリに，857 個の単語を分類したものである．また，JIWC に類似した辞書として，感情表現辞典 [6] がある．感情表現辞典は，2,200 語の単語を 10 感情に分類した辞書である．

表層的特徴とは，特定の基準に基づいて文を分割し，分析を行う手法である．ここで使用される特徴量には，形態素，文字 n-gram，品詞情報などが存在する．

トピック分析とは，単語同士の関係性を踏まえた分析手法であり，Latent Semantic Analysis (LSA) と，Latent Dirichlet Allocation (LDA) を用いて行われている．

2.3 文書の特徴ベクトル生成手法

文書の特徴ベクトルを生成する手法として doc2vec[7] が挙げられる．これまで特徴ベクトルの生成手法として広く利用されて来た Bag of Words (BoW) には， 1) 語彙の順序が失われる 2) 語の意味が無視される，という二つの問題点が存在する．しかし，doc2vec では，word2vec[8] と呼ばれる手法を利用する

表 1　各コーパスにおける対話数

	友人	初対面	合計
日本語話し言葉	103	139	242
名大会話	13	5	18
合計	116	144	260

ことで，これらの問題点を解決している．

word2vec とは，テキストデータにおいて，単語の周辺語の出現確率を学習することで，単語のベクトル表現を出力する手法である．doc2vec では，入力に単語 ID を付加することで，文書の特徴ベクトルを得ることが可能である．また，文単位から文章まで，幅広い文書長に対応することが可能である．また，doc2vec から出力された特徴ベクトル同士のコサイン類似度を算出することで，語の意味を考慮した文書同士の類似度計算が可能である．

3　予備分析

本稿では，親疎判別に必要な言語要素を明らかにし，その後，それらの言語要素を使用した発話対の判別手法を提案する．そのため，心理的現象と関連を持つ多様な言語的要素に関して，どの要素が親疎と関連があるのかを分析する必要がある．このために，話者同士の関係性が異なるコーパスに対する予備分析を行う．

3.1　使用コーパス

本稿では，親疎両方の対話を含むコーパスが必要となるため，話者同士の関係として初対面と友人の双方が含まれる二つのコーパスを利用する．

一つ目は，日本語話し言葉コーパス（トランスクリプト・音声）2011 年版 [9] である．このコーパスは，言語学研究を目的として多様な場面での自然対話を 66 時間分収録し，テキストデータに書き起こしたものである．

二つ目は名大会話コーパス [10] である．このコーパスは，129 会話分の日本語母語話者同士の雑談を文字化したコーパスである．

表 1 に，これら二つのコーパスにおける対話数を示す．

ただし，初対面同士の対話と友人同士の対話の数が異なるため，分析を行う際は，友人同士の対話より 116 件を無作為抽出し，対話数を揃える．

3.2　言語要素の数値化

本分析では，2.2 節で述べた言語要素を全て数値化する．まず，言語要素のうち，n-gram，単語，品詞情報については BoW を構築する．ただし，文書長の影響を排除するため，BoW の要素は，出現割合とする．また，n-gram に関しては，n の値を 1 ずつ増加させ，判別性能が極大となる値を採用する．また，言語要素のうち品詞情報，単語に関しては，JUMAN++ [*1]による形態素解析を行い，その解析結果を利用する．

一方，単語共起については，形態素を入力とした時の LDA, LSA の出力を利用する．また，双方について，トピック数を 1 から順に 1 ずつ増加させ，判別性能が極大となる値を採用する．

最後に，JIWC については，まず，コーパスと辞書の双方に対して JUMAN++ による形態素解析を行った上で，辞書の各カテゴリに属する単語の出現頻度を集計し，全単語数で割ることで割合を算出する．

3.3　分析手法

3.2 節で述べた手法によって特徴量を算出した場合，分析対象となるデータ数と比較して特徴量の数が多くなる．したがって，過学習に対処する手法が必要となる．また，説明変数間に相関が存在すると考えられることから，これに対処する手法が必要となる．

そこで，本稿では，Elastic Net ロジスティック回帰分析を用いる．この手法は，過学習を防ぎながら変数選択を行うことが可能であり，かつ，説明変数間の相関に対処することが可能である [11]．目的変数を初対面同士の対話とその他の対話の 2 値とし，説明変数を 2.2 節で述べた言語要素全てとして行い，変数選択を行う．なお，判別性能の評価は，10 分割交差検証法により適合率と再現率の調和平均（F 尺度）を算出することにより行う．なお，Elastic Net による変数選択を適切に行うため，全ての変数を平均 0，標準偏差 1 に標準化する必要がある．

3.4　分析結果と考察

3.2 節で述べた分析を行った結果，3,207 個の変数が選択され，F 尺度の平均値は 0.98 だった．また，モデルにおいて影響力の強い変数は，「だよ」「です」などの敬語の有無に関係する形態素だった．

以上の結果から，F 尺度の平均値が 1 に近いため，網羅的に言語要素を選出できたと考えられる．また，親疎の判別は主に敬語の有無によって行われていると推察できた．

4　提案手法

本稿の目的である発話対の親疎判別では，親疎判別に必要な言語要素の多くが一度も出現しない．そのため，判別モデルにおいて影響力を持つ説明変数のうちほとんどが 0 となり，判別性能が低下する問題が生じる．

この問題を解決するため，発話対を格納したデータ（QADB）より，親疎が類似している発話対（類似発話対）を取得し，特徴ベクトルを算出する手法の提案を行う．これにより，0 でない説明変数を増加させること

[*1] 日本語形態素解析システム JUMAN++：http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN++ （2018/1/16 閲覧）

が可能となる．本稿では，類似度計算に利用する手法に応じて二種類の手法を提案する．提案手法の処理は，QADB 内の発話対との類似度を算出，類似発話対を利用した特徴ベクトルの算出，という順に行う．本節では，これらの処理を順に説明する．

4.1 QADB 内の発話対との類似度算出

本節では，QADB より，親疎判別の対象となる発話対と親疎が類似した発話対を取得する方法を述べる．本稿では，類似度の算出に二種類の手法を利用する．

共起による類似度算出： 一つ目は，QADB に対する類似度計算として一般的によく利用される文書ベクトル同士の内積である [12]．親疎判別の対象となる発話対のうち発話文の文書ベクトルを $\boldsymbol{x^Q}$ とし，QADB 内の発話対 i における発話文の文書ベクトルを $\boldsymbol{y_i^Q}$ とする．この文書ベクトルは，3.4 節で選択された 3,207 個の変数の値を要素に持つ．この時，類似度 s_i^Q は，式 (1) で算出される．

$$s_i^Q = (\boldsymbol{x^Q})^t(\boldsymbol{y_i^Q}) \tag{1}$$

また，親疎判別の対象となる発話対のうち応答文の文書ベクトルを $\boldsymbol{x^R}$ とし，QADB 内の発話対 i における応答文の文書ベクトルを $\boldsymbol{y_i^R}$ とすると類似度 s_i^R は，式 (2) で算出される．

$$s_i^R = (\boldsymbol{x^R})^t(\boldsymbol{y_i^R}) \tag{2}$$

doc2vec による類似度算出： 二つ目は，2.3 節で述べた doc2vec を利用して類似度を算出する手法である．本稿において文書ベクトルは，3.4 節で選択された 3,207 個の変数のみを要素として持つことから，共起する場合が少ない．そのため，共起頻度を利用した類似度計算が困難な可能性がある．そこで，doc2vec を利用し，意味を考慮した類似度を算出する．具体的には，類似度 s_i^Q は，$\boldsymbol{x^Q}$，$\boldsymbol{y_i^Q}$ を doc2vec で特徴ベクトルに変換し，コサイン類似度を算出したものとする．ただし，doc2vec の学習には，QADB を利用する．また，s_i^R は同様の処理を $\boldsymbol{x^R}$ と $\boldsymbol{y_i^R}$ を利用して行ったものとする．

4.2 類似発話対を利用した特徴ベクトルの算出

一般に，対話において，応答は発話の親疎を反映していると考えられる．例えば，親密度の高い発話に対しては，親密度の高い応答が行われる場合が多いと考えられる．したがって，4.1 節において算出した類似度 s_i^Q が高い発話対を QADB 内から抽出し，その応答文における言語要素の出現傾向を変数化することで，0 でない説明変数を増加させることが可能であると考えられる．親疎と関係を持つ各言語要素を j とし，QADB 内の応答文における言語要素の出現傾向を表す変数を f_j^A とする時，この値は，式 (3) で計算される．

$$f_j^A = \frac{\sum_{i=1}^{m} s_i^Q y_{ij}^A}{m} \tag{3}$$

ここで，式 (3) における i は，QADB 内の各発話対を表しており，m が QADB 内の発話対の数を表している．また，y_{ij}^A は，3.2 節で述べた手法を用いて，QADB 内の応答文における各言語要素 j を変数化したものである．同様に，QADB の発話文における言語要素の出現傾向を f_j^Q とすると，この値は，s_i^R を使用して，式 (4) で計算される．

$$f_j^Q = \frac{\sum_{i=1}^{m} s_i^A y_{ij}^Q}{m} \tag{4}$$

式 (5) に示すように，これらの値を説明変数ベクトル $\boldsymbol{x}$ の要素とする．これにより，0 でない説明変数を増加させることが可能であると考えられる．

$$\boldsymbol{x} = [f_1^Q, ..., f_j^Q, ..., f_n^Q, f_1^A, ..., f_j^A, ..., f_n^A] \tag{5}$$

5 評価実験

本節では，提案手法の有効性を評価する．そのため，評価のために使用するデータの概要，評価実験の方法，結果と考察を順に述べる．

5.1 評価実験の使用データ

提案した発話対の親疎判別手法を評価するため，まず，3.1 節で述べたコーパスより発話対を抽出する．表 2 に，抽出された発話対の数を示す．ただし，初対面の発話対と友人同士の発話対の数が異なるため，初対面の発話対より 44,608 件を無作為抽出することにより，各発話対の数を揃えた．この結果，合計して 89,216 件の発話対が抽出された．さらに，本稿の提案手法では，親疎のラベルがついたデータの他に，ラベルなしデータである QADB を用意する必要がある．そこで，コーパスから抽出した発話対のうち，半数をラベル付きデータ，半数を QADB とする．これらの操作をすることで，親疎のラベルがついた発話対が 44,068 件，ラベルがついていない QADB が 44,068 件用意できた．

表 2 各コーパスにおける発話数

	友人	初対面	合計
日本語話し言葉	43,263	47,892	91,155
名大会話	2,176	1,345	3,521
合計	44,608	50,068	94,676

5.2 評価手法

提案手法の有効性を評価するため，3.4 で選択された言語要素を説明変数とする場合（既存 1），doc2vec を利用して生成した特徴ベクトルをそのまま説明変数と

表 3　評価実験の結果

	既存 1	既存 2	提案 1	提案 2
F 尺度	0.820	0.812	0.769	0.800

する場合（既存 2），4.1 節で述べた式 (1) と式 (2) で算出される類似度を利用して算出した特徴ベクトルを説明変数とする場合（提案 1），doc2vec によって算出された類似度を利用して特徴ベクトルを算出し，説明変数とした場合（提案 2）を比較する．

判別手法は，目的変数を初対面の発話対を 0，友人の発話対を 1 とした 2 値変数に設定し，Elastic Net ロジスティック回帰分析を行う．また，判別性能の評価指標として，F 尺度を算出する．

5.3　評価実験の結果と考察

表 3 に，評価実験を行った結果を示す．表 3 より，既存手法と比較して，提案手法の方が判別性能が低い結果となった．

以上の結果より，提案 2 の F-尺度が，提案 1 の F-尺度と比較して高かったことから，提案手法の判別性能は，発話対の類似度算出方法に影響を受けると考えられる．したがって，提案手法の判別性能が低かった原因として，類似度の算出法が適切でなかったことが挙げられる．

6　おわりに

本稿では，逐次的親密度推定の実現に向けて，発話対の親疎判別を実現する手法を提案した．しかし，提案手法は，既存手法よりも低い評価を得た．

今後の課題として，doc2vec のハイパーパラメタチューニングを行い，より正確に類似度計算を行うことや，類似度に閾値を設定し，ノイズとなる発話対を除去することが挙げられる．

謝辞

本稿の一部は独立行政法人日本学術振興会科研費 JP16H02908，JP26280115 の助成を受けたものである．

参考文献

[1] 宮澤幸希, 常世徹, 桝井祐介, 松尾智信, 菊池英明. 音声対話システムにおける継続欲求の高いインタラクションの要因. 電子情報通信学会論文誌 A, Vol. 95, No. 1, pp. 27–36, 2012.

[2] Cory D. Kidd and Cynthia Breazeal. Robots at home: Understanding long-term human-robot interaction. In *Proceedings of International Conference on Intelligent Robots and Systems*, pp. 3230–3235, 2008.

[3] Kazuyuki Matsumoto, Kyosuke Akita, Minoru Yoshida, Kenji Kita, and Fuji Ren. Estimate the intimacy of the characters based on their emotional states for application to non-task dialogue. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 327–333, 2015.

[4] Rumen Iliev, Morteza Dehaghani, and Eyal Sagi. Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition*, Vol. 7, No. 2, pp. 265–290, 2015.

[5] 柴田大作, 若宮翔子, 伊藤薫, 荒牧英治. JIWC: クラウドソーシングによる日本語感情表現辞書の構築. 言語処理学会 第 23 回年次大会 発表論文集, pp. 771–774, 2017.

[6] 中村明. 感情表現辞典. 東京堂出版, 1993.

[7] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196, 2014.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] 宇佐美まゆみ. BTSJ による日本語話し言葉コーパス（2011 年版）. 『人間の相互作用研究のための多言語会話コーパスの構築とその語用論的分析方法の開発』平成 20-22 年度科学研究費補助金基盤研究 B（課題番号 20320072）研究成果, 2011.

[10] 藤村逸子, 大曽美恵子, 大島 ディヴィッド義和. 会話コーパスの構築によるコミュニケーション研究. 藤村逸子, 滝沢直宏（編）, 言語研究の技法：データの収集と分析, pp. 43–72. ひつじ書房, 2011.

[11] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 67, No. 2, pp. 301–320, 2005.

[12] Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. Dialog system using real-time crowdsourcing and twitter large-scale corpus. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 227–231, 2012.