

ビジネスメールの敬語および形態素に着目した 送受信者間における親疎関係判定モデルの構築

寺本 優香[†] 楠 和馬[†] 蒲原 智也^{††} 波多野賢治^{††}

[†] 同志社大学大学院文化情報学研究所 〒610-0394 京都府京田辺市多々羅都谷1-3

^{††} 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

E-mail: †{teramoto,kusu,kambara,hitano}@ilab.doshisha.ac.jp

あらまし 日本語における敬語の運用規則は非常に複雑であるが、ビジネスシーンにおいては相手との人間関係を考慮し、適切な敬語を選択する能力が求められる。情報学の分野では、計算機を用いて敬語の文法規則を取得し、ユーザの敬語使用を補助する研究が行われている。本稿では、ビジネスシーンにおける親疎関係の敬意表現への影響を明らかにするため、ビジネスメールの例文を使用して、メールの送受信者間の親疎関係の判定を正確に行うモデルデータの入力となるデータ形式について調査し、過去に構築した親疎関係判定モデルの判別精度向上を目指す。

キーワード 敬語, 親疎関係, 上下関係, ビジネスメール

1. はじめに

敬語を含む敬意表現は、ビジネスの場をはじめとするさまざまなシーンにおいて円滑なコミュニケーションを行うための重要な働きを担っている。これは日本語が他の言語に比べ、敬語や改まった形式の表現(以下、敬意表現)の文法が非常に複雑な言語であり、それらを用いる際には、話し手、聞き手、話題内の登場人物など話題に関わる人物同士の関係性や、会話の内容、状況といった要因を考慮する必要があるためである [1]。

敬意表現と人間関係について、言語学の分野においてはポライトネス理論という考え方が存在する [2]。これは、相手を尊重する表現の丁寧さの度合いは、相手との社会的距離、力関係、相手と関わる際に与える負担の重さによって定まるというもので、世界標準の考え方とされている。一方、日本語の敬意表現に関する研究においては、年長者や経験者を敬う上下関係の概念、コミュニティの内部と外部を表すウチとソトの概念、そして相手との親密度を表す親疎の概念の三つが、敬意表現を決定する人間関係の要素と考えられている [1,3]。つまり、これらの人間関係の要素を考慮したうえで、考えられる表現の組合せパターンから相手と自分との距離感を適切に表す敬意表現を選ぶ必要がある。しかし、この運用規則にはしばしば曖昧さが残っており、敬意表現を用いる際の難しさとして指摘されている [3]。この難しさこそが、敬意表現の乱れや誤用の問題を生じさせていると考えられ、特にビジネスの場面における敬意表現の誤用は、社会的信用の失墜や深刻な誤解を招くこと可能性があるため、誤用を防ぐ方法が社会から求められている。

これらの課題を解決するために、近年は計算機を用いて敬意表現の運用規則を効率的に習得し、敬意表現の使用を補助するシステムの開発が行われている [4,5]。日本語における敬意表現の誤用を指摘するシステムを構築する際には、上下関係、ウチソト関係、親疎関係を把握したうえで、入力された敬意表現が適切であるかどうかを計算機が判断する必要がある。開発さ

れた敬語誤用判定システムでは、上下関係と組織におけるウチとソトの概念については考慮しているが、親疎関係については「相手との関係が親密である場合は敬意表現の規範はそれほど重要視されない」との考えから疎遠の場合のみを想定したシステム設計がなされている。

しかし、実際の人間関係における距離感には、前述したようにさまざまなバリエーションが存在するため、親疎関係の組合せを考慮すれば、文献 [4,5] のように疎の関係のみに固定することが問題となり得る場合が想定される。我々が過去に行った研究では、ビジネス上の関係性においても親疎関係に差が見られると仮定し、各人間関係においてどのような敬意表現が使われやすい傾向にあるのかを明らかにするために、メール例文内で使用されている敬語および形態素を索引語とし、ウチソト関係、親疎関係を説明するモデルを構築した [6]。本稿では、節に述べるこの判定モデルの問題点を解決するため、形態素解析後のデータに対し、説明変数に使用する語の複合語化、および品詞情報の追加に基づく索引語の細分化処理を加え、親疎ラベルの判定精度向上を目指す。

2. 関連研究

文献 [4,5] で開発されたシステムは、人手で話し手、聞き手、話題内の登場人物など話題に関わる人物同士の関係性をアノテーションし、状況ごとに使用される敬語表現との対応関係をデータベースに格納している。人物同士の関係性については、上下関係とウチソト関係を考慮している。テストデータを用いてシステムの評価を行った結果、95%の精度で誤用の指摘に成功しているが、アンケートによる有用性評価実験では、システムが採用している表現選択のルールは現実世界で使用されている敬語の文法と異なる印象があるという問題が指摘された。

この問題は、親疎関係を考慮することで改善する可能性が残されていると、我々は考える。なぜなら、実際の人間関係において疎の関係だけを考慮する状況は、現実には合わないと考え

表 1 メールデータに対するラベル付与と分類結果

ウチソト関係	親疎関係	親疎ラベル	定義	先行研究のデータ件数	本稿におけるデータ件数
社外	疎	社外一初対面	初めて関わる相手	47	53
	親	社外一既知	これまでに関わった相手	260	260
社内	疎	社内一部署外	異なるチームで働いている相手	47	53
	親	社内一部署内	同じ部、チームで働いている相手	39	53

られるからである。このため、親疎関係の有無が敬語の違いを表現している事実を統計的に確認することを目指す。

3. 事前実験

本節では、我々が文献 [6] で提案した親疎関係判定モデルについて解説する。このモデルは、ビジネスメールから抽出される敬語および形態素の定量データを用いて、表 1 に示した 4 種類の親疎ラベルを判定するモデルである。

3.1 モデル構築用データ

親疎関係判定モデルの構築にあたり、書籍として出版されているビジネスメール例文を使用した [7-18]。これらのデータを用いるメリットとして、出版社による校正を受けているため、正例データとして質が高いことが挙げられる、また、どのような状況下でそれらが書かれるべきなのかが明示されており、親疎ラベルのアノテーションがしやすいといった点が挙げられる。

3.2 定量データの抽出

各親疎ラベルに分類されたメールデータから敬意表現の抽出を行うため、JUMAN++ [19] による形態素解析を行い、各メールごとにその出現回数をカウントする。

あるメール $d_i (i = 1, 2, \dots, i')$ に出現する全形態素の個数を l_i としたとき、ある形態素 $t_j (j = 1, 2, \dots, j')$ の影響を表す値は $w_{t_j}^i$ はその出現回数 $n_{t_j}^i$ を用いて、以下のように表現される。

$$w_{t_j}^i = \frac{n_{t_j}^i}{l_i} \quad (1)$$

また、ロゴヴィスタ社製敬語辞典ソフトウェア「学研 敬語早わかり辞典」^(注1) に収録されている一つまたは複数の形態素からなる文字列の並びを表現を敬語と定義し、抽出を行った抽出されたある敬語 $h_k (k = 1, 2, \dots, k')$ があるメール d_i に $n_{h_k}^i$ 回出現した場合、敬語の定量データ $w_{h_k}^i$ は以下のように表現される。

$$w_{h_k}^i = \frac{n_{h_k}^i}{l_i} \quad (2)$$

3.3 判定モデルの構築

ビジネス上の関係性において、「社外一初対面」「社外一既知」「社内一部署外」「社内一部署内」の四種類の状況下を表す親疎ラベルを設定した。これは、ビジネスメールの例文を学習データとして用いたため、ほとんどが目上の相手を設定して作成されたものであり、上下関係を区別することが困難であったためである。また分類の簡素化のため、ウチソト関係と親疎関係の

有無でパタン分類をすることにした。敬語および形態素を索引語としたとき、その重みである定量データ $w_{t_j}^i$ と $w_{h_k}^i$ を説明変数とし、正準判別分析を適用したモデル式を作成した。「社外一初対面」「社外一既知」「社内一部署外」「社内一部署内」を親疎ラベルとして付与し、これを目的変数とする。 a_0 を定数、 $a_p (p = 1, 2, \dots, i', \dots)$ を係数としたとき、モデル式は式 (3) の通りに表される。ただし、今回は目的変数の種類が四つであるため、正準軸を求めるモデル式は三つ存在する。正準軸が表す三次元空間の学習データの散らばりから各親疎ラベルのカテゴリ空間を推定する。

$$L = a_0 + a_1 w_{t_1}^1 + \dots + a_{i'j'} w_{t_{j'}}^{i'} + a_{i'(j'+1)} w_{h_1}^1 + \dots + a_{i'(j'+k')} w_{h_{k'}}^{i'} \quad (3)$$

モデルの説明変数に使用する索引語の重みを、それぞれ敬語のみ、形態素のみ、敬語と形態素で定義した三種類のモデルを構築した結果、敬語のみで索引語を定義したモデルでは判定精度は非常に悪く、また敬語と形態素両方を用いたモデルでは、敬語の情報はほぼ活用されないことが判明した。

4. 提案手法

文献 [6] の問題点としては、以下の二点が挙げられる。

- 1) 複合語の一部を表す形態素を表現する索引語が説明変数として強く影響していた点、
- 2) メール の 話題 に 依存 する よう な 形態 素 である 索引 語 が 説明 変数 として 大きく 影響 を 与 えて いた 点、

このうち 1) に関しては、実際には異なる文脈で使われている形態素が、同一のものとして扱われている恐れがある。また 2) に関してはこのような索引語による判別は敬意表現の傾向を把握する上では有益ではなく、また、現実世界のメールにこれらの話題が多く含まれているということにはならないため、実用には適さない。

本稿では、親疎関係判定を行う際に、判別モデルの説明変数に使用する索引語の形式に着目し、複合語や品詞情報などを索引語に付与し、それらの重みを説明変数として扱う。この作業により、索引語の細分化を図り、さらのその処理の効果を調査する。

4.1 モデル構築用データ

我々が本稿で構築する親疎関係判定モデルにおいて、用いたメールデータの件数は表 1 に示した通りである。扱ったメール件数を追加した理由として、メール件数が少ないカテゴリの親疎ラベル判別精度が低かったことが挙げられる。そのため、新たに最近出版された書籍 [20-24] に掲載されているビジネス

(注1) : 学研 敬語早わかり辞典: https://www.logovista.co.jp/LVERP/shop/ItemDetail.aspx?contents_code=LVDGK07010(閲覧日 2018/1/10) .

表 2 タイプ A のモデルにおける学習データへの当てはまり

	L	評価値				計
		1	2	3	4	
実測値	1. 社外一初対面	48	3	1	1	53
	2. 社外一既知	6	47	0	0	53
	3. 社内一部署外	0	1	51	1	53
	4. 社内一部署内	0	0	7	46	53
計		53	53	53	53	212

表 3 タイプ B のモデルにおける学習データへの当てはまり

	L	評価値				計
		1	2	3	4	
実測値	1. 社外一初対面	47	5	1	0	53
	2. 社外一既知	4	48	1	0	53
	3. 社内一部署外	0	2	48	3	53
	4. 社内一部署内	0	0	9	44	53
計		51	55	59	47	212

表 4 タイプ A のモデルにおけるテストデータの判別結果

	L	評価値				計
		1	2	3	4	
実測値	1. 社外一初対面	9	1	0	0	10
	2. 社外一既知	1	9	0	0	10
	3. 社内一部署外	0	1	3	6	10
	4. 社内一部署内	0	0	2	8	10
計		14	7	9	10	40

表 5 タイプ B のモデルにおけるテストデータの判別結果

	L	評価値				計
		1	2	3	4	
実測値	1. 社外一初対面	8	2	0	0	10
	2. 社外一既知	1	9	0	0	10
	3. 社内一部署外	3	1	6	0	10
	4. 社内一部署内	0	0	3	7	10
計		9	12	9	7	40

メール例文を追加した。

4.2 追加した処理

本稿では文献 [6] でのモデルが抱える二つの問題点を解決するため、新たに処理を追加し、さらにそれらの効果について検証を行う。本節では、それらの問題点および対策について詳しく述べる。

処理 1: 語句の複合語化と品詞情報の追加

前述した問題 1) への対策として、連続する名詞を一つの名詞として扱う、いわゆる単語の複合化の処理を行うことが挙げられる。また同じ表記でも異なる意味を持つ語の存在が無視できないことが判明しているので、品詞情報を索引語に追加することで意味の違いを区別できるようにする。本稿では、これらの一連の処理を処理 1 とし、その効果について検証する。

処理 2: 定量データの抽出手法

問題 2) への解決策として、索引語の出現頻度を考慮したデータではなく、その出現有無のみに着目したデータを用いることが挙げられる。現在、我々の先行研究で作成した判定モデルで使用されている敬意表現の定量データは、形態素の種類ごとの出現頻度を含むデータである。しかし、規範的か否かに関しての判定であれば、単に索引語が出現しているかどうかの情報のみで判定することが可能であると考えられる。これにより、メールの話題に依存するような「開発」「取材」といった索引語の影響を抑えることができると考える。本稿では、この処理を処理 2 とし、索引語の定量データの抽出の際にこれらを適用した場合の効果を検証する。

索引語の出現有無のデータを用いて定量データの抽出を行う方法は以下のとおりである。あるメール $d_i (i = 1, 2, \dots, i')$ に出現する全索引語の個数を l_i としたとき、ある索引語 $t_j (j = 1, 2, \dots, j')$ の影響を表す値は $w_{t_j}^i$ はその出現回数 $n_{t_j}^i$ を用いて、以下のように表現される。

$$w_{t_j}^i = \begin{cases} 0 & (n_{t_j}^i = 0) \\ 1 & (n_{t_j}^i > 0) \end{cases} \quad (4)$$

処理 1 を行いビジネスメール例文内に出現する表現の定量データを集計した結果、639 種類の索引語を得た。定量化によって得られるデータは、列名を各索引語、行名を各メールとしたとき、メール内に出現する表現であれば 1、そうでなければ 0 の値を取るようなテーブルで表される。なお、本稿ではこれらの二値の行列データを、各セルの値の平均が 0、標準偏差が 1 となるような行列に変換してモデルに読み込ませる。

4.3 判定モデルの構築

索引語の有無を表す定量データを用いて、モデル構築を行う。モデル構築には多項ロジスティック回帰分析を用いる。文献 [6] 同様、目的変数のカテゴリが三つ以上の場合に用いられる手法である正準判別分析および、多項ロジスティック回帰分析を適用し、より分類精度の高かった手法による判定モデルを採用している。モデル式は式 (3) と同じものである。ただし、処理 1 の効果について考察するため、以下の 2 種類のモデルを比較する。

タイプ A: 処理 2 のみを行ったデータでモデルを構築

タイプ B: 処理 1, 2 両方を行ったデータでモデルを構築

構築したモデルの精度を見た結果、文献 [6] では正準判別分析が最も精度が高かったが、タイプ A, B のモデルにおいては、多項ロジスティック回帰分析のほうが精度が高かった。このため、タイプ A, B では、多項ロジスティック回帰分析をモデル構築の手法に採用する。なお、過学習対策として L1 正則化を行った。

4.4 評価実験と結果

文献 [6] と同様に、ビジネスマナーに関するサイト^(注2)から得られた送受信者間の関係が「社外一初対面」「社外一既知」「社内一部署外」「社内一部署内」に属するメールをそれぞれ 10 件ずつ収集する。これらをテストデータとして、構築したモデルにそれぞれ判定させることで評価実験を行う。

表 2, 表 3 はそれぞれ、今回新しく構築したタイプ A, タイプ B の判定モデルの結果を表している。タイプ A, タイプ B

(注2) : <http://email.chottu.net/> (2018 年 3 月 18 日 閲覧)

の当てはまりをそれぞれ F_{typeA} , F_{typeB} としたとき、それらは表 2, 3 より式 (5), 式 (6) と表すことができる.

$$F_{typeA} = \frac{48 + 47 + 51 + 46}{212} = 0.906 \quad (5)$$

$$F_{typeB} = \frac{47 + 48 + 48 + 44}{212} = 0.885 \quad (6)$$

タイプ A, B のモデルに対し行った評価実験の結果がそれぞれ表 4, 表 5 である. タイプ A, タイプ B の精度をそれぞれ A_{typeA} , A_{typeB} とすると, それらは式 (7), 式 (8) と表すことができる.

$$A_{typeA} = \frac{9 + 9 + 3 + 8}{10} = 0.700 \quad (7)$$

$$A_{typeB} = \frac{8 + 9 + 6 + 7}{10} = 0.750 \quad (8)$$

4.5 考 察

式 (7), 式 (8) は, 判定モデルに対して L1 正則化のためのパラメタ調整を行い, テストの精度が最も高かったパラメタで得られた結果であり, 今回構築したモデルは過学習対策済みであることから, 処理 1 を施した判定モデルのほうが精度が高い ($A_{typeA} < A_{typeB}$) ことが明らかになった. タイプ A, タイプ B の判定モデルにおいて説明変数を確認したところ, 処理 1 を行っていないタイプ A の場合, 「社」「株式」「部」「課」など, 部分的に出現しやすい形態素が判別に強く影響していることが明らかになった. これは, 我々の過去の研究でも課題として指摘されていた点である. 一方処理 1 を施したタイプ B の場合, 問題点は解決されていた.

一方, 「取引」「取材」など, メールの話題に依存すると考えられる形態素に関しては, 今回は影響を減らすことが出来なかった. これは, あるカテゴリの人間関係におけるメール例文に頻出する語句である. 例えば「取材」は, 「社外一初対面」のメール例文には, 取材の申し込みを行うシチュエーションが頻繁に登場するためであると考えられる. シチュエーションが同じようなメールが一定数を超えると, 処理 1, 処理 2 のみで抑えることが難しくなると考えられる.

5. おわりに

本稿では, ビジネスメールにおいて送受信者間の人間関係ごとに最適な敬意表現を提案するシステムの構築を最終目標とし, 敬意表現から親疎関係を判定するモデルを構築した. 文献 [6] で構築したモデルでは, 敬意表現以外の話題依存な語句や, 送受信者の所属, 役職名などが親疎関係の判定に大きく影響を与えていたが, 本稿ではデータ整形を行いそれらの影響を小さくすることで, 親疎関係判定モデルの精度は大きく向上し, また問題となった語句の影響はある程度抑えることができた.

一方, 頻出の話題に関連する索引語の一部は, 影響のある説明変数として残っている. この問題に対応するためには, 話題に依存する索引語の影響を抑える処理を新たに加える必要があると考える.

謝 辞

本研究の一部は日本学術振興会科学研究費助成事業基盤研究 (B) (課題番号: 15H02701) の助成を受けて実施された. ここに記して謝意を表す.

文 献

- [1] 菊池康人. 敬語. 講談社, 1997.
- [2] Penelope Brown and Stephen C. Levinson. *Politeness - Some Universals in Language Usage*. Cambridge University Press, 1987.
- [3] 文化審議会. 敬語の指針, 2007.
- [4] 白土保, 丸元聡子, 村田真樹, 井佐原均. 日本語発話文における敬語の誤用を指摘するシステムの開発. 自然言語処理, Vol. 13, No. 3, pp. 243-260, 2006.
- [5] 白土保, 丸元聡子, 村田真樹. 日本語の敬語誤用判定システム-判定ルール の妥当性の程度の数値化による柔軟な判定. 計量国語学, Vol. 28, No. 1, pp. 1-20, 2011.
- [6] 寺本優香, 楠和馬, 塩井隆円, 波多野賢治. メール送受信者間の親疎関係判定モデルの構築. DEIM Forum 2017 論文集, 2017.
- [7] ビジネス総合研究会. しっかり役立つビジネス文書実例集. 研究会法研, 2007.
- [8] 下條一郎. 史上最強のビジネス書 書き方&マナー. ナツメ社, 2014.
- [9] 神垣あゆみ. 考えすぎて書けない人のための 1 分間メール術. フォレスト出版, 2009.
- [10] 鈴木あつこ. すぐに使えて応用がきく ビジネス文書文例事典. 新星出版社, 2010.
- [11] 鈴木真理子. 就活から役に立つ 新社会人のためのビジネスメールの書き方. ソフトバンククリエイティブ株式会社, 2012.
- [12] 小川順子. 言いたいことが確実に伝わる メール の書き方 ビジネス・コミュニケーション術. あすか出版, 2016.
- [13] 木山泰嗣. カドが立たない 断りメール・謝りメール. 清文社, 2012.
- [14] 中川路亜紀. ビジネスメール文書術-知らない恥をかくルールから, そのまま使える文例まで. ダイヤモンド社, 2015.
- [15] 西出ひろ子. 仕事の基本 正しいビジネスメールの書き方. 日本能率協会マネジメントマナー, 2016.
- [16] 神垣あゆみ. メールは一分で返しなさい! 日本能率協会マネジメントマナー, 2009.
- [17] 平野友朗, 直井章子. カリスマ講師に学ぶ! 実践ビジネスメール教室. 日経 BP マーケティング, 2015.
- [18] 石黒圭. 正確に伝わる! わかりやすい文章の書き方. 日本経済新聞社, 2012.
- [19] Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2292-2297. ACL, 2015.
- [20] 山崎政志. いちばん伝わる! ビジネス文書の書き方とマナー. 高橋書店, 2010.
- [21] 坂本直紀, 深津伸子. ビジネスで使える メール文例集. 新日本法規, 2014.
- [22] 日本語書研究会. 超早引き! ビジネス書の書き方 例 500. 主婦と生活社, 2015.
- [23] 鈴木真理子. ビジネス文書 メールの基本. すばる社, 2016.
- [24] 芦屋広太. 気になる「症状」治します! ビジネス文書クリニック. 日経 BP 社, 2015.