# メール送受信者間の親疎関係判定モデルの構築

寺本　優香† 　塩井　隆円†† 　楠 　和馬†† 　波多野賢治†††

† 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
†† 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
††† 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,shioi,kusu}@ilab.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

**あらまし**　敬語表現はコミュニケーションにおいて重要な役割を担っており，敬語使用を支援するためのツールの開発が盛んに行われている．しかし，人間関係の重要な観点の一つである親疎関係と敬語表現にどのような関係があるのかについての研究はほとんどなされていない．本研究では，ビジネスメールの例文のデータを用いて，人物間の親疎関係を判断するモデルの構築を行い，親疎関係と敬語表現の関連性についての知識獲得を目指す．

**キーワード**　敬語，ビジネスメール，親疎関係

## 1. はじめに

敬語表現とはコミュニケーションにおいて他者への尊重を表す目的で用いられる表現の一つである．そのため，日本語における敬語表現は，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，相手と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 [?, ?]．しかし，敬語表現を用いる際には，年長者や経験者を敬う上下関係の概念や，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を考慮に入れ，相手ごとに自分との距離感を適切に表す表現を選ぶ必要がある．この運用規則の曖昧さが敬語表現を用いる際の難しさとして指摘されており [?]，しばしば敬語表現の乱れや誤用の問題が生じている．特に，ビジネスの場面における敬語表現の誤用は，社会的信用の失墜や深刻な誤解を招くことが考えられる．

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語表現の運用規則を効率的に習得し，敬語表現の使用を補助するシステムの開発が行われている [?, ?, ?]．敬語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，入力された敬語表現が適切であるかどうかを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．しかし，親疎関係については親と疎の二値のみを想定したうえで,「相手との関係が親である場合は敬語の規範はそれほど重要視されない」との考えから疎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様々なバリエーションが存在する．上下関係，ウチソト関係そして親疎関係の組合せを考慮すれば，既存研究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ため，敬語表現を始めとする表現手法にどのような影響を及ぼしているのかを明らかにする必要がある．

本研究では，ビジネス上の関係性において,「社外ー初対面」「社外ー既知」「社内ー部署外」「社内ー部署内」の四種類の状況下で親疎関係に差が見られると仮定する．これは使用するデータの特性上，上下関係を区別することが困難であったため，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ためである．この仮定の下，ビジネスマナーに関する書籍に掲載されているビジネスメールの例文を学習データとして用いた親疎判定モデルを作成し，その妥当性について考察を行う．

## 2. 関連研究

1. でも述べたように，ビジネスの場面での会話を想定し敬語の誤用を指摘するシステムの開発が白土らにより行われている．このシステムは，事前に人手で作成した敬語を含む文章から，話し手，聞き手，文の主語の上下関係，ウチソト関係を抽出し，それらの関係と使用されている敬語を対応付けその関係をデータベースに格納している．このデータベースを利用して，入力された人間関係の条件と敬語の組合せがデータベース内に存在しない場合は敬語の誤用と判断している．

テストデータを用いた実験では，95% の精度で誤用の指摘に成功している．しかし，アンケートを用いた評価実験では，実際に世間で使用されている敬語の規範と修正システムが採用している敬語の規範との乖離が見られたという報告がなされている．これは，上下関係，組織のウチソト関係の概念には考慮しているが，親疎関係に関しては全て疎を前提としていることに原因がある可能性が残されている．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られるからである．このため，親疎関係の有無が敬語表現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．

そこで本研究では，人物間の親疎関係を判断するためのモデル構築を行い，親疎関係と敬語表現の関連性を明らかにするための知識獲得を目指す．

## 3. 親疎関係判定モデルの構築

本研究では，ビジネス上の関係性において,「社外ー初対面」「社外ー既知」「社内ー部署外」「社内ー部署内」の四種類の状況下で親疎関係に差が見られると仮定した．この理由は，敬語

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 | データ件数 |
|---|---|---|---|---|
| 社外 | 疎 | 社外ー初対面 | 初めて関わる相手 | 47 |
| | 親 | 社外ー既知 | これまでに関わった相手 | 260 |
| 社内 | 疎 | 社内ー部署外 | 異なるチームで働いている相手 | 47 |
| | 親 | 社内ー部署内 | 同じ部，チームで働いている相手 | 39 |

は主体とその相手，もしくは話題中の人物との社会的関係を表す言語表現であることを考えると，ポライトネス (politeness) を実現するために用いられるからである．Brown と Levinson によって確立された「ポライトネス理論」[?] では，ポライトネスは対話者との力関係と社会的距離，そして対人配慮の負荷の三つの社会的変数の大小に基づいて行われていると主張している．今回用いたデータの特性上，メールを送る際に特に注意が必要である目上の相手を想定したメール例文がほとんどであった．このため，ウチソト関係と親疎関係を，力関係以外の 2 変数と対応させ，上記四種類の状況を考慮することにした．

### 3.1　モデル構築用データ

実際のメールは業務上の機密情報を多く含むため，分析が可能な形での提供を受けることが難しい．そこで本研究では，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．書籍を選ぶにあたり，前述の四種類の状況を判別できるように紙媒体のビジネス文書の例文とビジネスメールの例文とを明確に区別している書籍を使用した [?,?,?,?,?,?,?]．これは，多くのビジネスに関するマナー本では，メールの場合，内容を効率的に伝達するため，簡潔な文章を用いることを推奨する一方で，伝統的な紙面上の文章では格式張った形式で書くことを推奨しているからである．したがって，メールに用いる表現であることを明記している例文のみをビジネスメールの例文データとして用いている．最終的にはこれらの書籍から，332 通の社外メールと 109 通の社内メールのデータを抽出することができた．また，これらメールデータに対し，親疎関係を表現するラベルとして前述の四種類のラベルを手動で付与したところ，表 1 の通りとなった．

メール例文が社内向けであるか社外向けであるかについては，書籍において明示されていることが多かったため，それに従いラベル付与を行った．また親疎ラベルは，メールの例文に「新たな取引を申し込むためにメールを送る場合」というように，メールを送る際のシチュエーションが設定されていることが多かったため，それに従ってラベルの付与を行った．その一方でこのような設定が明記されていないメール例文の場合は，メール本文の「昨日はありがとうございました」や「いつもお世話になっております」といった記述から付与すべきラベルを推測した．最終的に人手による決定，推測が困難なメール例文は社外メールに 23 通，社内メールに 23 通含まれていたが，これらはモデル構築には使用しないものとした．

### 3.2　定量データの抽出

本研究では定量的データ分析を実施するために，3.1 で述べたラベル付きメールデータから形態素と敬語表現の抽出を行う．形態素の抽出には，京都大学大学院情報学研究科の黒橋・河原研究室で開発されている JUMAN++ [?] を用い，各メールごとにその出現回数をカウントした．あるメール $d_i(i = 1, 2, \ldots, i')$ 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 $t_j(j = 1, 2, \ldots, \mathrm{j}')$ の定量データ $w^i_{t_j}$ はその出現回数 $n^i_{t_j}$ を用いて，以下のように表現される．

$$w^i_{t_j} = \frac{n^i_{t_j}}{l_i} \tag{1}$$

一方，敬語表現の定量データについては，形態素の定量データの計算と同様，メールごとにその出現回数を抽出し，メール内の全形態素数で正規化を行う．敬語表現は抽出するためのツールが存在しないため，文化庁文化審議会の答申である「敬語の指針」[?] やそれに基づいて開発され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) に収録されている表現を敬語として定義し，自作プログラムによる敬語表現の抽出を行った抽出されたある敬語表現 $h_k(k = 1, 2, \ldots, k')$ があるメール $d_i$ に $n^i_{h_k}$ 回出現した場合，敬語表現の定量データ $w^i_{h_k}$ は以下のように表現される．

$$w^i_{h_k} = \frac{n^i_{h_k}}{l_i} \tag{2}$$

### 3.3　モデルの構築

3.2 によって抽出された定量データを用いて，モデル構築用データを四種類の親疎ラベルに分類するためのモデル構築を行う．

3.1 で述べたように，メールデータには既に「社外ー初対面」「社外ー既知」「社内ー部署外」「社内ー部署内」のラベルが付与されているが，この四つの状況下を親疎レベルとして表現すれば表 2 のように定義できる．

表 2　親疎ラベルと親疎レベルの対応

| 親疎ラベル | 親疎レベル |
|---|---|
| 社外ー初対面 | 1 |
| 社外ー既知 | 2 |
| 社内ー部署外 | 3 |
| 社内ー部署内 | 4 |

これは社外の人物ほど，また顔を合わす頻度が低いほど親疎関係は薄くなるからである．

したがって，統計的にモデル構築を行う場合は，親疎レベルを目的変数 $L$ に，そして形態素と敬語表現の定量データ $w^i_{t_j}$ と $w^i_{h_k}$ を説明変数として用いればよい．一般的にどのような分析

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

を行うべきかはさまざまな意見が存在するが，文献 [?] によると，各データ (本研究の場合はメールデータ) がどのカテゴリに属しているかが明確であるデータが存在する場合は判別分析を用いることができるとされているため，本研究においてもそれに倣うことにした．特に本研究の場合はカテゴリ数が 4 であるため，正準判別分析 (Canonical Discriminant Analysis) を適用するためのモデル式は，$a_0$ を定数，$a_p(p = 1, 2, \ldots, i', \ldots)$ を係数とした場合，以下のようになる．

$$\begin{aligned} L = a_0 + a_1 w_{t_1}^1 \\ + \ldots + a_{i'*j'} w_{t_{j'}}^{i'} + a_{i'*(j'+1)} w_{h_1}^1 \\ + \ldots + a_{i'*(j'+k')} w_{h_{k'}}^{i'} \end{aligned} \tag{3}$$

当然，式 (3) を用いれば，新規データの各カテゴリへの所属を推定することも可能であるため，インターネット上に公開されているメールデータを用いて性能評価を行うこともできる．

### 3.4 Wilks' Lambda

多変量解析を行う場合，式 (3) のようなモデル式を定義すると，説明変数が多い場合は，モデルの解釈や目的変数の値を予測する場合に実用的ではない場合が多い．そのために，説明変数を選択するために事前知識を元に変数を指定したり，全ての説明変数の組合せから最適なものを選択したり，規則に従って説明変数を逐次選択する方法が用いられる．一般的には，ステップワイズ法を用いて統計的に最も予測率が高いと考えられる説明変数から順にモデルに適用していく方法が採られるが，最終的には選択された説明変数によって構築されたモデルが有用であるかどうかをさまざまな検定法によって評価する．

この場合に使用される検定法は，モデル式に正準判別分析を使用していることから多変量分散分析の規準で行われるべきである．このため，複数のカテゴリの重心が全て重なっているかどうかを「各カテゴリの平均に差はない (平均値ベクトルが等しい)」とする帰無仮説と，「少なくとも一組のカテゴリ間の平均には差がある (平均値ベクトルは異なる)」とする対立仮説を設定し，それを検定することになる．こうした検定法には，

- Wilks' Lambda
- Lawley-Hotelling Trace
- Pillai's Trace
- Roy's Maximum Root

が存在する [?] が，本研究では，さまざまな統計解析ソフトウェアでも利用できる Wilks' Lambda を利用した検定法を用いて変数選択を行い，そうして選択した変数でモデル構築を行った．

モデル構築に使用したメールデータは，親疎ラベルが「社外ー既知」の件数だけが突出して多かった (表 1 参照) ため，カテゴリ内のメール件数の偏りが構築されたモデルの精度に影響することを防ぐために，他の親疎ラベルが付与されたメール件数と同等の 47 件をランダムに抽出し，計 180 件のメールデータを用いてモデル構築を行った．また，モデルの説明変数は次の三種類，1) 形態素のみ，2) 敬語のみ，3) 形態素＋敬語 を設定し，それぞれ Wilks' Lambda を利用した検定法により説明変数を選択した．

モデル構築に使用したデータの中で最もモデルの当てはまりが良かったのは説明変数に形態素のみを用いた場合（以下，モデル 1) とする）であり，形態素の説明変数は 42 種に選択された．図 1 は正準判別分析により導出される二つの正準軸が成す空間にモデル構築に利用したデータを分布させた結果とそれぞれのカテゴリ空間を推定した結果である．

モデル構築用のデータに対しての判別精度は 98 % と高精度な判別を可能にし，正準判別分析によって得られた結果は図 1 のような空間を形成する．図 1 左上は第 1 正準軸（LD1），第 2 正準軸（LD2）により作られる空間であり，図 1 右上は第 1 正準軸（LD1），第 3 正準軸（LD3）により作られる空間である．また，図 1 下側は上側の図にカテゴリごとの 95 % 信頼楕円を描いた図である．あるカテゴリに属するメールデータが対応する楕円の中に 95 % の確率で分布するという見方である．

次に図 2 はモデル 1) の第 1 正準軸および第 2 正準軸が成す 2 次平面上に，形態素 42 種がそれぞれどのカテゴリの判別に寄与しているか判断を支援するための図である．形態素ごとに対応する線分の長さおよびその線分が伸びている方向に従ってどのカテゴリの判別に寄与しているのか判断できる．図 2 より，形態素「御」は親疎レベル 1,「格別」や「納入」などは親疎レベル 2,「疲れ」や「長」などは親疎レベル 3,「平素」は親疎レベル 4 の判別に最も寄与していると判断できる．第 1 正準軸および第 2 正準軸がともに 0 に近い座標に位置している場合は判別に重要な変数では無いということも観測が可能である．

### 3.5 Web 上のメールデータによるモデルの評価

モデル 1) が学習データ以外のメールデータにも当てはまるかどうか確認するために評価実験を行う．この実験で利用するメールデータは，本研究で取り扱った親疎ラベルを判断することが可能な Web サイト (注2) から，各カテゴリ毎に 10 件ずつ取得した．

表 3 はモデル 1) の判別結果を表している．したがって，モデル 1) の全体の判別精度 $A_{tbl3}$ は正確度を用いて，

$$A_{tbl3} = \frac{8+3+4+8}{40} = 0.575 \tag{4}$$

と表すことができる．

表 3 判 別 結 果

| | | 評価値 | | | | 計 |
|---|---|---|---|---|---|---|
| | $L$ | 1 | 2 | 3 | 4 | |
| 実測値 | 1 | **8 ( 80%)** | 1 (10%) | 1 (10%) | 0 ( 0%) | 10 ( 25%) |
| | 2 | 4 ( 40%) | **3 (30%)** | 0 ( 0%) | 3 ( 30%) | 10 ( 25%) |
| | 3 | 0 ( 0%) | 2 (20%) | **4 (40%)** | 4 ( 40%) | 10 ( 25%) |
| | 4 | 1 ( 10%) | 0 ( 0%) | 1 (10%) | **8 ( 80%)** | 10 ( 25%) |
| 計 | | 13 (32.5%) | 6 (15%) | 6 (15%) | 15 (37.5%) | 40 (100%) |

正確度が 0.575 は判別精度としては，それほど良いとは言えない．これは 3.4 で述べたとおり，書籍から取得できたメールデータのカテゴリ間に偏りがあり，学習データが不足するカテ

(注2)：ビジネスメールの書き方．http://email.chottu.net/ （2017 年 2 月 17 日閲覧）

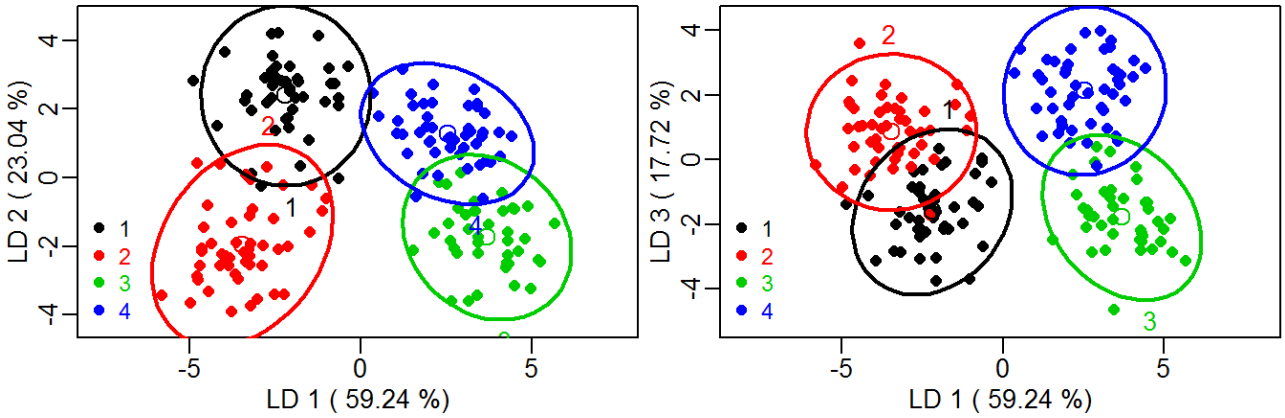


図 1　モデル 1) の正準空間およびデータの分布

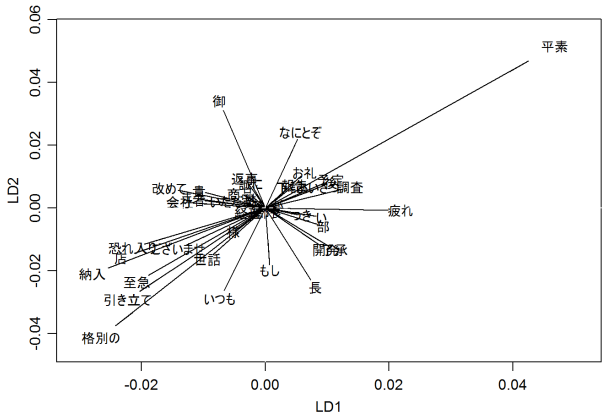


図 2　形態素 42 種ごとの各カテゴリへの寄与

表 4　判別結果 (ウチソト関係単独)

| | | 評価値 | | 計 |
|---|---|---|---|---|
| | $L$ | 1, 2 | 3, 4 | |
| 実測値 | 1, 2 | **16 (80%)** | 4 (20%) | 20 ( 50%) |
| | 3, 4 | 3 (15%) | **17 (85%)** | 20 ( 50%) |
| 計 | | 19 (47.5%) | 21 (53.5%) | 40 (100%) |

表 5　判別結果 (親疎関係単独)

| | | 評価値 | | 計 |
|---|---|---|---|---|
| | $L$ | 1, 3 | 2, 4 | |
| 実測値 | 1, 3 | **13 (65%)** | 7 (35%) | 20 ( 50%) |
| | 2, 4 | 6 (30%) | **14 (70%)** | 20 ( 50%) |
| 計 | | 19 (47.5%) | 21 (53.5%) | 40 (100%) |

ゴリがあったためである．今後は判別精度を高めるために親疎レベル 2 以外の学習データをより多く取得する必要がある．また図 2 から,「開発」および「部」という二つの形態素が変数選択されているが，これらはそもそも「開発部」という会社内の所属を示す語である．他にも所属を表す語は学習データ中にも含まれているが,「開発」よりも少ないか，他のカテゴリのメールにも存在していたため変数選択で選択されなかったことが考えられる．これは学習データに依存する問題であるため，より一般的なモデルを構築するためにも，形態素解析の際に所属を表す一つの語として認識するなど，学習データの高品質化が必要である．

一方，ウチソト関係や親疎関係をそれぞれ単独で考慮し判別精度を評価したところ，ウチソト関係単独は表 4，親疎関係単独は 5 のようになった．したがって，それぞれの判別結果の正確度 $A_{tbl4}$，$A_{tbl5}$ は，

$$A_{tbl4} = \frac{16+17}{40} = 0.775 \tag{5}$$

$$A_{tbl5} = \frac{13+14}{40} = 0.675 \tag{6}$$

と計算することができる．

この結果から，ウチソト関係，親疎関係単独での判別は概ね上手くいっているものの，それらを合わせて考慮した場合は判別精度が良くなかったと言える．このことから，親疎レベル 2, 3 の判別精度の向上を図るための新たな方策を考える必要がある．

## 4. おわりに

本研究では，ビジネスマナーに関する書籍に掲載されているビジネスメールの例文を正準判別分析の学習データとして用いて，メールデータの親疎判定モデルを構築し，その精度を評価した．

構築したモデルの中では形態素のみを用いて構築したモデルが最も学習データの当てはまりがよく，98 % の判別精度が得られた．本研究における仮定の正しさが確認されたことにより，敬語表現には親疎関係を考慮する必要であることが確認できた．一方，学習データ以外のメールデータに対してもウチソト関係と親疎関係の有無の判別が可能かを確かめた．その結果ウチソト関係，親疎関係単独の場合は高精度な判別が可能であったが，それら両方を考慮した際の判別は精度が良くなかった．つまり，ウチソト関係，親疎関係ごとに表現の違いがあることの確認はできたが，トピック依存の語句の影響が強く見られ，親疎レベル 2，3 の判別が困難であった．このため親疎レベル 2，3 の敬語表現にどのような差があるかに関しては今後分析していく

必要がある．

今後の課題としては，親疎レベル２以外のメールデータが不足しているため学習データを増やす必要がある．また，所属を表す語や会社名を表す語をそれぞれ認識するよう工夫した上で，再度，正準判別分析により判別モデルを構築する必要がある．

さらに，データの都合でポライトネス理論における力関係を表す社会的変数を今回は考慮できなかったが，この社会的変数を考慮にいれることが可能なデータを用いて敬語表現が 8 パタンで表現できることを確認し，敬語表現のパタン分類ができる分類器の構築を行う必要がある．

## 謝　　辞

本研究を行うにあたり，同志社大学文化情報学部の矢野環教授には，親疎関係判定モデルの構築に際し，有益なコメントを頂いた．また，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