言語表現を利用したメッセージの送り手の性別判定

3B−3

今村　賢治　　堀井　統之　　大山　芳史
NTT情報通信網研究所

1.はじめに

電報等のメッセージの中には、我々が読んだとき、送り手の性別を感じさせるものがある。例えば、「おまえもがんばれよ」という句が入ったメッセージの送り手の性別は男性、「あなたもがんばってね」という句なら女性と感じることができる。このように、同じ意味であるにも関わらず、送り手の性別の違いを感じるということは、メッセージに何らかの言語的特徴があることを示している。本稿では、言語表現という特徴を用いて、メッセージの送り手の性別を判定する方式について検討した結果を報告する。

2.性別判定のための言語表現

メッセージ約1300に関して、男性及び女性に「あなたなら内容を変えずに、どう書くか」という観点で書き換え実験を行い、原文の種別で分類した結果を表1に示す[1][2]*。書き換えられた部分は、男性が書き換えたものであれば書き換え元が女性表現、書き換え先が男性表現もしくは男性でも女性でも使用する中立表現、女性が書き換えた部分はその逆であると考えられる。このように考えた場合、書き換えが頻繁に起こった種類は、男性表現及び女性表現の特徴的部分となる。表1では、丁寧さの変更、人を表す名詞(人称代名詞、固有名詞、人を意味する一般名詞等)、文末表現の3種類がそれに相当している。以上の考察から、本稿では、丁寧さ、人を表す名詞、文末表現の3種の言語表現からメッセージの送り手の性別を判定することとする。

しかしながら、丁寧さに関しては、機械的に処理を行う場合、独立した言語表現として捉えることは困難である。なぜなら、丁寧さに影響を与える言語表現として、体言に付く「御」や、謙譲語・尊敬語・丁寧語等があるが、謙譲語・尊敬語・丁寧語については文末表現との区別が困難であり、体言の「御」については、それが人を表す名詞(特に一般名詞)に付加されている場合の区別が困難だからである。そのため、今回は丁寧さに関しては、メッセージを読んだ人が主観的に付与した数値(1〜5の5段階、5が最も丁寧)を利用することにした。

表1．メッセージ書き換え実験結果

	男性→女性	女性→男性
丁寧さの変更	320(52%)	141(28%)
人を表す名詞	146(24%)	121(24%)
文末表現	111(18%)	140(28%)
標準からの逸脱形	8(1%)	31(6%)
俗語	39(6%)	5(1%)
古語・文語	14(2%)	
卑語	1(0%)	
省略等		26(5%)
美化語		16(3%)
副詞		17(3%)
接続詞	7(1%)	
強意語		3(1%)
感嘆詞	1(0%)	
未分類	31(5%)	7(1%)

*男性→女性表現書き換えについては、文献[1]では約1700メッセージであったが、今回は女性→男性表現書き換えと統一し、再整理した。

以下、人を表す名詞、文末表現各々に関して、男性表現、女性表現のルールを作成し、それをメッセージに適用して決定したメッセージ全体の言語表現と、メッセージを人が判断して送り手の性別（男性、女性および不明の3種類に分類）を付与したデータとの比較実験を行った結果を述べる。人が判断した性別のうち、「不明」とは、男性でも女性でも送れるメッセージであることを示す。なお、以下の実験に使用したメッセージは、結婚、誕生日、母の日、父の日のお祝い電報例文626で、うち398メッセージが書き換え実験に使用したものである。

2.1.文末表現の影響

文末表現とは、本稿では各文の最終文節の表現と定義する。文末表現は、主として用言と助動詞、終助詞の列であるが、体言と助動詞列、省略が起こった場合には、体言と助詞の場合もある。

まず、文末表現のみを利用し、メッセージの送り手の性別を判定する実験を行った。文末表現はルール化して用いたが、そのルール作成の手順は以下のとおりである。

1.メッセージ書き換え実験の書き換え部分のうち、文末表現に関する部分を取り出し、これを形態素の列としてパターン化する（以下、これを文末表現パターンと呼ぶ）。例えば、「わかっ／て／いる／わ／よ」という文末は、動詞音便形＋接続助詞「て」＋補助動詞「いる」終止形＋終助詞「わ」＋終助詞「よ」とパターン化する。そしてこれを男性表現、女性表現、中立表現に分類し、男性表現パターン、女性表現パターンの原本とする。

2.実験対象メッセージのすべての文末表現を文末表現パターンとして、1.の中立表現パターンに追加する。

3.一つの文末表現パターンは、複数のメッセージに現われるので、中立表現の各パターンについて、人手で付与したメッセージの送り手の性別をカウントする。

4.3.で得られたカウントが、男性>>女性であるものは男性表現パターン、女性>>男性であるものは女性表現パターンとする。文末表現に関しては、女性および不明メッセージに現われていないものを男性表現パターン、男性および不明メッセージに現われていないものを女性表現パターンとした。

以上の手順で作成した男性表現パターン104と、女性表現パターン126を使用し、メッセージ全体で送り手の性別判定の実験を行った。判定規則は、メッセージの文末表現のうち、男性表現と女性表

A sex identification method of a message sender using linguistic characteristics
Kenzi IMAMURA, Motoyuki HORII, Yoshifumi OOYAMA
NTT Network Information Systems Laboratories

表２．文末表現による送り手の性別判定結果

		文末表現による判定			計
		男性	女性	不明	
人手による判定	男性	１０２	９	１３８	２４９
	女性	４	１０３	１１７	２２４
	不明	１０	１４	１２９	１５３
計		１１６	１２６	３８４	６２６

現が使われている回数を各々カウントし、多いほうの性別がメッセージの送り手の性別であるとした。実験結果を表2に示す。

人手で付与した性別と、文末表現で判定した性別が一致したものは53%となった。

２．２．人を表す名詞の影響

人を表す名詞とは、本稿では、人称代名詞・固有名詞・人を表す一般名詞等のことである。人称代名詞は単語自体に(「俺」が男性表現等)、固有名詞は「さん」「君」などの接尾辞に、一般名詞は単語と接頭辞・接尾辞などに(「奥様」「お坊ちゃん」等)、性別を感じさせる表現が現われる。

この人を表す名詞に関しても、2.1.節で述べた文末表現パターンと同様な手順でルールを作成した。ただし、4.項の出現回数のカウントにおける男性>>女性または女性>>男性の判断は、出現頻度が3倍以上のものを男性表現、女性表現とした。

このように作成した男性表現パターン36、女性表現パターン34を使用し、文末表現と同様の実験を行ったところ、表3の結果が得られた。

表３．人を表す名詞による送り手の性別判定結果

		人を表す名詞による判定			計
		男性	女性	不明	
人手による判定	男性	９２	１７	１４０	２４９
	女性	１３	１０６	１０５	２２４
	不明	１４	２６	１１３	１５３
計		１１９	１４９	３５８	６２６

人手で付与した性別と、人を表す名詞で判定した性別が一致したものは50%となり、文末表現で判定した場合とほとんど同じ割合となった。しかし、文末表現の場合と比べ、性別判定を誤ったもの（人手で付与した性別が男性にも関わらず、女性であると判定した場合、あるいはその逆）は626中の30メッセージ（5%）と増えている。これはルール作成の際に男性>>女性または女性>>男性の判断を、文末表現の場合ほど厳密に行わなかったからである。

３．組み合わせた場合の判定

３．１．線形判別関数の算出

第2章で作成した言語表現パターン等を組み合わせて送り手の性別を判定するにあたり、判定方法として線形判別関数[3]を利用することとした。線形判別関数は、重回帰式の定数項を変化させたものである。

文末表現、人を表す名詞に関しては、説明変数x_e, x_hとして以下の値を用いて言語表現パターンと一致した個数を頻度に直した。

$$Xe, Xh = \frac{\text{男性表現数} - \text{女性表現数}}{2(\text{男性表現数} + \text{女性表現数})} + 0.5 \qquad (1)$$

丁寧さに関しては、1〜5の5段階データなので、説明変数x_pは(2)式と設定した。目的変数であるメッセージの送り手の性別yは、男性を1、女性を0、不明を0.5として線形判別関数を算出した。

$$Xp = (\text{丁寧さ} - 1) / 4 \qquad (2)$$

(1)文末表現、人を表す名詞

まず、説明変数を文末表現、人を表す名詞に限り、線形判別関数を算出したところ、以下のとおりとなった。

$$f(Xe, Xh) = 0.639Xe + 0.417Xh \qquad (3)$$

判別境界は、人手により男性のメッセージと判定されたものの判別値$f(x_e, x_h)$の平均Mと不明と判定されたものの判別値の平均Nとの中点を男性／不明の境界、不明と判定されたものNと女性と判定されたものFの中点を女性／不明の境界とした。すなわち、

$$\left.\begin{array}{l} \text{男性}: f(Xe, Xh) \geqq (M+N)/2 = 0.605 \\ \text{女性}: f(Xe, Xh) \leqq (F+N)/2 = 0.402 \end{array}\right\} \qquad (4)$$

とした。

(3)式のx_e, x_hの係数は、送り手の性別判定における文末表現、人を表す名詞の重要性を示している。従って、送り手の性別を感じさせるような言語表現は、人を表す名詞より文末表現の方が重要であると言える。

(2)文末表現、人を表す名詞、丁寧さ

文末表現、人を表す名詞、丁寧さの3種を組み合わせて線形判別関数を算出したところ、以下のとおりとなった。

$$f(Xe, Xh, Xp) = 0.621Xe + 0.410Xh - 0.316Xp \qquad (5)$$

$$\left.\begin{array}{l} \text{男性}: f(Xe, Xh, Xp) \geqq (M+N)/2 = 0.526 \\ \text{女性}: f(Xe, Xh, Xp) \leqq (F+N)/2 = 0.320 \end{array}\right\} \qquad (6)$$

３．２．性別判定実験

前節で算出した線形判別関数を用い、メッセージの送り手の性別判定実験を行った結果を表4に示す。人手で付与した性別と、言語表現を利用して判定した性別が一致したものは、文末表現・人を表す名詞で行ったもの（(3)(4)式を使用）も、3種類すべてを組み合わせたもの（(5)(6)式使用）も、どちらも62%であった。丁寧さを含めた場合と含めない場合の結果がほとんど同じだったのは、丁寧さの要素が文末表現および人を表す名詞に吸収されてしまったからであると考えられる。

表４．組み合わせによる送り手の性別判定結果

		文末表現・人を表す名詞を用いた判定			文末表現・人を表す名詞・丁寧さを用いた判定			計
		男性	女性	不明	男性	女性	不明	
人手による判定	男性	１４２	１７	９０	１４１	１９	８９	２４９
	女性	１３	１４９	６２	１２	１５０	６２	２２４
	不明	１９	３７	９７	１９	３７	９７	１５３
計		１７４	２０３	２４９	１７２	２０６	２４８	６２６

また、62%という適合率は決して高いとはいえない。今回は、言語表現パターン作成に用いた例文で性別判定実験を行ったので、性別に影響する言語表現はほとんど抽出できたはずである。にも関わらずこのような結果が出たことは、言語表現による性別判定の限界を示している。さらに良い結果を出すためには、メッセージで使われている話題等の意味も考慮する必要がある。

４．まとめ

言語表現として、文末表現、人を表す名詞、丁寧さに着目し、メッセージの送り手の性別判定を行った。そして、約6割のについて、人手で付与した送り手の性別と一致した結果を得た。今後は、話題等のメッセージの意味も考慮して判定精度を上げてゆく。

参考文献

[1] 堀井他「メッセージにおける言語表現の分析とその生成」、情処NL研(78-14), 1990

[2] 今村他「メッセージの女性→男性表現変換の検討」、第44回情処全大(3Q-7), 1992

[3]奥野他「多変量解析法<<改訂版>>」、日科技連出版社, 1981など