# 人を表す指示語に対する性別判定

梁 寛侑　　松本忠博
岐阜大学 工学部 電気電子情報工学科

## １．はじめに

今日の機械翻訳は日々の翻訳技術の進化によってその精度は日々向上している．しかしながら今日においても完璧な翻訳とは言えず，課題が未だ多い．その一例として，「日本語の指示語に対する翻訳」というものがある．現在の機械翻訳，特にユーザーが利用しやすい翻訳サイトの翻訳を見ると，指示語（あいつ・こいつなど）は性別を問わず使用でき，そのため明確に性別が別れる英語の代名詞に正確に対応ができていないのが現状である．例えば，「あいつは看護婦だ．」という文章が，ある翻訳サイトでは，「That fellow is a nurse.」と翻訳される．この翻訳結果は間違いではないが，ナースは女性であるので，「She is a nurse.」と翻訳されるのが望ましい．

本研究では文中の性別の手がかりになる文章表現，品詞などを用いて，性別の不明確な指示語の性別を推定し，機械翻訳の精度向上に寄与することを目的としている．

## ２．指示語について

指示語とは，話し手のいる場所，状況を元にもの，人や事柄を示す語である．一般的に「こそあど言葉」と言われる．「こ」で始まる「これ」などのコ系列は「近称」と呼ばれ，主に話し手と同じ場にあるもの（事象）もしくはいるものを示す．「そ」で始まるソ系列は「中称」と呼ばれ，状況はコ系列と同じく話し手と同じ場にあるもしくはいるものを示すが，コ系列よりも話し手との距離は離れる．「あ」で始まるア系列は「遠称」と呼ばれ，話し手と同じ場にないものもしくはあっても距離があるものを示す．また記憶の中のものを示すこともある[4]．これら指示語に共通する点は，ものの性別に囚われずに使用できる点である．

本研究では，指示語の中でも人間を示す「あいつ・こいつ・そいつ」や「あの・この・その＋人」を対象とする．以降これらの指示語を「人称の指示語」と呼ぶ．

## ３．性別判定に有用な語句・表現の収集

本研究では，文章中の性別を表すことができる語句や表現を用いて性別判定を行う．そのため各語句や表現とその性別値で構成される辞書を作成した．
性別値は以下のように付与した．

- 男性的である→1
- どちらかと言えば男性的である→0.75
- 中性的である→0.5
- どちらかと言えば女性的である→0.25
- 女性的である→0

性別判定は文中に現れる語や表現の性別値の平均を用いて行う．

次に，性別判定に有用な語句の品詞と表現を以下に示す．

### 名詞

名詞は，性別が顕著に現れる職業や事柄を用いる．例えば，「化粧」や「看護婦」などがそれに当てはまる．収録単語は 174 単語である．

例）「あの人は看護婦です」→女性

### 形容詞

形容詞は，性別特有の表現を用いる．例えば，「艶やか」，「男臭い」や「男勝り」が該当する．収録単語は 30 単語である．

例）「あいつ男勝りなんだ．」→女性．

### 終助詞

終助詞は，話者の性別を顕著に表す語である．「〜わ」や「〜の」は女性的であり，「〜のか」や「〜ぜ」は男性的である[1],[2],[3]．収録単語数は 25 単語である．

例）「これはすごいぜ」→話者は男

### 名詞・代名詞

名詞や代名詞に関しては，人名は性別によってつける名前が顕著に違いかつ特徴が大きい．しかし例外的に「司」や「薫」などは男女ともに用いられることもある．収録単語数は 14, 664 単語である．

例）「太郎」→男，「春子」→女

人称代名詞も男女で用いられるものが決まっており重要な品詞である．収録単語数は 24 単語である．

例）「俺」→男，「うち」→女

### 動詞

動詞を含んだものは，まず「名詞」＋「動詞」パターンは「マニキュアを塗る」というような，性別特有の動作を行う表現を示す．「名詞」＋「に・と・を」＋（名詞）＋「動詞」パターンは「彼女に告白する．」といった動詞の対象の性別が動詞の主体の逆の性別になる表現である．最後に動詞単体であるが，「嫁ぐ」などの性別特有の動詞を扱う．各辞書の収録単語数は，「名詞」＋「動詞」のパターン辞書は 77 パターン，「名詞」＋「に・と・を」＋（名詞）＋「動詞」のパターン辞書は 63 パターンである．動詞単体辞書は 2 単語である．

### 命令形

命令形は，命令形の最後につくものに性別的な特徴が見られるためそれを用いる．収録単語は 2 単語である．

例）「〜ちょうだい」→女，
「〜くれ」→男性．

その他に以下のような処理を行う．

まず名詞の職業についての表現である．これは，一般的に男性的な職業で女性を表す時は職業に「女性」をつけて表す．例，「女性消防士」．次に名前に関しての表現であるが，時に女性が代名詞に「俺」を使うケースが見られる．この場合，そのあとに女性的な特徴に名前が出現した場合，代名詞の重みを下げることによりこのケースに対応する．

## ４．性別判定手法

性別判定を，先述の辞書を用いて行う．まず対象文の形態素解析を行い，その形態素に別れた文章を探索し，辞書に登録された単語と品詞が一致すれば，その単語に付与されている性別値を指示語，もしくは話者や目的語の性別値に付与する．文章の探索が終わった時点の性別値の平均を目的のものの性別値とし，性別判定を行う．判定は性別値が 1 に近ければ男性，0 に近ければ女性とする．形態素解析は日本語解析システム「ibukiC」を用いて行う．

また判定手法には，「単独文」用と「複数文」用があり，「単独文」は指示語を含んだ文章を対象とし，文章中の指示語の性別を先述の手法で判定する．「複数文」は指示語を含んだ文(以降対象文と呼ぶ)の指示語が示すものは，対象文の前後の文章の話者や目的語になることが多い点に注目し，前後文の話者と目的語の性別値を対象文の性別値を比較し対象文の指示語の内容を推定するものである．

## ５．評価実験

今回は指示語を含む人間が性別を判定可能な文 100 文を対象に単独文の判定を行う．判定手法は，先述の単独文を対象とした手法で行う．複数文に関しては，サンプルの不足により実施しなかった．対象とする文は，アンケートや「現代日本語書き言葉均衡コーパス-KOTONOHA」から収集した．

結果は，100 文中 96 文正解であった．判定に成功した例と失敗した例を以下に示す．

### 成功例

- 「この人彼女に振られたらしいよ．」
  結果　男性
  動詞を含むパターン表現による判定．

- 「あの人は女嫌いだ．」
  結果　男性
  名詞による判定．

- こいつ，真田信繁だろ．
  結果　男性
  人名による判定．

- 「あいつったら，バカな女ですよ．」
  結果　女性
  名詞による判定．

### 失敗例

- 「あの人彼が好きらしいわよ．」
  結果　男性

本来は，「彼」を「好き」の「名詞」＋「名詞」パターンによる判定を行うことで判定できるが，「名詞」＋「名詞」パターンが未実装なため，誤判定してしまった．文中の「彼」のパラメータのみ拾ってしまっている．

- 「あいつ，私の寝込みを襲うっての．」
  結果　中性
  私(この文では女性)の寝込みを襲うという表現であいつは対照の男性が正解であるが，動詞を含むパターン表現が現状誰々「の」という表現に対応していないため誤判定してしまった．

- 「そいつは啓祐だ．」
  結果　中性
  啓祐という人名が辞書に収録されていないため誤判定をしてしまった．

## ６．終わりに

今回，指示語の機械翻訳の精度向上に有用な，性別判定のシステム加えて辞書の作成を行った．今回の実験では，96%と高い割合で正しい判定を行うことができた．これは性別判定に有用な品詞の基本的なものを抑えて辞書の作成ができた結果であると考える．しかし残りの 4%に関しては，未実装の文章表現や品詞，単語の出現により精度が下がってしまった．これらの要因は辞書の追加や，単語の追加を行うことにより解決できる問題であるので，今後は文章表現や品詞の辞書のさらなる充実と単語数を増やす必要がある．

## 参考文献


[1] 安田芳子・小川小百合・品川なぎさ「現代日本語における男女差の現れと日本語教育」日本語教育研究会論文集 7, pp.73-85 (1994)
[2] 太田淑子「談話にみる性差の様相」横浜国立大学教育紀要 32, pp.329-342 (1992)
[3] 米澤昌子「終助詞の使用頻度と性差傾向ーシナリオを資料としてー」同志社大学留学生別科紀要 5, pp. 49-60 (2005)
[4] 金井勇人「引用された談話において自信を指す指示語について」埼玉大学国際交流センター,国際交流センター紀要 3, pp.15-24 (2009)