4W-04

単語分散表現における女性標示語のステレオタイプの定量化

陳　蕾思†　　杉本　徹‡
芝浦工業大学大学院理工学研究科†　　芝浦工業大学工学部‡

1. はじめに

人工知能における差別と公平性の問題は重要な課題の一つである。Bolukbasi らの研究における、単語ベクトル計算で $\overrightarrow{man} - \overrightarrow{woman} \fallingdotseq \overrightarrow{programmer} - \overrightarrow{homemaker}$ という結果[1]は、人間社会のジェンダーバイアスが機械学習の結果の公平性に影響を与えていることを証明した。

一方で、伝統文化の影響を受け、日本語では男性視点から女性を描く言葉が多い。男性中心社会の価値基準で、女性の希少性を強調した「女性標示語」である[2]。例えば、「大学生」という言葉は性別が特定されていないのに対し、女性の大学生を表す「女子大生」という特定の単語がある。コーパスに含まれるこのような女性標示語に女性へのステレオタイプがあるか、人工知能の公平性に影響が与えているかに関する研究はまだ少ないのが現状である。

そこで、本研究では、単語分散表現を用いた日本語の女性標示語のベクトル化により、女性標示語と性格用語の関係を分析し、女性標示語に隠されている女性へのネガティブなステレオタイプを分析する。

2. 単語分散表現を用いた女性標示語のステレオタイプの定量化

2.1 単語分散表現

ニューラルネットワーク言語モデルの word2vec は、単語のベクトルを生成することで、非構造化テキストを構造化された数学モデルに変換することを可能にする。単語ベクトルにより、単語の意味だけでなく、コサイン類似度計算を通して単語間の関係も把握することができる。そこで、本研究では、女性標示語の語彙にステレオタイプがあるかどうかを明らかにすることを目的として、女性標示語とポジティブ・ネガティブな性格形容詞の関係を計算するため、word2vec による単語分散表現を用いる。

Quantifying stereotypes of gender-specific words in word embedding
†Leisi Chen ‡Toru Sugimoto
†Graduate School of Engineering and Science, Shibaura Institute of Technology
‡Faculty of Engineering, Shibaura Institute of Technology

2.2 コーパス

本研究では、日本語 Wikipedia の記事データから作成したコーパスを用いた。Wikipedia は誰でも匿名で編集ができ、大量の記事があるという特徴を持っている。そのため、社会で特定の身分の女性がどのように認識されているかを一定のレベルで反映していると考えられる。

2.3 女性標示語と性格特性用語の選定

Google 1gram, 2gram, 3gram の中で「女」という文字から始まるもので出現回数が 1 万回以上のものから「女～」「女性～」「女子～」のような 40 個の女性標示語を選び本研究の実験対象とした。また、これらの女性標示語に対応する性別中立語 40 語（例：「女子大生」に対する「大学生」）も実験対象として選定した。

本研究では女性標示語に含まれるネガティブな印象を直感的に捉えるために、性格特性を表す単語を用いる。性格特性用語リスト[3]における、男女いずれかの選択率が 70%以上の性格特性用語と、男女とも選択率が 10%以下の性格特性用語の中から、ポジティブな意味の性格用語とネガティブな意味の性格用語をそれぞれ 15 個抽出した。

選定した女性標示語と性別中立語、性格特性用語の分散表現をコーパスから word2vec を用いて生成する。コーパスの分かち書きには MeCab を使用した。例えば「女子大生」という単語が「女子大」と「生」に分割されないように、あらかじめ MeCab の単語辞書に女性標示語をすべて追加してから分かち書きを実行した。

2.4 女性標示語のステレオタイプの可視化

本研究では、Bolukbasi らの研究[1]を参考にして、性格特性用語と女性標示語の関係を 2 次元平面における単語の分布として表す。ここで平面の X 軸は性別方向への射影を表し、X 軸の正方向は女性標示語、負方向は性別中立語の方向を表すとする。そのために、ある女性標示語とそれに対応する性別中立語の分散表現の差ベクトル v_{gender} を求めて、性格特性用語の分散表現 v と v_{gender} の内積の値をこの性格特性用語の X 座標と

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

する。

平面のY軸は性別定義語かそれ以外の単語かを表す。まず性別定義語とそれ以外の単語を分類する線形分類器としてサポートベクターマシン（SVM）を訓練し、この SVM の決定関数の値を性別特性用語の Y 座標とする。

例として、女性標示語を「女子大生」、性別中立語を「大学生」としたときの各性格特性語の 2 次元平面上の位置を図 1 に示す。

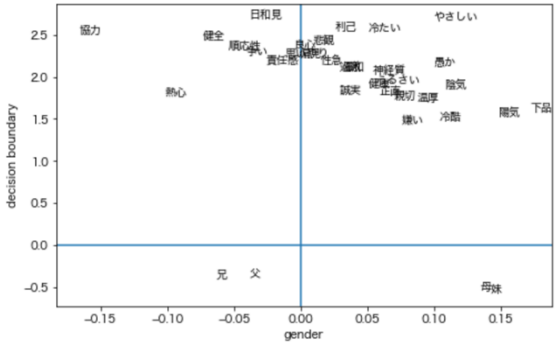

図 1.「女子大生」と性格特性語の関係

3. 女性標示語のステレオタイプの分析

女性標示語のステレオタイプを分析するために、40 個の女性標示語ひとつひとつに対して、前節で説明した可視化の結果X軸の正方向に位置する、つまり v_{gender} との内積が正の値となる性格特性用語の数をポジティブな用語とネガティブな用語に分けて数える．同様に、X 軸の負方向（性別中立語の方向）に位置する性格特性用語の数も数える。結果を表 1 と表 2 に示す。

表 1　ポジティブな性格用語の分布の比較

ポジティブな性格用語（個）	女性標示語（個）	性別中立語（個）
0〜5	23	9
6〜10	15	20
11〜15	2	11
性格用語数の平均	4.9	8.6

表 2　ネガティブな性格用語の分布の比較

ネガティブな性格用語（個）	女性標示語(個)	性別中立語(個)
0〜5	14	21
6〜10	13	9
11〜15	13	10
性格用語数の平均	7.6	6.2

表より、ポジティブな性格用語は女性標示語よりも性別中立語と関係が強いものが多く、ネガティブな性格語は女性標示語の方が関係が強いものが多いことがわかる。

また、同じペアの女性標示語と性別中立語の比較も行った。表 3 と表 4 は、それぞれポジティブ・ネガティブな性格用語のうちX軸の正方向に位置する用語数と負方向に位置する用語数を比較した結果を示している。やはり先ほどの分析結果と同様の傾向があることがわかる。

表 3　同じペアにおけるポジティブな性格用語の数の比較

ポジティブな性格用語の数	単語ペア数
X 軸正の語数＞X 軸負の語数	14
X 軸正の語数＜X 軸負の語数	26

表 4　同じペアにおけるネガティブな性格用語の数の比較

ネガティブな性格用語の数	単語ペア数
X 軸正の語数＞X 軸負の語数	24
X 軸正の語数＜X 軸負の語数	16

4. 結論

本研究では、単語分散表現におけるポジティブおよびネガティブな性格用語と女性標示語の関係を分析した。その結果、女性標示語は性別中立語よりもネガティブな性格印象と関係が深いことが分かった。つまり、女性標示語が使われた場合、特定の身分の女性グループが批判を含んだ見方で扱われる可能性が高いことが示唆された。その結果、特定の身分の女性に対するネガティブなステレオタイプが形成され、このネガティブなステレオタイプも機械学習で捉えられ、結果として女性を差別する不公平な結果になる可能性がある[4]。今後は、この影響を改善するための女性標示語を対象としたデバイアス手法を提案したい。

参考文献

[1] Bolukbasi Tolga et al. "Man is to computer programmer as woman is to homemaker? Debias in word embeddings", Advances in neural information processing systems 29 (2016): 4349-4357.
[2] 徐微潔. "日本語における女性標示語「女子〜」", 日本語と日本文学 2 (2013): 22-37.
[3] 内山理恵. "形容詞チェックリストによる青年の自己概念における性差", 日本性格心理学会発表論文集 5, 日本パーソナリティ心理学会 (1996).
[4] Leavy, Susan. "Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning", Proceedings of the 1st international workshop on gender equality in software engineering (2018).

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.