SentencePiece を用いたキャラクターの特徴語抽出

岸野望叶
茨城大学工学部情報工学科
17t4036s@vc.ibaraki.ac.jp

古宮嘉那子
茨城大学理工学研究科工学野情報科学領域
kanako.komiya.nlp@vc.ibaraki.ac.jp

1 はじめに

自然言語処理の分野では、ジェンダー、または年代などのキャラクターの特徴表現に注目した研究がなされている。しかしこれらの特徴語は、主に年代やジェンダーの分類にとどまり、個々のキャラクターの特徴的な話し方については注目されてこなかった。そこで本研究では、年代やジェンダーに限らずどのような言葉がキャラクター性を表しているかを抽出、分析した。

Mecab などの従来の形態素解析器では、キャラクターのセリフのような砕けた表現や、辞書に載っていないキャラクター固有の語尾などの分割が上手くできない。そのため、SentencePiece を使ってセリフを分割することで頻出する文字列の並びとしてセリフから特徴的な部分を切り出すことでよりキャラクター性をとらえた表現の抽出を行う手法を提案する。また、TF・IDF により、性別、年代別、キャラクター別に特徴づける部分の重み付けを行い、それぞれの特徴にあった言語表現の抽出を行った。

2 関連研究

宮崎ら [1] は、部分的に発話を書き換えることによって話者のキャラクタ性を表現する技術の実現を目指し、話者のキャラクタ性に寄与する言語表現の基礎的分析を行った。また、宮崎らのその後の研究 [2] では対話エージェントに発話にキャラクタ性を付与し、かつバリエーションを豊かにする方法として、各文節の機能部を目的のキャラクタに適した確率で適した表現に書き換え、キャラクタ性が読み手に伝わるかの実験を行った。また、音変化表現に注目し、それらを収集、分類を行った論文もある ([3])。また、奥井ら [4] はポインタ生成機構から、複数の異なるキャラクター応答を参照しながら少量データで学習する研究を行っている。

3 SentencePiece による特徴的な単語の抽出

キャラクターのセリフは、語尾にキャラクターの特徴を表す言葉がついていたり、砕けた表現だったりと、辞書に載っていない単語も多く出現する。ゆるキャラの「ふなっしー」のように、語尾に「〜なっしー」とつけるようなものや、「危ない」という単語を「危ねぇ」と変化していたりなどが例に挙げられる。そのため、辞書を使った既存の形態素解析では上手く特徴をとらえキャラクターのセリフ分割することが難しいと考えられる。そこで本稿では、SentencePiece によってキャラクターコーパスを分割し、それによって得た単語を TF・IDF によって重み付けすることで、キャラクターの特徴と言える部分を抽出する手法を提案する。

手法の評価のため、求めた TF・IDF の値を入力ベクトルとしてサポートベクターマシン（SVM）で性別、年代の分類を行い、既存の形態素解析器と比較した。

4 コーパス

インターネットから 22 作品、100 人のキャラクターのセリフを収集した。以下このセリフ集をキャラクターコーパスと呼ぶ。収集方法には、以下の三つを利用した。

1. インターネットに載っているアニメ・ゲームのセリフのまとめサイトから収集
2. アニメ動画サイトから得る
3. 漫画の電子書籍から、文章検出アプリを使ってテキスト化

キャラクター選ぶ際は、作品内でセリフの多いキャラクターを優先的に集めた。また、メインキャラクターとして出てくるキャラクターの年代は少年や青年に分類される者が多くなることが予想されたため、老人や子供に分類されるキャラクターでセリフの量が多いキャラクターは優先的に選出した。

5　キャラクターの特徴語抽出の実験

5.1　SentencePiece による特徴語抽出

SentencePiece とは、文章から直接分割方法を学習し、文章をサブワードへ分割するものである。サブワードとは、事前に単語の出現頻度を調べ、出現頻度の低い単語は文字やより小さい単語に分解するという考え方である。つまり SentencePiece では、辞書に載っている単語ではなく、出現頻度が高い文字列を１つの単語として扱う。この実験では、SentencePiece の Python モジュールを利用した。[1)]

SentencePiece による特徴語抽出の手順は以下の通りである。まず作成したキャラクターコーパス全体を入力とし SentencePiece で分割モデルを作成する。この際単語数は 1 万個に設定した。また同時に単語リストも作成した。作成した単語リストから、漢字 1 文字の単語を削除した。これは漢字 1 文字の単語はキャラクター性を表すことはないと考えたからである。次に作成した分割モデルを使ってキャラクターコーパスを分割する。作成した単語リストと分割したキャラクターコーパスを使って TF・IDF を求める。TF・IDF は次の式で求めた。

$$tf(t,d) = \frac{n_(t,d)}{\sum_{s \in d} n_(s,d)} \tag{1}$$

$tf(t,d)$: 文書 d 内のある単語 t の TF 値
$n_(t,d)$: ある単語 t の文書 d 内での出現回数
$\sum_{s \in d} n_(s,d)$: 文書 d 内のすべての単語の出現回数の和

$$idf(t) = \log \frac{N}{df(t)} \tag{2}$$

$idf(t)$: ある単語 t の IDF 値
N : 全文書数
$df(t)$: ある単語 t が出現する文書の数

$$TF・IDF = tf(t,d)・idf(t) \tag{3}$$

性別ごとの特徴語の抽出は、割合を使った方法と、TF・IDF を使ったやり方の二通りの方法で行った。割合を使う方法は、

$$\frac{n_(t,d1)}{n_(t,d2)} \tag{4}$$

$n_(t,d1)$: ある単語 t の文書 $d1$ 内での出現回数
$n_(t,d2)$: ある単語 t の文書 $d2$ 内での出現回数
の式で求めた。この際、d1 と d2 は男性または女性の全キャラクターのセリフ集とする。片方の出現回数が 0 回の単語は、0 の代わりに 0.001 で計算を行い、スムージングを行った。(4) 式は数値が大きくなるほど、その単語は片方の性別のみによく使われていることを示す。

性別の TF・IDF 値を求めるときは、どちらかの性別の全キャラクターのセリフを 1 文書とし、対する性別は各キャラクターのセリフをそれぞれ 1 文書とする。年代別の際は、各年代のキャラクターのセリフを 1 文書とした。キャラクターごとの際は、キャラクターコーパス全体を全文書、各キャラクターのセリフを 1 文書とした方法と、キャラクターコーパスを作品ごとに分け、作品ごとのセリフを全文書とし、その中での各キャラクターのセリフを 1 文書とする方法の 2 通りで行った。以降前者を全文書、後者を作品別文書と略記する。

5.2　Mecab による特徴語抽出

キャラクターコーパスを Mecab を使って分割し、分割されてできた単語を集めた Mecab 単語リストを作成した。SentencePiece と同様に、性別、年代別、キャラクター別で TF・IDF を求めた。

6　結果

得られた単語の TF・IDF 値上位 10 単語を表 1-4 に示す。また、キャラクター別の特徴は、一部を結果例として載せる。例として挙げるキャラクターは、アニメ「約束のネバーランド」からエマ、アニメ「エヴァンゲリオン」からシンジ、ゲーム「ドラゴンクエスト」からヤンガスである。

7　SVM による分類実験

求めた TF・IDF 値を入力として SVM を用いて性別、年代の分類をした。性別は、男性、女性、その他の 3 種類に分類をおこなった。アニメなどに登場するキャラクターには、性別不明のキャラクターもいるが、そのような場合はその他に分類する。年代は、1. 子供（〜12 歳）、2. 少年（12 歳〜17 歳、18 歳の学生）、3. 青年（18 歳〜29 歳）、4. 大人（30 歳〜49 歳）、5. 老人（50 歳〜）の 5 種類に分類を行った。

実験は 5 分割交差検証により行った。比較に用いる既存の辞書を使った形態素解析器には Mecab を

1）　https://github.com/google/sentencepiece/blob/master/python/README.md

SentencePiece				Mecab			
割合男性	TFIDF 男性	割合女性	TFIDF 女性	割合男性	TFIDF 男性	割合女性	TFIDF 女性
オイラ	でござる	わよね	わね	オイラ	ござる	うふふ	あたし
でがす	でござるよ	のよね	わよね	げす	オレ	ギルルン	かしら
でげす	オイラ	うふふ	のかしら	アッシ	ざる	わたくし	アルス
でがすよ	だぜ	あたし	わよ	おっちゃん	アルス	フローラ	、
ですな	でござるな	なのね	アルス	おいおい	オイラ	ギルルルル	私
でげすよ	でがす	なのかしら	のよね	ララァ	げす	カバン	しら
だよな	でげす	なさいよ	うふふ	フム	ぜ	キュルル	リュカ
オレたち	でがすよ	あたしたち	だわ	やしょ	ウイル	そうね	たし
オレは	アルス	ないわね	のね	うーむ	俺	ハラケン	ましょ
ウイルさん	ですね	なのよ	リュカ	つけろ	僕	立花	ウイル

表 1　性別特徴語

SentencePiece					Mecab				
子供	少年	青年	大人	老人	子供	少年	青年	大人	老人
オイラ	アルス	リュカ	でがす	でござるよ	オイラ	アルス	リュカ	げす	ござる
アルス	奉太郎	ウイル	でげす	でござるな	アルス	会長	オレ	アッシ	アルス
オイラたち	会長	○○	でがすよ	でござる	おっちゃん	けど	ウイル	兄貴	ござろ
のおっちゃん	ですね	流星様	でげすよ	アルスどの	ボク	折木	けど	やしょ	フム
ヤサコ	あたしたち	シンジ君	アッシ	でござるか	ヤサコ	ウイル	○	おっさん	ござっ
じっちゃん	よね	リュカさん	アッシは	ようでござるな	じんた	太郎	陛下	いやー	ござら
ねえお父さん	お兄さま	うふふ	アッシら	でござるが	けど	奉	流星	ウイル	アイラ
天沢さん	折木さん	わよね	でがすね	でござろう	電脳	オレ	うふふ	拳	メルビン
おっちゃん	めんま	オレたち	んでがす	でござるぞ	デンスケ	四宮	棕櫚	アッ	フーム
オイラも	千反田	ですわ	ですな	でござるなあ	ハラケン	千反	翡翠	ドルマゲス	許せん

表 2　年代別特徴語

SentencePiece			Mecab		
エマ	シンジ	ヤンガス	エマ	シンジ	ヤンガス
ノーマン	ミサトさん	でがす	ノーマン	僕	げす
レイ	綾波	でげす	ハア	サト	がす
コニー	僕	でがすよ	ママ	綾	兄貴
ギルダ	僕は	でげすよ	ギルダ	EVA	アッシ
ママ	父さん	兄貴	コニー	A	アッ
フィル	EVA	アッシ	出荷	アスカ	貴
ハウス	逃げちゃだめだ	アッシは	レイ	スカ	シ
ハアハア	僕が	アッシら	フィル	波	おっさん
出荷	使徒	でがすね	マン	父さん	ドルマゲス
発信器	リツコ	んでがす	ハウス	ミ	やしょ

表 3　キャラクター別特徴語全文書

エマ	シンジ	ヤンガス
ノーマン	ミサトさん	でがす
私たちの	僕	でげす
私たち	父さん	でがすよ
うん	僕は	でげすよ
それって	綾波	アッシ
んだね	んだ	アッシは
と思う	んだよ	アッシら
レイ	僕の	でがすね
え	なんだ	んでがす
よね	僕が	んでげす

表 4　SentencePiece キャラクター別特徴語作品別文書

利用した。この実験では、sklearn をライブラリとして利用した[2)]。SVM の入力は、キャラクター 1 人を 1 ベクトルとした。性別の分類では SentencePiece で得られた全単語の TF・IDF 値を入力した場合、TF・IDF 値の男女それぞれ上位 1000 単語を入力とした場合、Mecab で得られた単語を入力とした場合で比較する。年代の分類では、SentencePiece で得られた全単語の TF・IDF 値を入力した場合、TF・IDF 値の各年代それぞれ上位 500 単語を入力とした場合、Mecab で得られた単語を入力とした場合で比較を行った。

性別、年代の分類結果は以下の表 7-8 に示す。

男性	女性	その他
43	55	2

表 5　性別データ数

子供	少年	青年	大人	老人	不明
17	36	24	13	4	6

表 6　年代別データ数

8　考察

表 1 から SentencePiece を用いて抽出した特徴語のほとんどが性別に特徴的な単語であること、また、特に一人称や語尾が特徴として抽出されやすいことが分かった。また SentencePiece では、「でがす」「わよね」「のね」などのような辞書に載っていない表記も取ることができた。年代別の特徴語は、SentencePiece、Mecab どちらも固有名詞が多く、特徴的な表記を取ることができなかった。その理由としては、年代ごとに共通する表記が少なかったことが考えられる。また、大人世代と老人世代は一見固有名詞が少なく特徴語が取れているようにみえるが、これは年代の特徴というより、その年代の中のあるキャラクターの特徴である。このような結果になった要因として、大人世代と老人世代は人数が少なかったこと、その中でもキャラクターのセリフ量にばらつきがあり、セリフの多いキャラクターの特徴になってしまったことが挙げられる。キャラクター別の特徴語は、こちらも SentencePiece、Mecab 両方で固有名詞や一般名詞がよく出てきた。しかし、Mecab のほうでは「アッシ」という単語が「アッシ」だったり「シ」だったりと、文脈によって分割のされ方が変わってしまった。SentencePiece の手法では、キャラクターコーパスを作品ごとに分け、その中で TF・IDF 値をとることで、表 4 のように、より特徴とは言えない固有名詞や一般名詞を削除し、特徴語のみをとることができた。

2）　https://sklearn.org/

SentencePiece 全単語	SentencePiece 上位 1000 単語	Mecab 全単語
0.8	0.76	0.68

表 7　5 分割交差検証性別分類

SentencePiece 全単語	SentencePiece 上位 500 単語	Mecab 全単語
0.5	0.47	0.43

表 8　5 分割交差検証年代別分類

SVM の結果を見ても性別年代別共に SentencePiece で分割した全単語の TF・IDF 値を入力としたときが最もよい精度となった。

9　まとめ

本稿では、SentencePiece を用いてアニメなどのキャラクターのセリフを分割し、TF・IDF 値で重みづけをすることによって、既存の辞書を用いた形態素解析器ではとれなかったキャラクターに固有の特徴語を抽出したことを示した。また、抽出した特徴語を素性としてキャラクターの性別と年代を分類したところ、MeCab を用いてセリフを単語分割した場合に比べて、分類精度が向上したことを示した。

謝辞

本研究は、茨城大学の特色研究加速イニシアティブ個人研究支援型「自然言語処理、データマイニングに関する研究」に対する研究支援 および JSPS 科研費 17KK0002、18K11421 の助成を受けたものです。

参考文献

[1] 宮崎千明, 平野徹, 東中竜一郎, 牧野俊朗, 松尾義博, 佐藤理史. 話者のキャラクタ性に寄与する言語表現の基礎的分析. 言語処理学会 第 20 回年次大会 発表論文集, pp. 232–235, 2014.

[2] 宮崎千明, 平野徹, 東中竜一郎, 牧野俊朗, 松尾義博, 佐藤理史. 文節機能部の確率的書き換えによる言語表現のキャラクタ性変換. 人工知能学会論文誌, Vol. 31, No. 1, pp. DSF–515, 2016.

[3] 宮崎千明, 佐藤理史. 発話テキストへのキャラクタ性付与のための音変化表現の分類. 自然言語処理, Vol. 26, No. 2, pp. 407–440, 2019.

[4] 奥井颯平, 中辻真. ポインタ生成機構を用いたキャラクター応答生成の検証. 第 34 回人工知能学会全国大会論文集, pp. 1I4–GS–2–01, 2020.