

文からの感情判断における ベクトル化手法の検討

大澤 泰我[†]

マッキン ケネスジェームス[†]
[†]東京情報大学

永井 保夫[†]

研究背景

- 近年コミュニケーションロボットの需要が高まっている
- 対話システムと人との自然なコミュニケーションの実現が求められている
- 人との自然な対話の実現において、感情情報は必要不可欠

対話システムの問題点

- 対話の際に感じる心理的抵抗感や発話が機械的
- 解決する方法の一つとして、感情を考慮した対話の実現が考えられている[1][2]

[1]河野 泉,池邊 亮志,和氣 早苗,上窪 真一,岩沢 透,西村 健士.感情表現を用いた対話システムEDSの開発(1): システム概要と感情モデル.情報処理学会研究報告, “HI,ヒューマンインタフェース研究会報告 89”, pp.43-48,(2000)

[2]河野 泉,池邊 亮志,和氣 早苗,上窪 真一,岩沢 透,西村 健士.感情表現を用いた対話システムEDSの開発(2): システム概要と感情モデル.情報処理学会研究報告, “HI,ヒューマンインタフェース研究会報告 89”, pp.49-56,(2000)

研究目的

- 対話システムから受け取った入力文から感情を判断し、対話システムに感情を返す
- データセットを作り深層学習（ディープラーニング）を用いて感情判断器を作成する

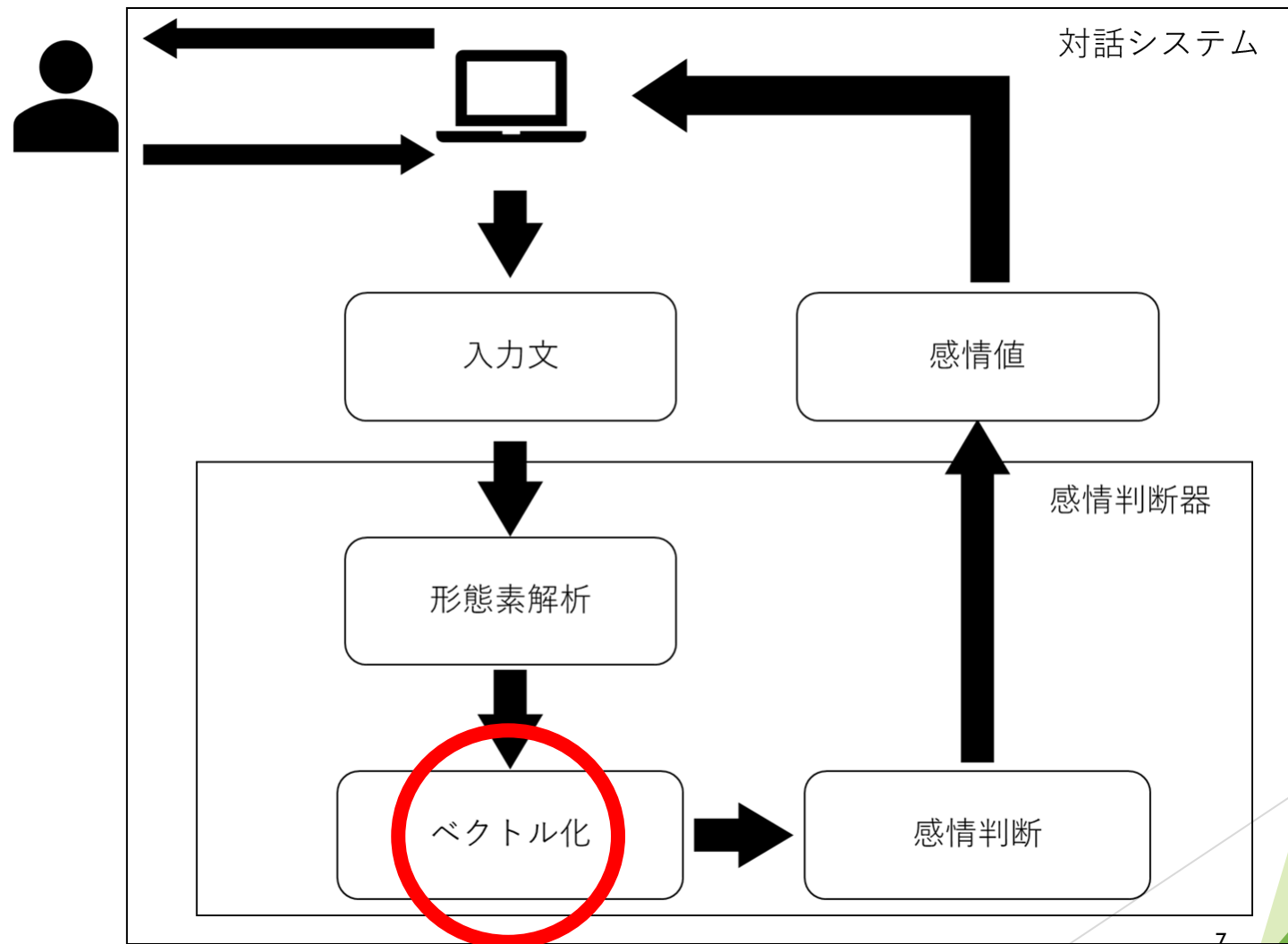
関連研究1

- ML-Ask[3]（感情的要素/感情表現分析システム）
 - 感情推定システム
- 感情表現辞書
 - 感情を表すとされる単語を定義した辞書
 - 2100語収録されている

研究の概要[4]

- データセット
 - TwitterAPIから集めたオリジナルの辞書
- 形態素解析
 - MeCab
- 学習
 - LSTMネットワーク

感情判断におけるベクトル化手法の比較



関連研究2

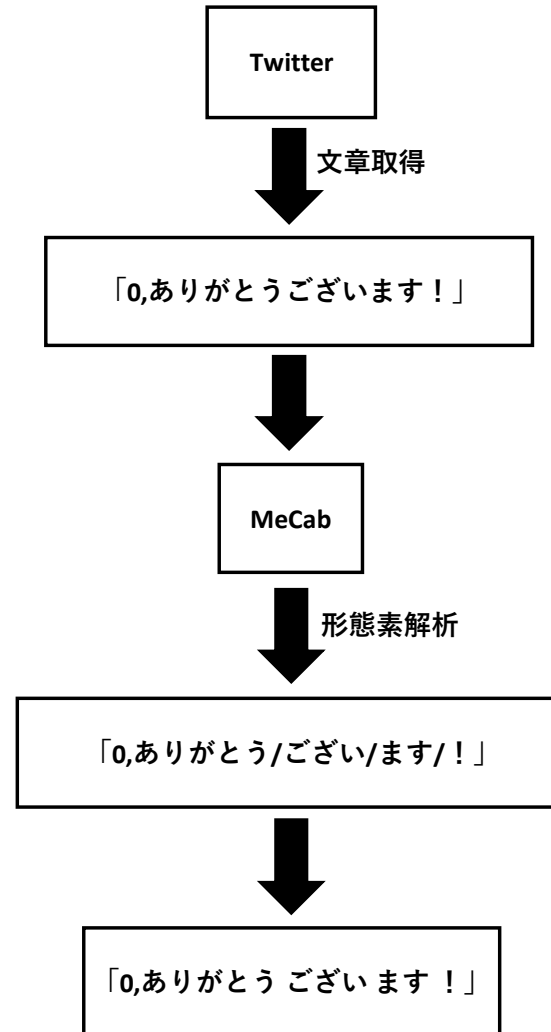
- 「コミックのセリフの感情分析[5]」
 - セリフが「ポジティブ、ネガティブ、ニュートラル」のどれに当たるのかを判断する
- ベクトル化手法
 - Word2Vec

ベクトル化手法について

- カウントベース
 - One-hot encoding
 - Bag-of-Words
 - TF-IDF
- 分散表現（推論ベース）
 - Word2Vec
 - FastText

データセットについて

例：喜びの顔文字「(^_^)」が含まれ文章を取得



データセットについて

カウントベース

- 感情表現辞書
 - 約2千件（単語）
- コミックスのセリフの感情分析
 - 約1千件（文章）
- 本研究
 - 約1万～（文章）

推論ベース

- データセット
 - 約 10万件～
 - 約100万件～

ベクトル化手法について

元データ	
テキスト1	私達はラーメンがとても大好きです。
テキスト2	私達は蕎麦がとても大好きです。

Bag-of-words									
単語	私達	は	ラーメン	蕎麦	が	とても	大好き	です	。
テキスト1	1	1	1	0	1	1	1	1	1
テキスト2	1	1	0	1	1	1	1	1	1

Word2Vec								
単語	私達	は	ラーメン	が	とても	大好き	です	。
テキスト1	-0.0149	0.0612	0.0134	0.0513	0.0483	0.0450	0.0449	0.0479
単語	私達	は	蕎麦	が	とても	大好き	です	。
テキスト2	-0.0149	0.0612	0.0344	0.0513	0.0483	0.0450	0.0449	0.0479

本研究の手法（従来手法）								
単語	私達	は	ラーメン	が	とても	大好き	です	。
テキスト1	1	2	3	4	5	6	7	8
単語	私達	は	蕎麦	が	とても	大好き	です	。
テキスト2	1	2	9	4	5	6	7	8

カウントベース手法の問題

▶ 次元数は単語数に比例する

データ										
テキスト1	あ	い	う	え	お	か	き	く	け	こ
テキスト2	さ	し	す	せ	そ	た	ち	つ	て	と
テキスト3	な	に	ぬ	ね	の	は	ひ	ふ	へ	ほ

Bag-of-words																		
単語	あ	い	...	か	き	...	さ	し	...	た	ち	...	な	に	...	は	ひ	...
テキスト1	1	1	...	1	1	...	0	0	...	0	0	...	0	0	...	0	0	...
単語	あ	い	...	か	き	...	さ	し	...	た	ち	...	な	に	...	は	ひ	...
テキスト2	0	0	...	0	0	...	1	1	...	1	1	...	0	0	...	0	0	...
単語	あ	い	...	か	き	...	さ	し	...	た	ち	...	な	に	...	は	ひ	...
テキスト3	0	0	...	0	0	...	0	0	...	0	0	...	1	1	...	1	1	...

Word2Vec										
単語	あ	い	う	え	お	か	き	く	け	こ
テキスト1	-0.0149	0.0612	0.0134	0.03439	0.0513	0.0483	0.0450	0.0449	0.0479	0.02139
単語	さ	し	す	せ	そ	た	ち	つ	て	と
テキスト2	0.0349	0.0121	0.0633	0.01432	0.0431	0.0556	0.0410	0.0358	0.0417	0.02918
単語	な	に	ぬ	ね	の	は	ひ	ふ	へ	ほ
テキスト3	0.0549	0.0312	0.0234	0.06439	0.0113	0.0683	0.0950	0.0249	0.0889	0.03139

本研究の手法（従来手法）										
単語	あ	い	う	え	お	か	き	く	け	こ
テキスト1	1	2	3	4	5	6	7	8	9	10
単語	さ	し	す	せ	そ	た	ち	つ	て	と
テキスト2	11	12	13	14	15	16	17	18	19	20
単語	な	に	ぬ	ね	の	は	ひ	ふ	へ	ほ
テキスト3	21	22	23	24	25	26	27	28	29	30

提案

➤ カウントベースのベクトル化手法の提案

- 単語の出現順序を維持
- 次元数を増やさない

- 「ある文書の中で何度も出現する単語は重要度が高いが、多くの文書に共通して出現する単語は重要度が低い」 [6]

提案

- ベクトル中の重要度の低い単語を切り捨てる

元データ					
テキスト1	今日	は	、	良い	天気
テキスト2	今日	は	、	きりさめ	だ
テキスト3	明日	が	、	雨	らしい
テキスト4	明日	が	、	晴れ	だって

ベクトル化したデータ					
テキスト1	1	2	3	4	5
テキスト2	1	2	3	6	7
テキスト3	8	9	3	10	11
テキスト4	8	9	3	12	13

提案手法適用前					
テキスト1	1	2	3	4	5
テキスト2	1	2	3	6	7
テキスト3	8	9	3	10	11
テキスト4	8	9	3	12	13

提案手法適用後					
テキスト1	1	2		4	5
テキスト2	1	2		6	7
テキスト3	8	9		10	11
テキスト4	8	9		12	13

ベクトル対応表	
今日	1
は	2
、	3
良い	4
天気	5
きりさめ	6
だ	7
明日	8
が	9
雨	10
らしい	11
晴れ	12
だって	13

提案手法を適応した結果

▶ 件数

	語数
総単語数	628,296
切り捨てた単語	76,908

12.2%

データセットの前処理手法による 正解率の比較

データセット	ベクトル化手法	正解率
前処理あり	従来手法	62.5%
	提案手法	65.4%
前処理なし	従来手法	60.5%
	提案手法	62.4%

分散表現を用いた手法との比較

- カウントベースと分散表現の次元数は同じ100次元

ベクトル化手法	正解率
カウントベース (提案手法)	65.4%
分散表現 (Word2Vec)	36.7%

結論

- 本研究の感情判断器には、カウントベースのベクトル化手法の方が、高い正解率が得られた
- データセットに前処理を適用した手法は、本研究で導入していた従来手法よりも約3%の正解率の向上が確認された

今後の予定

- 学習手法の再検討
 - 複数の機械学習手法による比較を行う
- Twitterからのデータ収集における感情辞書の有効性の検証

参考文献

- [1]河野 泉,池邊 亮志,和氣 早苗,上窪 真一,岩沢 透,西村 健士.感情表現を用いた対話システムEDSの開発(1):システム概要と感情モデル.情報処理学会研究報告,“HI,ヒューマンインタフェース研究会報告 89”, pp.43-48,(2000) .
- [2]河野 泉,池邊 亮志,和氣 早苗,上窪 真一,岩沢 透,西村 健士.感情表現を用いた対話システムEDSの開発(2):システム概要と感情モデル.情報処理学会研究報告,“HI,ヒューマンインタフェース研究会報告 89”, pp.49-56,(2000) .
- [3] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka, Kenji Araki. A system for affect analysis of utterances in Japanese supported with web mining. 日本知能情報ファジィ学会誌, vol.21, No.2, pp.194-213 (2009)
- [4] 大澤泰我, ケネス マッキン, 永井保夫. 文脈から感情を分析するための感情判断器の検討. 第 19 回情報科学技術フォーラム. F-026, (2020).
- [5] 盛舒峰, 渡辺裕. コミックのセリフの感情分析. 電子通信学会基礎・境界ソサイエティ/NOLTAソサイエティ大会, A-10-18 (2018).
- [6] 木村美紀. TF-IDFを用いた文書分類の試み. 文学研究論集第48号. 1-17, (2017).
- [7] 大澤泰我, ケネス マッキン, 永井保夫. 対話システムのための感情判断器の検討. 情報処理学会第82回全国大会. 1S-06, (2020).
- [8] Word2Vec. <https://code.google.com/archive/p/Word2Vec/>.

