5Q-01

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†　永井　保夫†
東京情報大学†

1. はじめに

コミュニケーションロボットの需要が高まる中,人と対話システムの自然なコミュニケーションの実現が求められている.人と対話システムの自然な対話の実現において,感情情報は必要不可欠である.本研究では,深層学習を利用して文章から感情を判断し,対話システムが感情的な会話を返すことで,人と対話システムの自然なコミュニケーションの実現を目的としている[7][8].ここでは,主に対話システムのための感情判断器を検討している.盛ら[1]は,感情分析を複数のニューラルネットワークモデルで学習をおこない,その精度の比較を行っている.ベクトル化には,Word2Vec によるベクトル化を採用している.中井ら[2]は,感情分析にナイーブベイズ SVM とロジスティック回帰を用いて学習をおこない,その精度の比較を行っている.その際のベクトル化には Bag-of-words と TF-IDF の両方を用いている.

本研究の感情判断手法では,Twitter 上に投稿された文章を学習し,構築したモデルを用いて感情判断を行う.顔文字付きの文を集め,データセットを作成し,顔文字に対応する感情情報を正解データとして学習させる.データセットは 58,619 件の文章から構成され,前処理手法として品詞の一部を省いている.感情判断手法は,Plutchik の感情の輪[3]に基づき,8 つの感情「喜び,信頼,恐れ,驚き,悲しみ,嫌悪,怒り,期待」に分類を行った.本研究では,学習モデルに LSTM ネットワークを用い,対話システムから入力された会話文に対して,会話文から感情を判断し,正しい感情値を返すことを示した.

本論文では,学習を利用して感情判断を行うのに必要とされるデータセットのベクトル化手法の比較を行う.ここでは,分散表現を用いないカウントベースのベクトル化手法と分散表現を用いたベクトル化手法を用いて学習精度の比較を行う.

2. 感情判断におけるベクトル化手法と問題点

2.1 ベクトル化手法の特徴

自然言語を機械学習で扱う場合,何らかの方法で文章をベクトルデータに変換するベクトル化手法が必要となる.ベクトル化手法は,One-hot encoding や Bag-of-Words,TF-IDF などに代表されるカウントベース手法と Word2Vec に代表される分散表現を生成する手法に大別できる.

One-hot encoding は,データ中に存在する単語（語彙）数の数だけ次元を用意して,各行に含まれている単語（語彙）に対応する次元を 1 に,それ以外を 0 にする手法である.

Bag-of-words は,古典的な非常にシンプルなモデルで出現単語に ID を付け文書の各単語の有無だけを集計する.

Word2Vec は,テキストコーパスを入力とし,出力として単語ベクトルを生成する[4].Word2Vec では,単語ベクトルを意味的な関係を表現することができ,自然言語処理の多くのシステムの改善が期待されている.

従来,我々は研究において,与えられた単語（語彙）に対して順番に数字を割り当てる最もシンプルな手法を用いてきた[7].

表 1 は,Bag-of-Words,Word2Vec,従来の研究で用いてきたベクトル化手法の適用例を示している。

表 1　ベクトル化手法の適用例

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

2.2 問題点

Word2Vec は,あらかじめ単語を学習しておく必要があり,未知語が入力された際にベクトルに変換できないという問題がある.本研究では,未知語を 0(ないもの)とすることで対応した.

Bag-of-words や TF-IDF などのカウントベースの手法は,単語（語彙）数が大きくなるほど次元数が大きくなってしまう.例えば,10 個の単語で構成されたデータセットが 10 件あった場合,Word2Vec や従来手法[7]では 1 データセット当たりの次元数は 10 次元だが,Bag-of-words では 1 データセット当たりの次元数は 100 次元となってしまう.また,Bag-of-words に用いられる One-hot encoding は単語の出現順序を考慮しないという問題もある.

3. 提案手法

大量の文章を取り扱うことは単語（語彙）数が増えることに繋がる.Bag-of-words では,単語の出現順序を考慮せず次元数が膨大になってしまう.本研究で取り扱うデータセットの文章数は 58,619 件であり,単語（語彙）数は 628,296 語となっている.ベクトル化に際して文章数の増加,すなわち,単語（語彙）数が増えた場合に,次元数を増やさない手法が必要となる.

そのため,本提案では,単語の出現順序を維持したまま,次元数を増やさないカウントベースの手法を提案する.

提案手法は,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

Text Vectorization for Emotion Classification.
Taiga Ohsawa†,Kenneth J Mackin†,Yasuo Nagai†
†Tokyo University of Information Sciences

が低い[5]という考え方に基づいたカウントベースの手法である.

表 2　提案手法の適用例

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

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

提案手法の適用により,データセット全体で多く存在するベクトル（次元）の切り捨てをおこなう.表 2 の左下の表中の 5 次元のベクトル（提案手法適用前）に対して赤色の「3」が切り捨てられ,右下の表中の 4 次元のベクトル（提案手法適用後）が作成されている.

データセットの総単語数 628,296 語のうち提案手法が適用できたのは,12.2%であった.切り捨てたベクトルはデータセットの中に 76,908 語存在し,そのすべてはデータセットから切り捨てられている.

提案手法では,ベクトル中の重要度の低い数を切り捨てることで,次元数を減らしたベクトルを作成している.

4. 評価実験

本研究では,LSTM ネットワークを用いて学習モデルを生成している.学習モデルは同じものを使用して,各ベクトル化手法を適用した際の学習精度の比較を行った.評価指標には正解率を用いた.

4.1 前処理手法による正解率の比較

本研究のデータセットに対する前処理手法では,表 2 の「、」などの句読点や「は」などの接続助詞をデータセットから省く品詞加工を行っている[8].

ここでは,前処理手法による正解率を比較するために,品詞加工を行った（前処理ありの）データセットと品詞加工を行なわなかった（前処理なしの）データセットの正解率の比較を行った.表 3 は前処理ありのデータセットと前処理なしのデータセットよる正解率を示している.

表 3　前処理手法の適用による正解率

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%

データセットに対して前処理を適用した提案手法では,従来手法の 62.5%に対して 65.4%と約 3%の正解率の向上が見られた.

また,データセットに対して前処理手法を適用しなかった場合,提案手法の正解率は 62.4%であり,データセットの前処理手法を適用した従来手法における正解率の 62.5%とほぼ同等の正解率であった.

4.2 分散表現を用いた手法との比較

一般的に多くのデータセットを扱う際に,分散表現を用いた手法が使われているため,提案手法と分散表現を用いた手法における正解率の比較を行った.表 4 は提案手法（カウントベース）と Word2Vec（分散表現）の正解率を示している.

本研究における深層学習を用いた感情判断においては,分散表現を用いた手法よりも,カウントベースを適用した手法の方が,正解率が高いことがわかった.

表 4　分散表現との比較

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%

5. 結論

本研究は,人との自然なコミュニケーションを行う対話システムに必要となる感情判断器の開発を目指している.

本論文では,文の感情判断におけるベクトル化手法の比較を行った.本研究における学習を用いた感情判断手法においては,分散表現を用いた手法よりも,提案した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.また,提案手法はデータセットの前処理手法を適用しなくても,前処理手法を適用した従来手法と同等の正解率を持つため,データセットの品詞加工なしでもある程度の正解率を得ることが可能と考える.

今後は,感情判断器の正解率を向上するため,ベクトル化手法だけでなく,学習手法を再検討し,複数の学習手法による比較を行う予定である.

参考文献

[1] 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).

[2] 中井諒馬,山田誠.感情分析における機械学習手法の比較検討.京都大学学術情報リポジトリ.ELCAS Journal,(2020).

[3] Robert Plutchik. “The nature of emotions”. American Scientist.Vol.89. Iss.4, pp. 344-350,(2001).

[4] Tomas Mikolov, Kai Chen,Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations,(2013).

[5] 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).

[6] 五十嵐光秋,坂地泰紀,和泉潔,島田尚,須田真太郎.極性を考慮したリスク発見に向けた因果関係ネットワークの構築.第 33 回人工知能学会全国大会.203-J-13-04,(2019).

[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] 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).

[9] Word2Vec. https://code.google.com/archive/p/Word2Vec/.