表記ゆれに着目したテキスト感情分類システムの改善

2019SE046 大西翼
指導教員：張漢明

1　はじめに

文章を読み上げる感情を指定することで感情を含んだ音声を合成する技術がある．文章に感情を付加する技術の問題点として，感情の指定を人手で行う必要性があげられる．この問題に対し坂ら [1] は，テキスト情報から感情値を判断し，音声合成の発話スタイルを自動で切り替えるシステムを提案した．このシステムは，文書分類の手法を用い学習したテキストを基に，入力テキストの感情を判断することができる．テキストのみを入力として感情を含んだ音声の出力を可能とした．

このシステムにはテキスト情報の読み取りにおいて改善すべき点がある．その一つに表記ゆれがある．表記揺れとは同じ意味の語でも異なる表記があることを指し，学習データセットに出てきた語が，感情を分類したいテキスト上で異なる表記をされた際に，学習した結果が反映されないといった問題が生じる．

本研究の目的は，表記ゆれ問題に着目した，感情分類システムの改善である．本研究では，坂ら [1] の提案したシステムの改善を行い，感情分類の精度向上を図る．

研究方針として，システムの感情分類部において，表記ゆれ問題を改善するために，形態素解析部の変更を行い，その結果を考察する．

2　先行研究・関連技術

坂ら [1] は，入力したテキストを，句読点分割などの前処理をしたのち，文書分類することで感情を割り当てた．以下システムの解説を行う．ただし，音声合成ソフトウェアについては本研究では扱わないため説明を省力する．

2.1　分類の前処理

入力したテキストに対し，文ごとに感情をラベル付けするため，句読点分割を行う．以下，分割後の短文をフレーズとして表現する．次に，形態素解析をフレーズごとに行う．形態素解析は，テキストを意味のある最小単位である形態素にまで分割し，それらがどの品詞であるのかを分類する．これは日本語の文について分かち書き，品詞厳選をする目的で行う．分かち書きとは，単語と単語の間に空間を入れることである．品詞厳選は，感情に関して関連の低い語句を取り除く目的で行う．

2.2　文書分類

感情をカテゴリとして文書分類を行い，形態素解析後のフレーズ一つ一つに対し，感情のラベルを付ける．分類は，大きく分けて「文書のベクトル化」「分類」という２つのフェーズがあり，以下の２つの手法を用いた．

- Bag of Words (BoW)
- ナイーブベイズ分類

BoW は，カウントベースで文書のベクトル化を行う．コーパスを基に単語の出現を記録する．コーパスとはテキストを大規模に集め，データベース化したものである．

ナイーブベイズ分類とは，過去の事例をもとに未知の文書があらかじめ与えられているどのカテゴリに属するかを決定する，教師あり分類手法である．坂ら [1] の研究，山本ら [2] の研究では感情コーパスを構築する目的も兼ねて，感情分類手法としてナイーブベイズ分類を提案している．

2.3　感情コーパス

感情コーパスとは感情ごとにテキストデータを集約したデータベースである．ナイーブベイズ分類はカテゴリごとに既知の文書が必要である．本研究では４感情「喜び」「平常」「怒り」「悲しみ」に分類するため，それぞれの文書を集約したデータセットを用いる．「日本語話し言葉コーパス」[3] から無作為に抽出したものが含まれ，これらは感情が定義されていないため，「日本語感情表現辞書」[4] をもとに感情のラベル付けを行っている．

3　システムの設計・問題点

システム作成には２つの過程がある．感情コーパス作成と感情判定部作成である．前者はナイーブベイズ分類器の前処理となる学習をし，後者は任意の入力テキストに対し学習結果をもとに感情分類を行うシステムを構築する．

3.1　感情コーパス作成

学習は，感情ごとのデータセットを形態素解析し，感情分類に有効だとみられる品詞をすべて集約する．単語，文書の集約は感情ごとに行い，それぞれベクトル化することでナイーブベイズ分類の既知の文書とする．また感情を問わず，データセットに出てきた単語を全て集約したものを単語辞書コーパスとして作成する．これは BoW を作成するためのものである．

3.2　感情判定部

入力テキストをフレーズ分割し，フレーズごとに形態素解析を行う過程である．フレーズに出てきた，感情に関する品詞を抜き出し BoW を行った後，ナイーブベイズ分類を行う．

3.3　形態素解析への依存

3.1 ，3.2 節で述べた 2 つの過程では，どちらも形態素解析を行った後，単語の集約，ベクトル化をする．その過程を**図 1** に示した．形態素解析は単語辞書の作成，感情

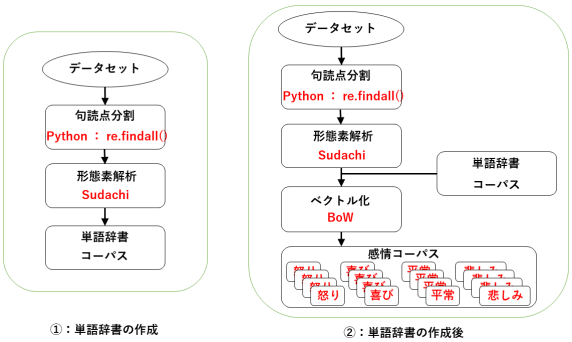

図 1　感情分類コーパス作成部

	janome	MeCab＋NEologd辞書	sudachi（＋形状詞）
名詞，動詞，形容詞	約0.712	約0.720	約0.705
名詞，動詞，形容詞，助動詞	約0.817	約0.821	約0.810
名詞，動詞，形容詞，助動詞，副詞	約0.824	約0.826	約0.827
名詞，動詞，形容詞，助動詞，副詞，接続詞	約0.805	約0.811	約0.821
全ての品詞	約0.837	約0.839	約0.842

図 2　システム変更と品詞による精度の違い

コーパスの作成にかかわり，入力テキストの解析にも使用する．このシステムにおける依存度は大きい．形態素解析の結果によって，単語の分かち書き，品詞，表記ゆれ問題などがすべて変わるためである．

3.4　表記ゆれ

表記ゆれとは，同じ意味の言葉であっても複数の表記があることを指す．これにより，同じ意味の語でも分類結果が異なるものになってしまう．複合語を分割をどの程度まで一語とみなすかという問題もこれに該当する．

3.5　sudachi の導入

本研究では，表記ゆれ問題解決のため，形態素解析部の変更をする．坂ら [1] の用いた形態素解析器「janome」は表記ゆれ問題が生じる．そこで，表記ゆれ問題に対処した形態素解析器「sudachi」を用いる．「sudachi」はワークス徳島人工知能 NLP 研究所が開発した形態素解析器である [5]．表記ゆれを同一視するための正規化が整備された辞書により表記ゆれ問題の改善を図ることが可能である．

加えて，3.3 節で述べたようにシステムは形態素解析部に大きく依存しており，変更により分類精度を向上させることができるとも考えた．

4　結果と考察

4.1　形態素解析器の変更による学習結果への影響

形態素解析部を「janome」から「sudachi」に変更したシステムを作成した．コーパスに登録された単語数を解析器変更の前後で比較すると減少がみられた．このことから，データセット内で表記のゆれていた語をより正規化して，上手くまとめられたのだと推測でき，表記ゆれ問題を改善をできたといえる．

4.2　精度検証

層化 K 分割交差検証を K=10 として行った．感情分類システムの精度が最も上層する品詞厳選の組み合わせを探るとともに，形態素解析部の変更がどんな影響を与えるか調べるためにそれぞれ比較を行う．その結果を**図 2** に示す．すべての品詞を対象に学習，分類するモデルは形態素解析器「sudachi」を用いたシステムの精度が高いという結果を得られた．また，辞書の変更によってもわずかな精度向上を得られた．

4.3　表記ゆれ対策の有効性

形態素解析部の変更により，品詞ごとの登録した単語数は変化が見られた．「sudachi」のシステムでは，「janome」のシステムと比較して，副詞では 63 語，接続詞では 37 語登録数が減少した．**図 2** に示した通り，それらの品詞を学習，分類に使用したモデルでは「sudachi」の方が精度を大きく改善する結果となった．これらは語句の表記ゆれ対策，すなわちまとめ上げにより，学習データ内に登場する語が少なくなるとともに，登場頻度が上昇することにより，学習効率が上昇したためだと推定できる．

5　おわりに

本研究では，坂ら [1] の提案したシステムに対し，表記ゆれ問題の改善を図った．改善は見られたが，感情分類の正確性はまだ足りないため，さらなる改善が必要である．学習データセットの調整や分類システムの変更を施し，感情を正しく分類できるよう精度向上を目指したい．

参考文献

[1] 坂瞭太郎，太田侑雅：音声合成の発話スタイルを切り替えるシステムの設計, 南山大学理工学部ソフトウェア工学科卒業論文 (2022)

[2] 山本麻由，土屋誠司，黒岩眞吾，任福継：感情コーパス構築のための文中の語に基づく感情分類手法，社会法人情報処理学会, (2007)

[3] コーパス検索アプリケーション「中納言」, 国立国語研究所：https://chunagon.ninjal.ac.jp

[4] 山本和英：SNOW D18 日本語感情表現辞書，長岡技術科学大学 電気電子情報工学専攻，(2018)

[5] 坂本美保, 川原典子, 久本空海, 一馬, 内田 佳孝（株式会社ワークスアプリケーションズ ワークス徳島人工知能 NLP 研究所）. 形態素解析器『sudachi』のための大規模辞書開発. 言語資源活用ワークショップ (2018)