見出し語の違いが感性情報分類に与える影響についての考察

情報科学科　宮川 知也　　指導教員：指導教員 山村 毅

1　はじめに

近年，単語の意味を低次元の密なベクトル (word embedding) で分散表現する研究に注目が集まっている．とりわけ，Mikolov ら [1] が提案した word embedding は，加法構成性 (“king”−“man”＋“woman”≈“queen” のように単語のベクトル間で意味に関する加減算が可能であること) を持つことで特に注目を集めている．本研究では，単語の word embedding を，その言語学的性質を考慮してより正確に求めることが，そうでない場合と比べてどの程度自然言語処理システムの性能に影響を及ぼすかを調べることを目的とする．具体的には，見出し語の違い (word embedding を求める単語単位の違い) が感性情報分類の精度にどのような影響を及ぼすかを調べる．

2　感性情報分類

我々は，新聞社説を対象に，感性情報を抽出・分類する研究を行なっている [2]．そこでは，新聞社説に含まれる感性情報を「メッセージ」もしくは「意見情報」と呼び，Bag-of-words で特徴表現された文を肯定，否定，疑問，助言，否定，メッセージ無しの 6 つのメッセージだと想定し, そのうちの一つに分類する問題として捉えている．新聞社説には毎日新聞の 2006 年の社説記事 9492 文を用いており、実験にはこれにあらかじめ人手で正解情報が付与されているものを用いている. 内訳は以下の通りである.

表 1: データの内訳

肯定	期待	疑問	助言	否定	無し	合計
308	417	357	1024	1138	6248	9492

3　見出し語

Mikolov らが提案した word embedding の獲得手法には word2vec がある. これはテキストファイルを学習データの文脈の前後関係して word embedding を求める方法である. この際，テキストファイル中の空白で区切られた部分を word embedding の対象になる「単語」と見なしている．本研究では，この word embedding の対象になる単語のことを「見出し語」と呼ぶことにする. 日本語を対象に word embedding を求める場合，文章を形態素解析して見出し語を空白で区切る必要があるが，この際，単語の屈折 (“書く” と “書か” を区別するか) や品詞 (格助詞の “が” と終助詞の “が” を区別するか) を考慮するかによって見出し語の求め方が異なる.

4　実験方法

本研究では感性情報分類としてデータ表 1 を 6 つのメッセージに分類することを考える. すなわち，文中の各単語の word embedding を求め，それらの算術平均

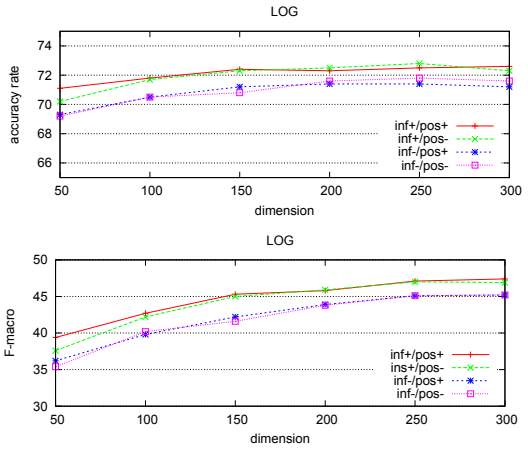

図 1: 分類結果, 正解率 (上),F 値マクロ平均 (下)

をその文の特徴表現とし，これを分類器へ入力して感性情報分類を行い，分類正解率と F 値マクロ平均を計算する．このとき、word embedding を求める際の見出し語の求め方の違い (単語の屈折の有無、品詞情報の有無) によって分類性能がどのように異なるのかを調べる. word embedding は，毎日新聞 2005〜2010 年の社説以外の記事を学習データとして用いて，word2vec[1]で学習させた．また，形態素解析には MeCab[2]を，分類器の実装には WEKA[3]をそれぞれ用いた．これら実験条件の下で，分類正解率と F 値マクロ平均の計算には 10 分割交差検定を用いた.

5　実験結果とまとめ

分類器に対数線形モデル (LOG) を用いた場合の結果を図 1 に示す (図中，inf，pos は見出し語を求める際の屈折，品詞情報の考慮を意味し，+，−はその有無を表す)．屈折がある場合の方 (赤線および緑線) が，分類正解率も F 値マクロ平均も高いことが分かる．また，品詞情報は分類性能の向上にほとんど寄与していないことが分かる (赤線 vs 緑線，青線 vs 紫線)．結論として感性情報分類には見出し語を考慮すると分類性能がよくなり，品詞情報の有無ではほとんど性能が向上しないといえる.

参考文献

[1] T. Mikolov et al.: “Efficient Estimation of Word Representions in Vector Space”, International Conference on Learning Representions Work shop, 2013

[2] 宮川知也, 藤巻直也, 山村毅: “意見情報の推定における特徴選択についての一考察”, 電気・電子・情報関係学会東海支部連合大会, D3–1，2016

[1] https://code.google.com/archive/p/word2vec/
[2] http://taku910.github.io/mecab/
[3] http://www.cs.waikato.ac.nz/ml/weka/