感性情報分類における見出し語の違いの影響について

宮川 知也　　山村 毅
愛知県立大学 情報科学部
{is121081@cis, yamamura@ist}.aichi-pu.ac.jp

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．同様の加法構成性を持つ word embegging を学習するモデルとしては，Pennington らによる GloVe[2] がある．

こういった word embedding は単語の意味を比較的によく表しているのではないかとの観察から，これを自然言語処理システムに取り入れることでその性能向上を図ろうとする研究がいくつか行われているが [3, 4, 5, 6]，その性能は期待したほど向上していない．これは，word embedding の自然言語処理システムでの利用の仕方に問題があると考えることもできるが，一方で，word embedding の正確性に問題 (意味を正確に表していない可能性．例えば Mikolov らが公開している word2vec のデモでは 5 月を表す May と助動詞の may を区別していない) があると考えることもできるであろう

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 新聞社説の感性情報分類

今日絶え間なく大量のデータが発信される情報社会では，処理をコンピュータ等によって自動化し効率化を測る必要性に迫られている．情報の中には Twitter などの SNS への投稿や通販サイトの商品レビュー，映画の感想といった感性情報もあり，これらを自動で処理して，例えば，意見や評価に関する情報を取り出すことができれば，極めて有用であると考えられる．

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 [7, 8, 9]．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 6 種類のメッセージ [1] のうちの一つに分類する問題として捉えている．このうち文献 [9] では，文に縮約処理 (主動詞とそれにかかる文節の主要部を残し他を削除する処理) を施して感性情報の判定に不要な部分を取り除き，それを Bag-of-words で特徴表現してナイーブベイズ分類器で分類することで，毎日新聞の 2006 年の社説記事 9492 文を対象にした評価実験で，分類正解率 79.0%，F 値マクロ平均 63.7%を実現している．

3 word2vec と見出し語

3.1 word2vec

word2vec は Mikolov ら [1] によって提案された word embedding の獲得手法である．入力層，中間層，出力層の 3 層からなるニューラルネットワークを用いて，単語とその周辺 k 単語の関係を学習するもので，CBOW モデルと Skip-gram モデルの二つがある．CBOW は，単語 w_t の周辺 k 単語 $(w_{t-k}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+k})$ からその単語 w_t を予測するものであり，Skip-gram は，CBOW とは反対に単語 w_t からその周辺 k 単語 $w_{t-k}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+k}$ を予測するものである．

いずれのモデルでも，学習後のニューラルネットワークの入力層と中間層の重みが word embedding に対応する．

[1] 肯定，否定，疑問，助言，否定，メッセージ無し

3.2 見出し語

word2vec では，テキストファイルを学習データとして用いて word embedding を学習するが，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，概ね次の 4 通りの見出し語の求め方がある．

- 屈折を考慮する+品詞情報あり
- 屈折を考慮する+品詞情報なし
- 屈折を考慮しない+品詞情報あり
- 屈折を考慮しない+品詞情報なし

4 見出し語の違いによる分類性能の比較

4.1 実験方法

新聞社説の文を対象に感性情報分類を行う．すなわち，文中の各単語の word embedding を求め，それらの算術平均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，3.2 で述べた，word embedding を求める際の見出し語の求め方の違いによって分類性能がどのように異なるのかを比較する．また，word embedding の次元数の違い (50〜300 次元まで 50 刻み)，および分類器の違い (対数線形モデル，ニューラルネットワーク，サポートベクトルマシン) による性能の違いについても調べる．

実験で使用する新聞社説および感性情報の分類カテゴリは，先行研究 [9] と同じものを用いた (毎日新聞の 2006 年の社説記事 9492 文を用いて,「肯定」「否定」「疑問」「助言」「否定」「(メッセージ) 無し」のうちの一つに分類する)．表 1 にその内訳を示す．

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，3.1 で述べた word2vec[2] で学習させた．また，形態素解析には MeCab[3] を，分類器の実装には WEKA[4] をそれぞれ用いた．

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

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．

4.2 実験結果

word embedding の見出し語の求め方の違いによって分類正解率，F 値マクロ平均がどう異なるかを，分類器ごとに word embedding の次元数を変えながら求めた結果を，それぞれ図 1，2 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．また，LOG は対数線形モデル，NN はニューラルネットワーク，SMO はサポートベクトルマシンを表す)．

これらの図から，どの分類器でも，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)．最大分類正解率は,「屈折あり+品詞情報あり」の見出し語から 150 次元の word embedding を求めて対数線形モデル (LOG) で分類した場合の 73.5%であった．また，最大 F 値マクロ平均は,「屈折あり+品詞情報あり」，300 次元，サポートベクトルマシン (SMO) の場合の 47.4%であった．

対数線形モデル (LOG) では次元数の上昇に伴い正解率と F 値マクロ平均が上昇している．特に F 値マクロ平均は大幅に増加している．ニューラルネットワーク (NN) も次元数の上昇とともに F 値マクロ平均は上昇するが，分類正解率は反対に低下している．サポートベクトルマシン (SMO) は，150 次元で最大正解率，100 次元で最大 F 値マクロ平均となるが，その後は次元数の増加とともに値が低下する．

4.3 考察

実験結果から，全体として，屈折を考慮した見出し語の方が分類性能が良いことが分かった．これは，屈折した語形を用いることでその単語の文脈をより正確に推定することができるため，より正確に word

[2] https://code.google.com/archive/p/word2vec/

[3] http://taku910.github.io/mecab/

[4] http://www.cs.waikato.ac.nz/ml/weka/

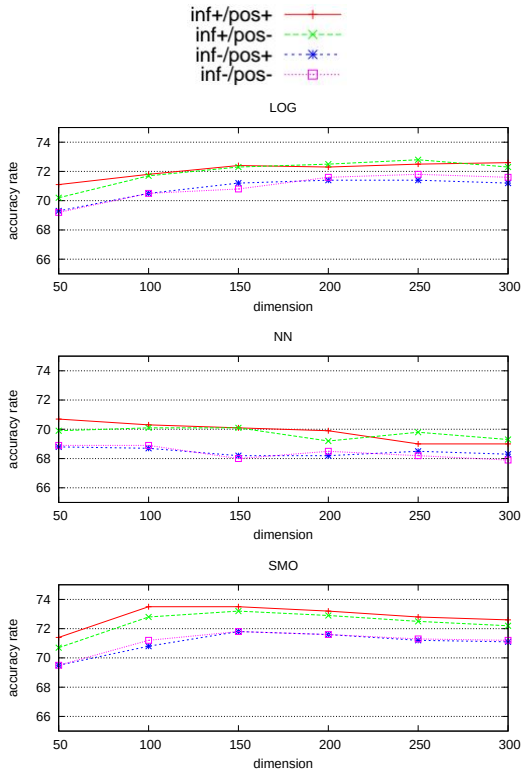

図 1: 見出し語の違いよる分類正解率の比較

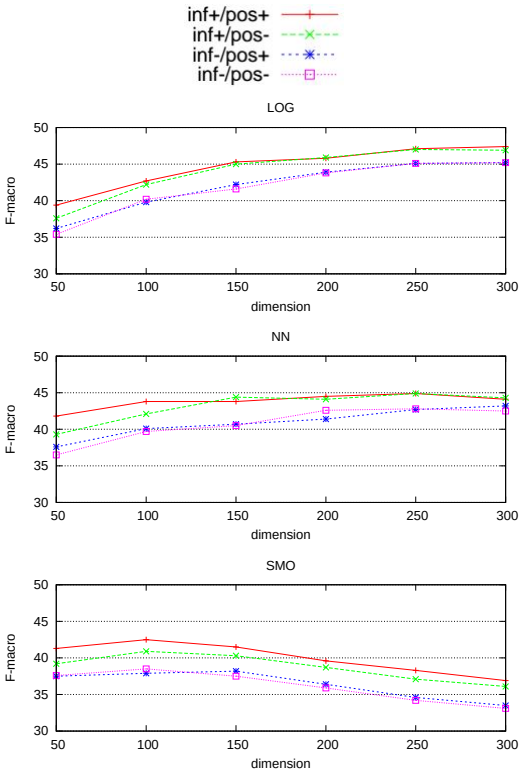

図 2: 見出し語の違いによる F 値マクロ平均の比較

embedding を求めることができるためではないかと考えられる．

一方，品詞情報に関しては，同様の理由から分類性能が向上することが期待されたが，実際には向上しない場合もあった．これは，日本語の場合，多品詞語が多くないことがその大きな要因ではないかと思われる (例えば，“書く” は通常，動詞でのみ出現するため，これを “書く (動詞)” としても，文脈推定の向上にはほとんど寄与しないと考えられる)．

図 3，4 は，図 1，2 の結果を，分類器間で比較したものである．どの見出し語の場合においても，概して対数線形モデル (LOG，赤線) の結果が一番よく，その性能は次元数を増やすとともに増加しているが，ニューラルネットワーク (NN，緑線) は性能が悪いことが分かる．また，サポートベクトルマシン (SMO，青線) は F 値マクロ平均が次元数の増加とともに減少していることが分かる．これらの理由については，今後 (データを増やすなどして) さらに詳しく調べる必要があるが，いわゆる過学習が生じているのではないかと考えられる．

5 おわりに

本研究では見出し語の違いが感性情報分類の精度に及ぼす影響について調査した．見出し語の求め方 (屈折の有無 × 品詞情報の有無) を変えて word embedding を求め，それを用いて，新聞社説の文に対する感性情報分類を行ったところ，屈折を考慮した見出し語の方が性能が良いことが分かった．一方，品詞情報を考慮した見出しについては，性能はほとんど向上しなかった．

感性情報分類実験において，各文をそこに含まれる単語の word embedding の算術平均として特徴表現したが，これは，語順情報を考慮しないものであるため，文の特徴表現としてはかなり “荒い” ものであった．このことが，先行研究 [9] と比べた場合の性能の低下を招いているのではないかと考えられる．

今後は，感性情報分類における各文の特徴表現に語順を考慮したものを考えるとともに，分類に用いるデータを増やして同様の実験を行う必要がある．

参考文献

[1] T. Mikolov et al.: “Efficient Estimation of Word Represention in Vector Space”, International

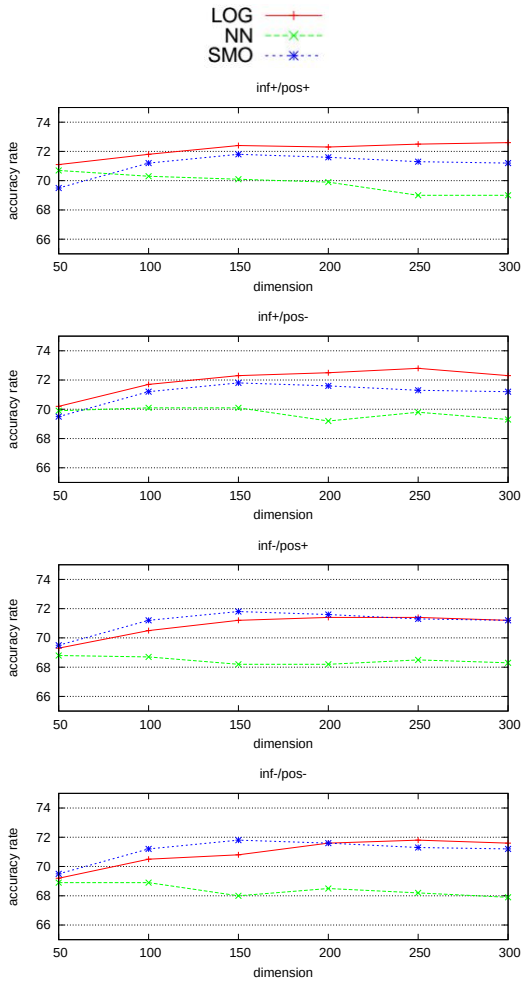

図 3: 分類器の違いによる分類正解率の比較

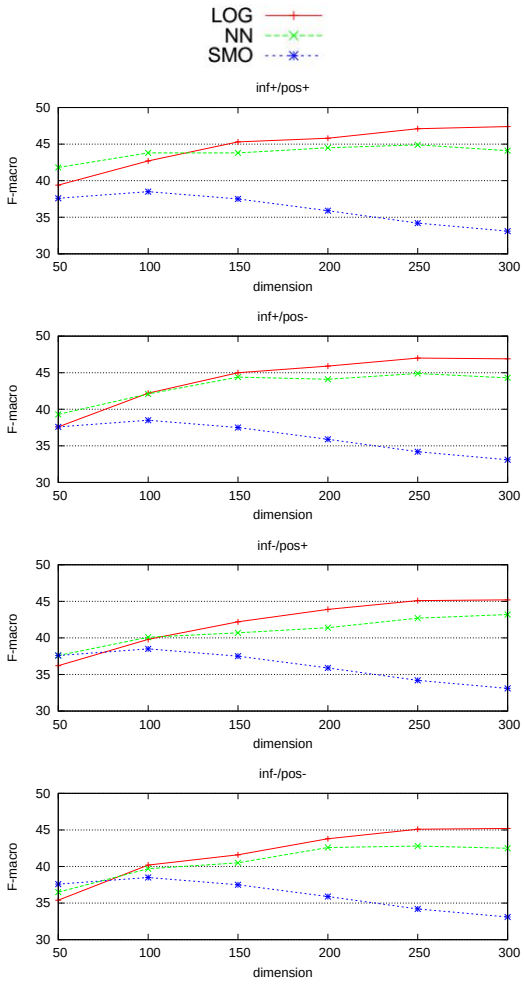

図 4: 分類器の違いによる F 値マクロ平均の比較

Conference on Learning Representions Workshop, 2013

[2] J. Pennington et al.: “GloVe: Global Vectors for Word Representation”, Empirical Methods in Natural Language Processing, 1532–1534, 2014

[3] J. Turian et al.: “Word Representations: A simple and general method for semi-supervised learning”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 384–394, 2010

[4] 加藤和平ほか: “word2vec と深層学習を用いた大規模評判分析”, 言語処理学会第 21 回年次大会, 525–528, 2015

[5] 野口正樹ほか: “分散表現を用いたヤフー知恵袋の要約”, 言語処理学会第 21 回年次大会, 1064–1067, 2015

[6] 山本翔馬ほか: “分散表現を用いた教師あり機械学習による語義曖昧性解消”, 情報処理学会研究報告, 2015-NL-224(17), 2015

[7] 三浦弦太, 石井絵理香, 山村毅: “ナイーブベイズ分類を用いた新聞社説の感性的情報の分類”, 電気・電子・情報関係学会東海支部連合大会講演論文集, L1–5, 2014

[8] 藤巻直也, 三浦弦太, 山村毅: “意見情報の推定における Bag-of-Words の有効性について”, 電気・電子・情報関係学会東海支部連合大会講演論文集, D4–3, 2015

[9] 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