word embeddingを利用した文の特徴表現の違いがメッセージ情報分類に与える影響について

情報科学科　加藤 里奈　　指導教員：山村　毅

1　はじめに

近年，自然言語処理に word embedding を取り入れた研究が行われるようになった．word embedding とは，意味の類似したものが近くに配置されるように，単語をベクトル表現したものであり，単語の意味表現としても使えるものである．この単語のベクトル表現を利用して文のベクトル表現を求めることができれば，文の意味表現が使えるようになる為，既存の自然言語処理の性能が向上できると考えられる．新聞社説を対象とし, 特徴量に word embedding を用いたメッセージ情報分類の研究に宮川ら [1] によるものがある．この研究では，文の表現方法に算術平均を用いているが，他の表現方法について調べることは価値があると思われる．

本研究では、新聞社説を対象に word enbedding を利用した文の特徴表現の違いがメッセージ情報分類に与える影響について調べる．具体的には，新聞社説の文の表現方法を様々に変化させて，メッセージ情報分類の精度を比較する．

2　分類カテゴリー

メッセージカテゴリーには，三浦ら [2] によって提案された，肯定，期待，疑問，助言，否定，非メッセージの 6 つを利用する．実験データは，2006 年前半の毎日新聞の社説 344 記事 9492 文を使用し，あらかじめ人手により正解を付与している．その内訳は，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	非メッセージ
308	417	357	1024	1138	6248

3　特徴表現

3.1　新聞社説

本研究では，各文の全単語を用いた「全文」，各文を係り受け解析し核となる部分だけを残した「全文（縮約）」，各文から助詞，助動詞と記号を削除した「全文（自立語のみ）」，各文の末尾から指定された数の単語を取り出した「文末」の 4 種類の社説を用いる．又，それぞれの社説においては，単語の屈折（“走る”と“走ら”を区別するか）と品詞情報の有無も考慮する．

3.2　表現方法

文の表現方法として，そこに含まれる word embedding の算術平均，要素ごとの最大値，要素ごとの最小値，連結の 4 種類を用いる．連結とは，単語が並んでいる順番にそのまま word embedding を連結させたものである．

4　実験方法

新聞社説を対象にメッセージ情報分類を行う．特徴表現として，社説の各文には 3.1 節の 4 種類の処理を施したものを用い，文の表現には 3.2 節の 4 種類を用いて全ての組み合わせを試す．ただし，連結においては次元数が大きくなりすぎてしまう為，文末のみに適用する．これらの特徴表現を分類器（ナイーブベイズ，対数線形モデル，サポートベクトルマシン，ニューラルネットワーク）にかけて分類を行い，10 分割交差検定により正解率と F 値マクロ平均を求め，評価する．形態素解析には MeCab を，word embedding の生成には word2vec[3] を使用し，学習データに 2010 年の毎日新聞の記事データを用いた．又，分類器の実装には WEKA を用いた．

5　実験結果とまとめ

それぞれ組み合わせた手法の最大正解率と F 値マクロ平均の結果を表 5 に示す．（①は全文，②は全文（短縮），③は全文（自立語のみ），④は文末を表している．）又，文末 2-10 単語の正解率の結果を表 3 に示す．

表 2　分類結果 (%)

	正解率	F 値 マクロ平均		正解率	F 値 マクロ平均
①平均	73.5	47.9	③平均	71.6	43.8
①最大	70.7	42.7	③最大	69.0	36.5
①最小	71.2	42.8	③最小	69.5	38.0
②平均	77.0	56.6	④平均	79.4	62.9
②最大	75.9	54.3	④最大	78.8	61.3
②最小	75.8	54.0	④最小	78.7	60.2
			④連結	78.6	63.2

表 3　文末×平均の正解率 (%)

10	9	8	7	6	5	4	3	2
77.2	77.5	78.3	78.9	79.2	79.4	79.3	75.8	70.2

実験結果より，新聞社説の処理の中で性能が良い順に文末，全文（短縮），全文，全文（自立語のみ）の順となった．この結果より，分類には付属語（助詞や助動詞）が大きく影響していると考えられる．

最も性能が良かった手法は，文末単語を利用した算術平均，及び連結したものであった．平均においては，4 種の新聞社説のどの場合においても高い結果となっている．又，表 3 より文末において 4-6 単語を用いることが word embedding を用いた特徴表現において最も有効であり，文の表現方法は，算術平均，及び連結したものが有効であると考えられる．

参考文献

[1] 宮川知也，山村毅『感性情報分類における見出し語の違いの影響について』 言語処理学会第 23 回年次大会発表論文集，pp.1046-1049，2017

[2] 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

[3] Tomas Mikolov，Kai Chen，Greg Corrado，Jeffrey Dean，“Efficient Estimation of Word Represention in Vector Space”， International Conference on Learning Representations Workshop，2013