E-007

Word2Vec による分類・推定における事前処理法の提案

DAO VAN TUAN† 佐藤 浩‡
防衛大学校理工学科研究科情報数理専攻†
防衛大学校電気情報学群‡

１．はじめに

現在，日本語には様々な言語データが存在する． Wikipedia をはじめ，ツイート，Yahoo!ニュース，新聞記事などのデータが，自然言語処理分野に活用されている．大量のデータに各種の統計的手法を適用することで良い結果が得られるが，精度があまり上がらない場合もある．この原因は手法の不完全性，または， データの質にあると考えられる．

研究においては，無料で大量の日本語データが得られる Wikipedia が多く用いられる．Wikipedia のデータを全て用いた研究[1,2,3,4]では，既存手法より分類･推定における良い結果が得られたものの，大きな差は見られなかった．名詞，動詞，形容詞のみを抽出して処理した研究もあるが[5,6],不要な情報が残ったため，結果はわずかしか向上しなかった．

大量のデータを用いた学習には，多くの時間が必要である[7]．本研究では，Wikipedia のデータの関連する記事から名詞，動詞，形容詞だけを抽出することで，質の高い学習データを作り出す手法を提案する.実験により，本手法が既存研究に比べ，短い学習時間で良い結果を得られることを示す．

２．Word2Vec

Word2Vec は Mikolov らにより発表された単語群のベクトル化手法である [8,9]．Word2Vec はニューラルネットワークを用いて単語の分散表現を計算する手法，およびそのオープンソース実装の名称である．Word2Vec では単語の意味を表現するベクトルが低次元で作られる．さらに，意味的に関連が強い単語はベクトルが近くなるという特徴を持つため，現在の自然言語処理分野において多く用いられている．

Word2Vec のモデルとして CBOW（周辺の単語から中心の単語を予測する）と skip-gram（中心の単語から周辺の単語を予測）の２つがあるが，Mikolov らの実験の結果 [8] では skip-gram モデルの方は精度がため，本研究は skip-gram モデルを用いて実験を行なう．

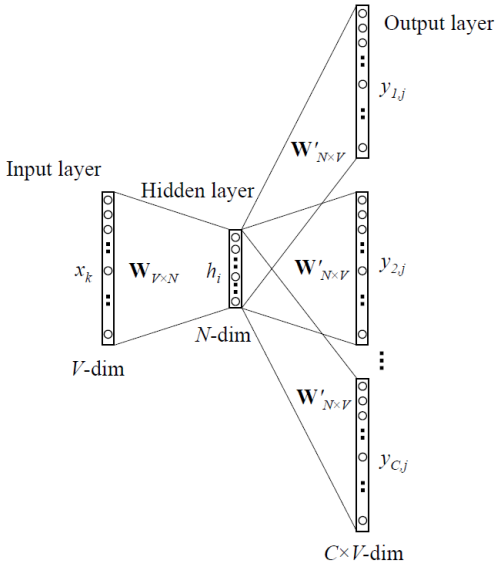

図 1. Skip-gram モデル

Skip-gram モデルでは単語w_tから複数語w_{t+j}が予測される確率を$p(w_{t+j}|w_t)$とおき，次の式で示される目的関数を最大にする単語ベクトルを学習する[8].

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}|w_t) \qquad (1)$$

ここで，T はコーパスの単語数，c は文脈のサイズを示す．$p(w_{t+j}|w_t)$の計算は下の式で定義される：

$$p(w_O|w_I) = \frac{exp(v'^{T}_{w_O} v_{w_I})}{\sum_{w=1}^{W} exp(v'^{T}_{w} v_{w_I})} \quad (2)$$

ここで，v_wは入力された語のベクトル，v'_wは出力される語のベクトルである．

Word2Vec による言語処理でよい精度を得るためには，パラメータの調整が欠かせない[7]．本研究では，パラメータ調整の対象を Size，Window，Min-count とした．ここで，Size はベクトルの次元，Window は文脈の最大単語数，Min-count は最低出現回数を表す．

適切なパラメータを得るため，訓練データから類似語の５ペアを選び，上記の３つのパラメータの値を決められたパラメータの範囲（表 1）で変更し，類似度が高いものを訓練データのパラメータとする．パラメータ調整の結果を表 2 に示す．

表 1. パラメータ選択の範囲

パラメータ	数値の範囲
Size	[50,100,200,300,400]
Window	[1,2,3,4,5,6,7,8]
Min-count	[1,2,3,4,5,6,7,8]

表 2. パラメータ調整の結果

パラメータ	数値
Size	50
Window	2
Min-count	1

３．提案システムの構築

本研究では，言語データとして Wikipedia2017 年 5 月（約 10 億単語）を用いた．データの単語単位への分割を行うための分かち書きにはオープンソースの日本語形態素解析器である MeCab[10]を使用した．また Web ページには，様々な新語や流行語などが含まれる．MeCab の標準辞書に存在しないこれらの単語に適切に形態素解析を適用するため，Web 上の言語資料から得た新語で標準辞書を拡張した mecab-ipadic-neologd[11]を単語辞書として用いる．

学習時間を短縮するため Wikipedia から名詞，動詞，形容詞の３つの品詞のみを抽出し，これを本研究における全コーパスとする．全コーパスからキーワードに関する記事のみを取得する．比較のため，

- 全コーパス
- 全コーパスからランダムな抽出（全コーパスの半分の語数）

を用いた実験を行なった．

４．実験

本節では，推定および分類のため，以下の２つの実験を行う．

実験１：検索補助の評価

実験 2：単語のポジティブ・ネガティブ分類

4.1 実験１：検索補助の評価

実験の手順を以下に示す：

1. キーワードを入力し，類似度の高い順に 10 単語を出力する
2. 出力された単語がキーワードと関連しているかを被験者に評価してもらう．
   (ア) 関連する：1
   (イ) 関連しない：0 とする

表 3. 入力“自衛隊”の出力結果

順	類似語	類似度
1	陸上自衛隊	0.803
2	航空自衛隊	0.780
3	陸自	0.738
4	自衛官	0.737
5	自衛隊員	0.736
6	防衛省	0.734
7	空自	0.695
8	防衛庁	0.692
9	在日米軍	0.679
10	自衛隊海外派遣	0.675

入力単語を“自衛隊”に対し,学習済みの Word2Vec により類似度が高いと判定された単語を表３に示す．

他の入力単語“日本，アメリカ，日米同盟など”10 単語を選んで実験した結果を表 4 に示す．

ここで,有用率は被験者 5 人による評価の平均値である．

表 4．検索補助の評価

手法	有用率
既存研究（tf-idf 法）	0.46
既存研究（細川ら）	0.57
全コーパス	0.62
ランダム	0.44
提案手法	0.75

全コーパスで学習するだけでは，既存研究（tf-idf 法）より 16 ポイント,細川らの研究[3]より，5 ポイントしか上がらないが，提案手法では，それぞれを 29 ポイント,18 ポイント上回った．

学習時間の結果を表 5 に示す．

表 5．学習時間の比較

手法	学習時間
既存研究（細川ら）	11
全コーパス	4
ランダム	2
提案手法	0.8

細川らは全ての Wikipedia のデータで学習するため，一番時間がかかった．品詞を限定することで，学習時間は半分以上減少することができた．本提案手法では，さらに一部の必要な情報だけで学習するため，時間は十分短縮ができると言える．

4.2 実験 2 ：単語のポジティブ・ネガティブ分類

分類対象としてポジティブ，ネガティブの値を持っている単語感情極性対応表[12]を用いた．対応表の単語には，-1 から+1 の実数値を割り当てられており，-1 に近いほどネガティブ，+1 に近いほどポジティブと考えられる．対応表における単語の数は，ポジティブ 5142 単語，ネガティブ 49983 単語である．提案手法の分類性能を比較するため，以下の手順で実験を行なった．提案手法では，Wikipedia から，感情に関連する記事のみを抽出し，Word2Vec の学習データとして用いた．

1. 対応表から絶対値の高いポジティブな単語のグループ（10 単語）とネガティブな単語（10 単語）を作る．
2. 判定したい単語について，両グループの平均ベクトルに近い方を，その単語のグループとする．
3. 判定したい単語の対応表を対照し，正解率を算出する．

手順のイメージを図 2 に示す．

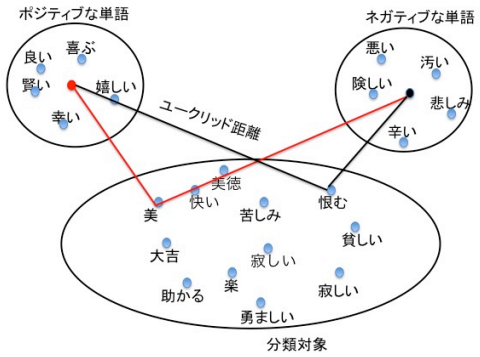

図 2．ポジティブ・ネガティブ分類

学習データがよいほど，よいポジティブベクトル，ネガティブベクトルを作ることができる．それぞれ Wikipedia の扱い方による正解率の違いを表 6 に示す．

表 6．ポジティブとネガティブ分類の評価

手法	正解率
全データ	0.70
ランダム	0.60
提案手法	0.85

表 6 の結果によると，Wikipedia から，必要とする品詞（名詞，動詞および形容詞）を抽出するだけでは正解率は高くない．ランダムに単語を減少することによって学習時間は早くなったが，精度が落ちてしまう．提案手法の場合は必要なデータを取得することによってよりよい結果が得られた．

５．おわりに

本研究では Wikipedia のキーワードに関する記事データから重要な品詞のみを抽出し，加えて Word2Vec のパラメータを調整することで，分類・推定において良い結果が得られることを示した．本手法は，一部のデータしか使わないため，学習時間を大幅に短縮することができた．その一方で，各キーワードによって異なる訓練データを取得しなければならないが，学習時間が短いため，再学習は問題とはならない．Word2Vec のパラメータ調整に関して，今回は三つの要素しか行わなかったが，今後，他のパラメータの調整を行う予定である．

参考文献

[1] 鹿島ら，“Word2Vec と Web 検索を用いた検索クエリ置換手法”．DEIM Forum 2017.

[2] 宮田ら，“統計的学習モデルを利用した日本語慣用句の意味的曖昧性解消”，情報処理学会第７９回全国大会.

[3] 細川ら，“分散表現を用いた Web 検索結果の自動タギング”，自然言語処理学会第７９回全国大会.

[4] 佐藤ら，“感情語に基づくことわざ推薦システム”，だ１５回情報科学技術フォーラム.

[5] 山本ら，“分散表現を用いた教師あり機械学習による語彙曖昧性解消 ”，情報処理学会研究報告， Vol.2015-NL-224 No.17

[6] 野沢ら，“Word2Vec を用いた代替食材の発見手法の提案”，電子情報通信学会技術研究報告, pp.41-46,2014-09

[7] 西尾泰和，“Word2Vec による自然言語処理”O’ Reilly Japan， 2014.

[8] Tomas Mikolov， Kai Chen， Grag Corrado， Jeffery Dean， “Efficient Estimation of Word Represention s in Vector Space” Cornell University Library arXiv. org， arXiv;1301. 3781v3[cs. CL]， 2013.

[9] Tomas Mikolov，T., Sutskever, I., Chen, K., Corrado G. and Dean, J.: Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems, pp.3111-3119, 2013.

[10] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.

[11] Mecab の最新辞書 https://github.com/neologd/mecab-ipadic-neologd

[12] 単語感情極性対応表 http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html