分散表現と特徴量の組み合わせによる文書分類法

DAO VAN TUAN† 佐藤 浩†
防衛大学校電気情報学群情報工学科†
{em55039, hsato}@nda.ac.jp

1 はじめに

計算機による自然言語処理では，単語と単語の関係を構築することが重要であり，様々な手法が提案されている．そして，そのための学習データとして，インターネット上に存在する大量の言語データが活用されて始めている．

単語間の関係構築については，単語をベクトル化し，その距離関係に着目した研究が近年注目を集めている．特に，Mikolov により提案された Word2Vec[1][2] は，単語を低次元で表現できる手法として多くの研究者により利用されている．Word2Vec においては, 意味の近い単語から生成されたベクトルは類似したベクトルとなる特徴を持つことが期待される．

分散表現の距離関係を文書分類に活用する研究例はいくつかある[3][4]が, 単語のベクトル化のためには，ある程度の量のデータが必要であるだけでなく，データの質も同時に重要である．

本研究では，分散表現による文書の分類にあたり，適切な特徴ベクトルを構築することが重要であると考える. Word2Vec は単語の意味関係を捉えることができるが，ベクトル化のための学習データが持つ特徴量を考慮しない．本研究は，TF-IDF 法を利用することでカテゴリの重要語を抽出し，カテゴリの特徴ベクトルをより適切なものとさせることを目的とする．実験結果より，提案手法は既存研究に比べ，短い学習時間で良い結果を得られることを示した．

2 特徴量及び分散表現

文書の特徴量を測るための代表的な手法に TF-IDF がある．TF-IDF はコーパス内の文書，または文書の集合に対する単語の重要度を評価するために使用される統計的尺度である．TF(t, d) はテキスト d 中の単語 t の出現頻度である．IDF は単語が出現する文書数の逆数であり，出現する文書数が少ない単語に大きな値を与える．

$$idf(t) = \log\frac{N}{df(t)} + 1 \qquad (1)$$

ここで，N はテキスト集合中の総テキスト数，DF(t)は単語 t が出現する文書頻度である．最終的な値，は次のように計算される．

$$tfidf(t,d) = tf(t,d) \times idf(t) \qquad (2)$$

TF-IDF 法を用いて文書分類のために重要となる単語を見つけることができる．

Word2Vec による単語のベクトル化には，CBOW(周辺の単語から中心の単語を予測)と Skip-gram(中心の単語から周辺の単語を予測)と呼ばれる 2 つのモデルがある．実験では Skip-gram モデルの方が高い意味精度を示したため，本研究は Skip-gram モデルを採用する．

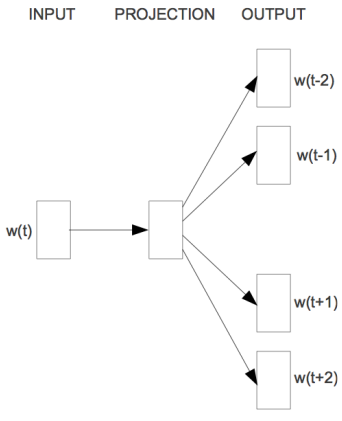

図 1 Skip-gram モデル

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$p(w_{t+j}|w_t)$とし，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．

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}|w_t) \qquad (3)$$

ここで，T はコーパスが持つ単語数， c は文脈のサイズを示す． $p(w_{t+j}|w_t)$は，以下の式で定義される．

$$p(w_O|w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W} \exp({v'_w}^T v_{w_I})} \quad (4)$$

ここで，v_wは入力される単語ベクトル，v'_wは出力される単語ベクトルである．

4 提案手法

提案する文書分類法の手順を以下に示す．

1. コーパス（日本語 Wikipedia）から，カテゴリごとに記事を収集し，前処理を行う．
2. 各カテゴリにおける単語の重要度を TF-IDF 法を用いて算出する．
3. 各カテゴリの記事から重要度の高い単語を抽出し，重複を除き，平均をとる → フィーチャカテゴリベクトルと呼ぶ．
4. フィーチャカテゴリベクトルと各カテゴリ内の単語ベクトルへの平均距離をカテゴリ閾値とする．
5. 判定する文書と各フィーチャカテゴリベクトルとのユークリッド距離を計算し，カテゴリ閾値範囲内にあるうちで最短のカテゴリに属すると判定する．

本手法は，（1）分散表現により単語間の関連を表現した上で，単語の重要性を統計的特徴量を利用して取り込み，（2）コーパスから必要な文書および必要な品詞のみを抽出して利用する，という２つの特徴を持つ．

5 実験

提案手法の分類精度を確認するため，Wikipedia のカテゴリ分類とニュースのカテゴリ分類という，長短２種類の文書に対する分類実験を行った．

Wikipedia のカテゴリ分類については，スポーツ，文化，経済，軍事からランダムに 10%のカテゴリ記事を抽出したものをテストデータ１とする．ニュース分類については，Web ニュースサイトの経済，スポーツから抽出した文書をテストデータ２とする．

既存研究の分類精度との比較を行った結果を表 1，2 に示す．表１，２より，本研究の提案手法は，丸井ら，及び加藤らの結果より分類精度が高い．分散表現のみならず，特徴量を考慮することにより，分類精度が高まることが分かった．

表 1 Wikipedia のカテゴリ分類精度

	正解率（precision）
丸井ら[3]	0.53
提案手法	**0.89**

表 2 ニュースの記事分類精度(F 値)

	経済	スポーツ
ナイーブベイズ	0.58	0.84
加藤ら[4]	0.75	0.90
提案手法	**0.87**	**0.91**

6 まとめ

本研究では．文書分類において，分散表現と特徴量を組み合わせた手法を提案し，既存研究よりも良い結果を得た．パラメータの調整により分散表現の精度はさらに上がる．今後は，自動的なパラメータ調整システムを考案し，分類精度のさらなる向上を目指す．

参考文献

[1] 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.

[2] 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.

[3] 丸井淳己, 萩原正人, “Category2Vec: 単語・段落・カテゴリに対するベクトル分散表現”, 言語処理学会第２１回年次大会, 2015.

[4] Ryoma Kato, Hiroyuki Goto, “Categorization of Web News Documents Using Word2Vec and Deep Learning”, IEOM2016.