Paragraph Vector の日本語弁別の応用と既存手法との比較

阪本　彩†　　小川　毅彦†　　林　誠治†
†拓殖大学　工学部　電子システム工学科

1. はじめに

テキストマイニングとは，文字列に対するデータマイニングである．カテゴリ分析の手法として主流である Bag of Words では同じ単語でも文脈により意味が異なる場合もあるため，テキストの意味を完全に把握することは困難である．正確にテキストのカテゴリ分類をするためには，単語の頻度だけでなく文章の文脈も読み取ることが必要である．本研究では，テキストのカテゴリ分類を正確に行うために，文章から推測する手法である Paragraph Vector と，単語から推測する手法である Bag of Words との比較を試みる．

2. 関連研究

2.1 Bag of Words

文章の類似度を測る際に，単語の頻度を利用する．ある複数の文章に対して，共通の単語が多く含まれる場合その文章は近い意味を持つと考える．

2.2 Paragraph Vector

Le [1] らが考案した自然言語のパラグラフのベクトルを教師なし学習で生成するアルゴリズムである．2013 年後半に Google Inc, の T. Mikolov が発表した手法で，近年，英語の応用は盛ん[2]であるが，日本語への応用と検証は不十分である．結果として，T. Mikolov は，Bag of Words の Error Rate が 8.1%に対し，Paragraph Vector は 3.82%であることを示している．また，計算時間に関しても利点があり，次元削減をし，計算時間を大幅に削減できる．

3. 目的

本研究では，Paragraph Vector の日本語への応用と Bag of Words との比較を目的とする．比較対象は，①精度，②計算時間とする．Bag of Words と Paragraph Vector それぞれに対して，同じ文章群（19000 個の文章）を Positive, Negative の 2 値分類し，Paragraph Vector の有用性を比較調査することが研究の目的である．

本研究では，Paragraph Vector と Bag of Words の精度と計算時間を比較する．精度が優れていても必要な計算機の記憶容量を考慮しない計算時間だと実用には難しい．今回は Bag of Words, Paragraph Vector ともに特徴ベクトルに変換した後，Support Vector Machine を使用する際に学習に要した時間と分類に要した時間を Python3 スクリプトにて測定する．データセットとしては，実験のテキストデータとしてレストランのレビューを収集した livedoor グルメ Data Sets[3]を使用する．

4. 実験

データは livedoor グルメデータセットの口コミから 5 星の評価のコメントを Positive とし，1 星，2 星のコメントを Negative とする．それぞれのデータ量を 19000 文章とし，計 38000 文章を収集した．特徴ベクトル作成に，今回評価対象の Bag of Words と Paragraph Vector それぞれを使用する．弁別にはサポートベクターマシンを使用する．学習に際して，学習量が弁別に影響を与えるため，データと学習データの比率を 10%から 90%まで 10%刻みで実験を行う．評価は，精度と学習・弁別時間の 2 つで行う．

5. 結果

90%学習した後 Positive に分類した Precision の結果を以下の図 1 に示す．グラフの横軸は文章量，縦軸は Precision の平均を表す．文章量の幅が大きいため横軸は対数目盛りに設定した．一部の文章量における精度が低下している原因としてデータ量が少ないことが考えられる．

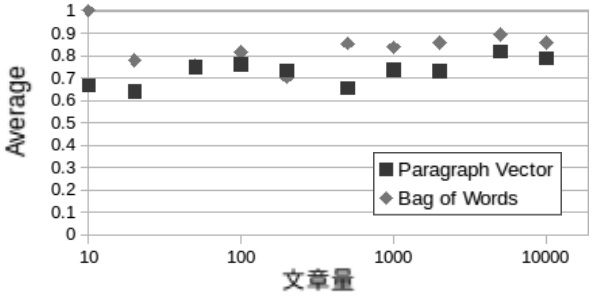

図 1: Positive の Precision 結果[90%学習]

6. 考察

6.1 精度比較

精度では文章量に関係なく Bag of Words が Paragraph Vector より上回る結果になった．しかしながら両手法の差は 0.1 程度でありデータセットによる誤差範囲と思われる．Bag of Words の結果のグラフは文章量にあまり依存せず水平に近いのに対し，Paragraph Vector のグラフは文章量が増えるごとに精度も上がっていることから，今後文章量が大きいデータが与えられた際に精度向上が期待される．

6.2 計算時間比較

文章量に対しての学習時間に着目すると，文章量が多いほど Paragraph Vector は Bag of Words より計算速度が速いことがわかる．学習量が 90%と 10%を比較した場合，具体的には文章量 19000 の場合，学習量 90%，10%の Paragraph Vector はそれぞれ 376.8[sec]，1.5[sec]でありおよそ 251.2 倍の差がある．また学習量 90%，10%の Bag of Words はそれぞれ 1943.8[sec]，8.5[sec]でありおよそ 228.7 倍の差がある．このことから計算時間にかかる増加率は Paragraph Vector は小さいことに加え，Paragraph Vector は Bag of Words より計算時間が短いことから計算時間の優位性が分かる．分類時間に着目すると Support Vector Machine は分類の際にランダムにデータを選択するため誤差が生じたと考える．

7. おわりに

本研究では livedoor グルメ Data Sets の口コミデータと口コミにつけられているラベルをデータセットとし，Paragraph Vector と Bag of Words の手法を用いて Positive, Negative の分類を行った．精度では Paragraph Vector よりも Bag of Words のほうが優れてはいたが，Bag of Words は Latent Semantic Indexing を使用することで次元圧縮を行わないと計算が出来なかった点と，Paragraph Vector のほうがおよそ 258.4 倍，計算時間が早い点があった．このことから文書のカテゴリ分類には精度と計算時間の兼ね合いを考慮することが必要であることが考えられる．

参考文献

[1] Q. Le and T. Mikolov Distributed Representations of Sentences and Documents. In Proceedings of The 31st International Conference on Machine Learning, 2014.
[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed Representations of Words and Phrasesand their Compositionality, In Advances in Neural Information Processing Systems, 2013.
[3] Livedoor グルメ, <http://blog.livedoor.jp/techblog/archives/65836960.html>.