BERT を用いたクラスタ分析による文章分類

Classification of Documents by Cluster Analysis using BERT

2022Y12　後藤 貴樹（Takaki Goto）
担当教員　出口 利憲（Toshinori Deguchi）・田島 孝治（Koji Tajima）

１．序論

近年、コンピュータやスマートフォンが普及したことで人々の多くがインターネットを用いている。これによってインターネット上に存在するデータは数を増し、それらから必要なデータを抽出することは難しくなっている。

このような問題の解決策としてテキストマイニングが挙げられる。これは日本語などの自然言語をコンピュータが理解できる形へと変換することによって文書データの中から自身の求める情報を抽出する技術である。

本研究は文書分類を用いて、テキストマイニングの処理の一部である次元削減をクラスタ分析によって行った場合の有効性を評価することを目的としている。

２．BERT

本研究では東北大学が作成した日本語 BERT モデル[1]を用いる。

BERT は 2018 年に Google によって発表された自然言語処理モデルである。単語などのトークンを高次元のベクトルとして表現することでコンピュータが処理できるようにする分散表現を用いている。このモデルの一番の特徴は Masked Language Model を用いて、一文を文頭と文末から文章を学習することで文脈を考慮した単語の分散表現を可能としている点である。これに加えて Next Sentence Prediction と呼ばれる二つの文が関係するかの学習を行うことによって文章を考慮した分散表現が可能となっている。

３．実験方法

この研究では livedoor ニュースコーパス[2]を用いてダウンロードした 9 ジャンルのニュース記事を分類の対象とする。

ダウンロードしたニュース記事を BERT によってトークン化する。このトークンの先頭から 128 個を分散表現によりベクトル化する。これにより生成された 128 つて以下の手順でクラスタ分析による次元削減を行う。

Table. 1 Average and CLS token results(genres).

	5genres	9genres
Vector Average	0.54	0.44
CLS Token	0.58	0.53

1. ウォード法を用いた階層的なクラスタリングを行う
2. ニュース記事内に存在するトークンを 0 と 1 で示した行列を作成
3. 1 の結果をもとにトークンがどのクラスタに属するかを 0 と 1 で示した行列を作成
4. 2 と 3 で作成した行列の積をとり、それをクラスタリングする

この手順によって 128 個の 768 次元のベクトルで表されていた一文書の情報を指定したクラスタ数にまで削減することができる。

このプロセスを以下の条件で行い、ジャンル数と文書数の関係を調べる。

- ジャンル数 3:1 ジャンルにつき 10 文書
- ジャンル数 5:1 ジャンルにつき 10 文書
- ジャンル数 7:1 ジャンルにつき 10 文書
- ジャンル数 9:1 ジャンルにつき 10 文書
- ジャンル数 3:1 ジャンルにつき 20 文書
- ジャンル数 5:1 ジャンルにつき 20 文書
- ジャンル数 7:1 ジャンルにつき 20 文書
- ジャンル数 9:1 ジャンルにつき 20 文書

この方法で得た行列をクラスタリングによってジャンル数にまで分類する。このクラスタ番号の最頻値をそのジャンルでの正解クラスタ番号として番号を割り振り、その番号の文書数を正解数として正解率を求めた。

４．実験結果と考察

このクラスタ分析による方法の評価のため、BERT で標準的に利用されるベクトルの平均値を用いた文章分類と CLS トークンを用いた文章分類を行い、それらの結果と比較した。この結果を Table.1 に示す。

これと同条件でクラスタ分析のクラスタ数を 10 から 300 まで 10 区切りで文章分類を行った。その結果を Fig.1 に示す。

Fig.1 から 9 ジャンルで行った文章分類よりも 5 ジャンルで行った文章分類の方が高い正解率になることがわかる。また、クラスタ数が増えるほどにその差がなくなり 0.40 ほどの正解率に収束していくことがわかる。

Table.1 に示した平均値を用いたものと CLS トークンを用いたものと比較し、クラスタ分析を用いた文章分類は 9 ジャンルの際はそれらより正解率が低く、あまり有効的な次元削減の方法とは言えなかった。また、5 ジャンルの際も平均ベクトルや CLS などでの文書分類に近い正解率となった。

また、9 ジャンルで行った実験で文書数を 10 と 20 で変更して行った結果を Figure.2 に、平均ベクトルと CLS トークンを用いたもので文書数を変更したものを Table.2 に示す。

Fig.2 からは同じジャンル数であれば使用する文書の量が少ないほうが高い正解率であることがわかる。

クラスタ分析による次元削減を用いた文書分類においてジャンル数が 5 である方の正解率が高い結果になったが、これは文章分類の方法によるものである可能性が高い。今回は次元削減後の行列をクラスタリングし、それを文章分類として用いている。この方法ではクラスタ番号の最頻値を利用しているのだが、重複した場合は別のものを割り当てるようにしている。このように重複した場合は正解率が著しく下がる上にクラスタリングによる分類は重複が多かった。重複が多かったためにジャンル数の母数が少ないほど正解率が高くなったと考えられる。そのためこの方法による文章分類ではあまり次元削減の有効性を調べるには向かない可能性が高い。

Table. 2 Average and CLS token results(documents).

	10documents	20documents
Vector Average	0.44	0.53
CLS Token	0.53	0.58

5. 課題

本研究ではニュースの本文をトークン化したものの最初から 128 個を使用して用いている。この中には単語の他に接続詞などの文書の特徴的でない部分が含まれている。それら情報を削り、ニュースに含まれている他の動詞や名詞などのトークンを用いることによってより特徴を残した次元削減が可能になると考えられる。

6. まとめ

今回の研究では BERT を用いたクラスタ分析による次元削減では平均ベクトルを用いた方法や CLS トークンを用いた方法と同等の結果が得られることがわかった。ただし、用いたトークンの数や種類といったものを調整することによってこれよりも効率的な次元削減を行える可能性がある。そのため、今後は今回の研究で得られた課題を解決し、適切な評価方法を確立する必要がある。

7. 参考文献

[1] livedoor ニュースコーパス, ロンウィット
https://www.rondhuit.com/download.html
[2] cl-tohoku/bert-japanese
https://github.com/cl-tohoku/bert-japanese

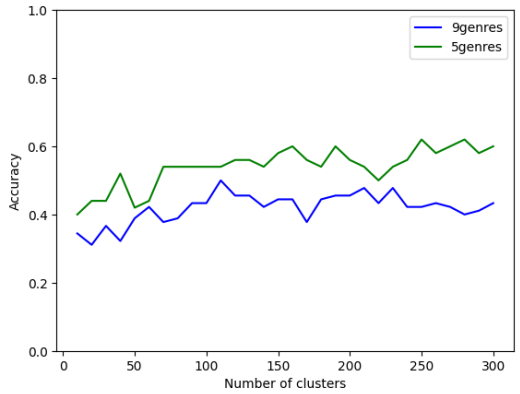

Fig.1 Cluster Analysis Results(genres).

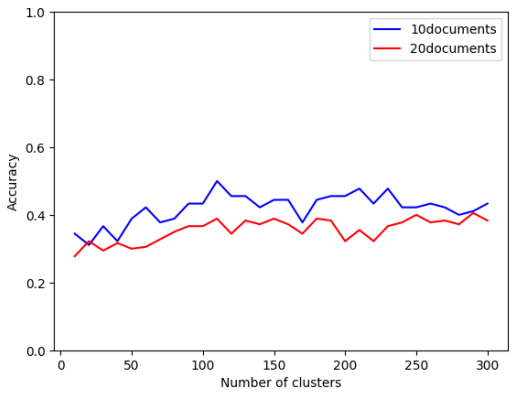

Fig. 2 Cluster Analysis Results(documents).