5W-03

BERT による文書分類の根拠提示手法の検討

坂和樹† マッキン ケネスジェームス† 永井 保夫†
Kazuki Saka Kenneth J Mackin Yasuo Nagai
東京情報大学‡

1. はじめに

近年,深層学習を用いたニューラルネットワークモデルが注目されている.そのひとつである BERT は自然言語処理において優れた成果を残している[1].しかし,BERT の出力は推論結果のみで推論の過程は全く明かされないことから推論過程の説明の必要性が指摘されている[2].

我々はこのような問題を解決するために,BERT の文書分類において推論の根拠となるような単語を提示する手法の提案を行う.

2. BERT による文章分類

本研究で使用した BERT モデルは東北大学乾研究室が公開しているモデル[3]を用いている.分類対象の記事に対して品詞や記号などの除外は行っていない. フレームワークは pytorch 1.10.2 を使用した.出力次元数は 512,単語ベクトル次元数は 768 に設定した.最適化関数は確率的勾配降下法を用いて学習率は 0.001 で学習した.損失関数はクロスエントロピー誤差検証を使用した.訓練データは livedoor ニュースコーパスの全カテゴリから 736 件のデータをランダムに抽出した.テストデータは訓練データ以外のデータをランダムに 736 件抽出し,分類精度は 88.7%だった.

3. BERT による文書分類の根拠提示手法

図 1 は BERT モデルによる文書分類の流れを示している. BERT モデルは入力された文章をベクトル表現に変換し,単語ごとに分割する.分割した単語ごとに Attention 値を計算する.最終的にカテゴリごとに分類スコアを計算し,それが最も高かったカテゴリを推論結果としている.

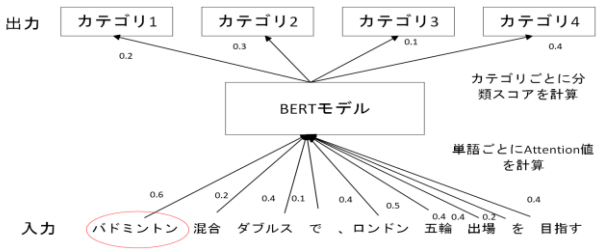

図 1:BERT による文書分類

BERT による文章分類の根拠提示手法としては,次の Attention 値手法, WD 値手法, Att*WD 値手法が提案されている[4].

(1) Attention 値手法

Attention 値手法では, 推論の際にモデルがどの単語に注目したかを示す Attention 値が高い単語を根拠語として提示する. Attention 値は BERT モデルが入力文に対して計算した値であり,モデルが推論するときにどの単語を重要視したかを示している.BERT モデルは, Attention 値を計算し出力することができる. Attention 値手法では,BERT モデルの 12 層の Transformer 層が計算した全てを取得して加算した値を Attention 値とした.

(2) WD 値手法

WD 値手法おける WD 値は, 入力文の一単語のみをモデルに分類させた時の分類スコアである. 通常の文書分類の際,入力は文章である.一方,WD 値を計算する時は入力文を単語ごとに入力して,分類スコアを計算する.分類スコアは分類するカテゴリの数だけ出力されるため,WD 値は分類対象となる文章の正解カテゴリの分類スコアを表す.図 1 の場合,”バトミントン”のみをモデルに入力して正解カテゴリであるカテゴリ 4 の分類スコアを WD 値とする.

(3) Att*WD 値手法

Att*WD 値は Attention 値の絶対値と WD 値の符号(正負)をもとに計算し, Att*WD 値が高い単語を根拠語として提示する.図 1 の”バトミントン”の Attention 値が 0.8,WD 値が 0.4 の場合,Attention 値の絶対値が 0.8,WD 値の符号は+なので Att*WD 値は+0.8 となる.WD 値の符号が-の場合に-0.8 となる.

4. 提案手法

Attention 値手法や Att*WD 値手法は提示された単語に助詞や助動詞などの意味のない単語が多く含まれている場合が多い.WD 値手法は入力文の時系列情報が欠落しているため,モデルの推論過程を十分に反映していない可能性がある.

それを解決するために,本論文では WD 値が高い単語の中から Attention 値が高い単語をモデルの推論

A study on Interpretability of Document Classification by BERT
Kazuki Saka ,Kenneth J Mackin,and Yasuo Nagai
‡Tokyo University of Information Sciences

の根拠として提示する手法を提案する.

本提案手法では，まず,入力文の単語を WD 値が高い順にソートする.次に,その配列の WD 値が高い上位 10%の中から Attention 値が高い単語を根拠として提示する.図 2 と表 1 は livedoor ニュースコーパスのスポーツウォッチの記事の例と,それを分類したときに各手法が提示した単語を示している.

バドミントン混合ダブルスで、ロンドン五輪出場を目指す、潮田玲子&池田信太郎の通称"イケシオ"ペア。潮田と小椋久美子の"オグシオ"人気を引き継ぐ格好で北京五輪の翌年 2009 年に大きな注目のもと結成が発表されるも、以後、全日本選手権では 2 年連続で優勝を逃すなど、国内最強ペアと呼ばれながらも結果を残せず苦しい戦いが続いた。・・・

図 2: 入力文の例

手法	単語
提案手法	、,転 機,な ど,続 い,以 後,語 っ,混 合,優 勝,全 日 本,昨 年
Attention	。,た,、,そ ん な,は,伝 え,二,人,も,れ る
WD	以 後,続 い,翌 年,北 京,出 場,ダ ブ ル ス,残 せ,結 果,昨 年,優 勝
Att*WD	。,た,、,そ ん な,は,伝 え,人,も,れ る,だ が

表 1:各手法が提示する単語

5. 提案手法の評価と考察

本提案手法の評価は，次の分類スコアの減少量と話題を想起させる単語数を用いて行う.評価に使用したデータは livedoor ニュースコーパスの独女通信,ムービーエンター,Smax,スポーツウォッチの 4 つのカテゴリから 10 件ずつランダムに選択した.

(1) 分類スコアの減少量
この評価指標[4]では，それぞれの根拠提示手法が提示した単語を’[MASK]’というトークンに変換し,未変換の文章の分類スコアとの減少量とした.表 2 は 4 つのカテゴリの分類スコアの減少量の合計を示したものである.

表 2:分類スコアの減少量

手法	減少量
提案手法	4.09
Attention	5.72
WD	3.41
Att*WD	5.72

(2) 話題を想起させる単語数
話題を想起させる単語とはその単語を提示することでカテゴリの話題を連想させる単語である.例えば,スポーツカテゴリの話題を想起させる単語は’野球’,’試合’,’選手’があげられる.これらの単語数を人手により評価し,評価指標とした.その結果が表 3 であり，Attention 値に寄与している Attention 値手法や Att*WD 値手法と，WD 値に寄与している WD 値手法や提案手法の間で大きな差があることが確認できた.

表 3:話題を想起させる単語数

手法	単語数
提案手法	117
Attention	18
WD	118
Att＊WD	18

2 つの評価指標の結果から Attention 値に寄与している手法である Attention 値手法と Att*WD 値手法は分類スコアを大きく減少させているが，提示する単語はカテゴリの話題を想起させづらいことが分かった.これに対して，WD 値に寄与している WD 値手法と提案手法は，話題を想起させる単語数は多く,分類スコアの減少量は Attention 値手法や Att*WD 手法ほど大きくならかった.

さらに,提案手法は話題を想起させる単語数が最も多い WD 値の 118 語に次ぎ 117 語である.分類スコアの減少量は WD 値手法を上回っている.これらのことから提案手法が提示する根拠語の妥当性は高いと考えられる.

6. おわりに

我々は BERT の文書分類における根拠提示手法の検討を行った.従来の根拠提示手法では，2 つの問題点が明らかになった.1 つ目は提示される単語に意味のない単語が含まれる点，2 つ目はモデルの推論過程を十分に反映していない点である．本提案では，WD 値が高い単語の中から Attention 値が高い単語を提示することで，根拠として妥当性の高い単語の提示手法を示すことができた.

参考文献

[1]Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv:1810.04805, (2018).

[2]Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, Mark Neerincx, ”Evaluating XAI: A comparison of rule-based and example-based explanations”, Artificial Intelligence, Vol.291 103404 (2021).

[3]"BERT models for Japanese text", https://github.com/cl-tohoku/bert-japanese (accessed 2023/1/06).

[4]為栗敦生 高橋良颯 山口実靖, “BERT における文書分類の判断根拠の提示に関する一考察”,情報処理学会研究報告 Vol.2022-NL-252 No.2 (2022).