BERTによる文書分類の根拠提示手法の評価

総合情報学科　j19191 永井ゼミ　坂和樹

研究背景

- BERTによる文書分類の過程はブラックボックス
- 根拠提示手法を確立することでモデルに説明性を付与することができる

- Evaluating XAI: A comparison of rule-based and example-based explanations　Jasper van der Waa a,b,*, Elisabeth Nieuwburg a,c, Anita Cremers , Mark Neerincx

文書分類とは

- 学習したモデルに分類したい文章を入力し、あらかじめ設定したカテゴリにモデルが分類する

- USB3.0対応で爆速データ転送！　9倍速のリーダー/ライター登場

ITカテゴリ

判断根拠とは

- 文書分類においてその判断の根拠となる単語
- モデルはどの単語をもとにして分類しているのか

判断根拠とは

- USB3.0対応で爆速データ転送！　9倍速のリーダー/ライター登場

　USB3.0が登場してから今年で４年目となるがパソコン側でのUSB3.0ポート搭載が進んでも対応機器がなかなか充実していない現状がある。

USBやパソコンなどの単語が多いからITカテゴリ

従来研究

- 為栗ら「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.2-2022/6/29

- BERTにおける文書分類タスクでどの単語が重要であったかを提示する手法を提示して性能評価している
- 提案された手法が適切に判断根拠を提示できているという結果

従来研究の手法

- Attention値法
  - 入力文に対して各単語のAttention値を計算し、それが大きい単語を分類根拠として提示する。
- WD値手法
  - 入力文内の一語を分類モデルに入力し、その分類スコアが高い単語を分類根拠として提示する。

- Att*WD値手法
  - Attention値を大きさ、WD値を正負とした値を計算し、それが大きい単語を分類根拠として提示する。

Attention値とは

- BERTモデルが計算した入力文のどの部分を分類において重要だとしているか表す値

- ライブラリに実装されている機能を利用してモデルが計算したAttention値を取り出すことができる

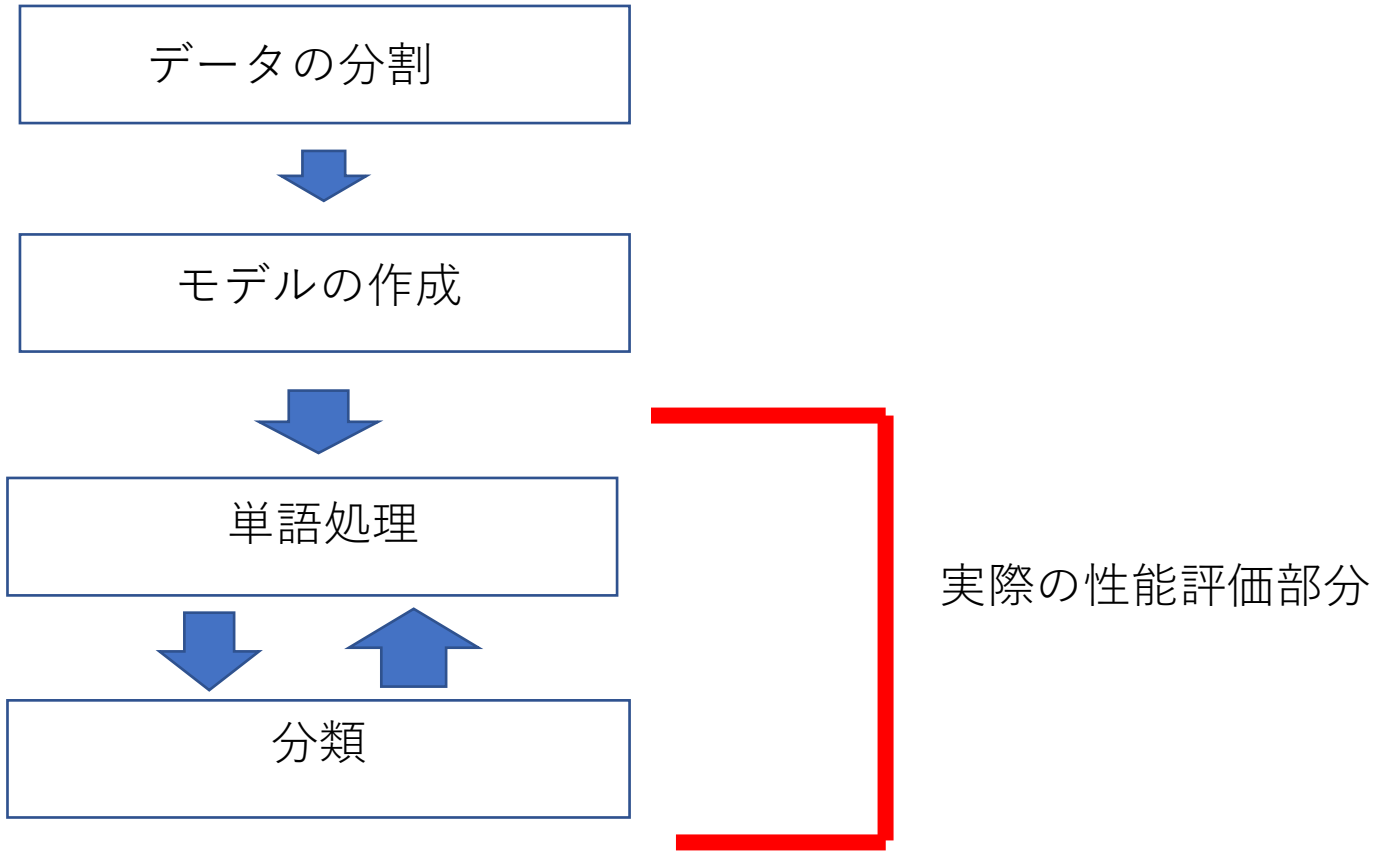
新たに提案する手法

- 手法1
  - 入力文を句点「。」で区切ってモデルに入力し、Attention値が高い単語を分類根拠とする。
- 手法2
  - モデルの分類スコアが高い単語の中からAttention値が高い単語を選択し、分類根拠とする。

実行環境

- フレームワーク:Pytorch
- BERT:　事前学習モデルは東北大学の日本語BERTモデルを使用
- 開発環境：Anaconda jupyter notebook

性能評価の流れ



実験の流れ

- １回目：プレーンな文章を入力
- ２回目：１単語を処理した文章を入力
- ３回目：２単語を処理した文章を入力

⋮

- 分類スコアの変化を記録する
- 単語に施す処理はマスク化と削除の２種類

２回目に入力する文章

- [MASK]対応で爆速データ転送！　9倍速のリーダー/ライター登場

　[MASK]が登場してから今年で４年目となるがパソコン側での[MASK]ポート搭載が進んでも対応機器がなかなか充実していない現状がある。

マスク化した単語：USB3.0

３回目に入力する文章

- [MASK]対応で爆速データ転送！　9倍速のリーダー/ライター登場

　[MASK]が登場してから今年で４年目となるが[MASK]側での[MASK]ポート搭載が進んでも対応機器がなかなか充実していない現状がある。

マスク化した単語：パソコン

評価指標

- 単語の処理語数ごとに減少する分類スコアの速さ
- 速ければその手法の性能が良い

- 分類スコア：大きいほどモデルが確信をもって判断している

実験の内容　テキストの一例

500MB以上アップロードしたユーザーには、無料容量が500MB分プラスされるという得点もある。

無料容量の上限は5GBだが、通常の使い方では2GBなので、使ってみる価値はある。■Skydrive、iOS版が

登場 Microsoftの「Skydrive」は、iOS版クライアントがリリースされた。機能はファイルの閲覧が主

だが、ファイルの共有を設定する機能もある。ただ、ファイルが存在するのに「移動または削除されています」

と表示され、ファイルを閲覧できないこともある。このような場合には,「リンク」ボタンから「更新」を

選び、キャッシュを更新するとよい。

実験の内容　WD値手法が提示した単語

- ファイル、アップデート、機能、実現、クライアント、ストレージ、##op、容量、ユーザー、デジ

- 左の単語からテキストから除外することでテキストの情報量を減少させていく

実験の内容　２回目の入力

- 500MB以上アップロードしたユーザーには、無料容量が500MB分プラスされるという得点もある。

無料容量の上限は5GBだが、通常の使い方では2GBなので、使ってみる価値はある。■Skydrive、iOS版が

登場 Microsoftの「Skydrive」は、iOS版クライアントがリリースされた。機能は[MASK]の閲覧が主

だが、[MASK]の共有を設定する機能もある。ただ、[MASK]が存在するのに「移動または削除されています」

と表示され、[MASK]を閲覧できないこともある。このような場合には,「リンク」ボタンから「更新」を

選び、キャッシュを更新するとよい。

実験の内容　10回目の入力

- 500MB以上アップロードした[MASK]には、無料容量が500MB分プラスされるという得点もある。

無料[MASK]の上限は5GBだが、通常の使い方では2GBなので、使ってみる価値はある。■Skydrive、iOS版が

登場 Microsoftの「Skydrive」は、iOS版[MASK]がリリースされた。[MASK]は[MASK]の閲覧が主

だが、[MASK]の共有を設定する[MASK]もある。ただ、[MASK]が存在するのに「移動または削除されています」

と表示され、[MASK]を閲覧できないこともある。このような場合には,「リンク」ボタンから「更新」を

選び、キャッシュを更新するとよい。

実験の内容

- マスク処理と削除の2パターンで実験する
- 5つの手法　×　2パターン　で10通りの値を評価する
- テストデータは200件

研究結果

手法	最初のスコア	最終スコア
Attentionマスク	10.65	7.90
Attention削除	10.65	9.10
WDマスク	10.65	8.58
WD削除	10.65	9.18
Atwdマスク	10.65	7.90
Atwd削除	10.65	7.85
提案手法１マスク	10.65	8.58
提案手法１削除	10.65	8.63
提案手法２マスク	10.65	8.03
提案手法２削除	10.65	8.06

- Atwd手法が最もスコアがスコアが低くなり、それに次いでAttentionマスク、提案手法２がスコアが低かった

実験結果　提示されて単語

- 従来研究の数値的な評価指標だけでなく各手法が提示した単語の内容にも着目した

実験結果　提示された単語

- データの本文の一例

- 9日深夜、フジテレビ「すぽると！」では、リーグ3位からクライマックスシリーズを勝ち上がり、日本シリーズを制した千葉ロッテマリーンズをフューチャーした。二試合連続で延長までもつれたシリーズ第6戦、第7戦より「10時間39分の死闘の分岐点」と題し、ロッテ主力メンバーが試合を振り返った同特集。センターを守り、決勝打を放った岡田幸文は、「今日の9回ですね。9回表、何かあるなと思いました、正直。1点勝ち越して、9回守りにいくぞと。このままじゃ終われないなというのは、センターを守っていたときに、和田さん（中日のトップバッター）のタイミングの取り方で気づきました」という。事実、その和田にスリーベースを許し、延長戦へ突入することになったが、レフトを守る清田育宏は、「最終回、和田さんの打球を僕と岡田さん、二人で追ってしまって、ツーベースで止めないといけないところをスリーベースにしてしまったので、（小林）宏之さんにはすいませんと」と語った。

研究結果　提示された単語

- 各手法が提示した単語
- Attention値
  - '[SEP]', '。', 'た', 'し', 'は', '、', '日', 'で', '深夜', '##チャー'
- WD値
  - '放っ', '中日', '試合', '試合', 'ロッテ', '制し', 'ところ', '戦', 'に', '勝ち'
- Atwd値
- '[SEP]', '。', 'た', 'し', 'は', '、', '日', 'で', '深夜', '##チャー'

研究結果　提示された単語

- 提案手法1
  - '勝ち越し', 'は', 'まし', 'と', 'た', '正直', '9', '放っ', '回', 'し'
- 提案手法2
  - 'ロッテ', '戦', '中日', '試合', '語っ', '6', '主力', '放っ', '守り', ‘に’

考察

- 本研究で定義した評価指標でいえばAtwd値手法が最も優れた根拠提示手法だった
- 提示した単語を見てみると、WD値手法や提案手法2が提示する単語が分かりやすい
- 提示した単語の内容を加味した評価指標が必要なのではないか

参考文献

- 「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.2-2022/6/29
- Evaluating XAI: A comparison of rule-based and example-based explanations　Jasper van der Waa , Elisabeth Nieuwburg a,c, Anita Cremers , Mark Neerincx