E-044

BERT を用いた分類モデルによる冗長な質問文の要点抽出
Extraction of main point from lengthy questions with classification model using BERT

中山 研一朗† 正田 備也†
Kenichiro Nakayama Tomonari Masada

1. はじめに

近年 AI とくに自然言語処理を用いたチャットボットが世の中に広まっている。幾つかの製品は、あらかじめ質問と回答の組合せを登録しておき、利用者が質問すると、類似した質問を特定し、ペアである回答を表示する分類モデルを用いている。企業内でチャットボットを利用した事例から、利用者による質問の多くは、状況説明や前提条件などを含んでいるが、事前登録した質問では想定しきれず、分類の精度を下げる要因となっている。その為、冗長な質問文から説明などを除き、質問のみを抜き出す仕組みがあれば、チャットボットの分類精度を向上できると考えた。

2. 先行研究

石垣ら[1]は、インターネットサイト上の QA サイトにおいて質問に補足が付与され、要旨の把握が難しいことを指摘し、質問の要約を生成する手法を提案した。しかし、質問と要約の対を獲得するコストが高いものとなっている。山田ら[2]は、様々な文書を 16 種類の情報タイプに分類する手法を提案しているが、今回の課題に対しては複雑なものである。より容易にデータを獲得し、シンプルに構築できるモデルを提案する。

3. 提案内容

3.1 提案モデル

今回提案するモデルは図 1 の通りで、複数の文からなる質問を句点と疑問符というルールで複数の文章に分解する。分解後の各文が、質問か質問以外かを判別する二値分類モデルを、機械学習を用いて構築する。質問と分類された文をチャットボットに渡せば、回答精度が向上すると考える。

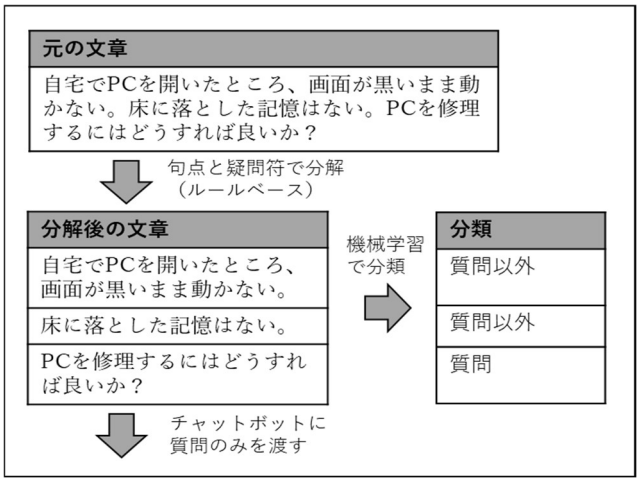

図 1 提案モデルのイメージ

3.2 使用技術

自然言語処理の技術には、BERT や GPT-3、BERT の派生モデルなど、優れた技術が多数ある。日本語の処理において高い精度をあげていることから、NICT が提供する BERT 日本語 Pre-trained モデル（BPE 無し）[3]を採用した。

3.3 使用データ

Yahoo 知恵袋から、最初の質問とベストアンサーを 1 つのペアとして、4 ジャンル（ドラマ、レシピ、株式、不動産）から各 4,000 ペア、合計 16,000 件のペアを取得した。大量のデータに対して、質問か質問以外の分類ラベルを自動的に付与できることから本データを用いた。

ジャンル	最初の質問	ベストアンサー
ドラマ、レシピ、株式、不動産	16,000件	16,000件

自動で分類ラベル付け

テキスト		分類ラベル
最初の質問	16,000件	質問
ベストアンサー	16,000件	質問以外

ホールドアウト法でデータ分割

種別	件数
訓練データ	19,198件（2件UTF-8変換エラー）
検証データ	6,400件
テストデータ	6,400件

図 2 使用データ

4. 実験結果

4.1 Fine Tuning

図 2 の訓練データを用いて、NICT の Pre-Trained モデルに対して、Fine Tuning を行った。検証データによる分類精度は、96.23%、テストデータによる分類精度は 93.56%となった。検証データの分類精度は、BERT 付属のソースコードで算出した値であり、内部的には TensorFlow の tf.metrics.accuracy を用いている。tf.metrics.accuracy は、エポックを跨いで計算される為、分類精度として比較が難しい。テストデータの分類精度は、Fine Tuning 後のモデルで分類予測をして、その結果の正解数÷総数から計算している。この後の比較には、テストデータの分類精度を用いる。

想定より分類精度が高いことから、Fine Tuning 後のモデルが末尾に疑問符が付いている文を質問と分類している可能性を考えた。NICT の BERT Pre-Trained モデルの語彙リスト（vocab.txt）を確認したところ、疑問符は含まれていない。つまり、疑問符は質問かどうかの判断に使われていないことがわかる。

4.2 モデル評価

次に、図 1 の通り分解した文章を、Fine Tuning 後のモデルを用いて、質問か質問以外か分類予測を行う。Yahoo 知

† 立教大学大学院 人工知能科学研究科
Rikkyo University, Graduate School of Artificial Intelligence and Science

恵袋のデータを用いたが、分解することで分類ラベルを自動的に付与することができない。その為、目視により、1000 件分の正解データを用意した。

Fine Tuning 後のモデルによる予測精度は、82.80%となり、Fine Tuning 時の予測精度 93.56%から低下した。複数の文からなるデータを用いて、Fine Tuning を行い、そのモデル評価に、分解後のデータを用いた為と考える。

4.3 Fine Tuning 二回目

Fine Tuning に用いるデータを見直すこととした。具体的には、データを 10 倍に増やした上で、単一文のみを残し、それ以外は捨てることとした。

種別	ドラマ	レシピ	株式	不動産
原文の質問と回答	40,000件	40,000件	40,000件	40,000件
↓				
単一文の質問（最初の質問）	11,201件	5,735件	3,690件	2,279件
単一文の回答（ベストアンサー）	9,518件	390件	5,381件	9,360件

図 3 見直し後の訓練データ

再度、Fine Tuning を行ったところ、テストデータによる分類精度は、98.36%に改善した。一回目の Fine Tuning では、複数の文からなるデータを用いており、質問か質問以外かの判別は難しかったが、単一文ではそれが明確になり、分類精度が向上したと考えられる。

4.4 モデル評価二回目

二回目の Fine Tuning 後のモデルに、一回目の評価と同じデータを用いて分類予測したところ、分類精度は 93.90%となり、大幅に向上した。質問と質問以外に分類するモデルとしては十分と考える。

Fine Tuning 後のモデルが分類を誤ったデータの一部を表 1 に挙げる。文末が「か」や「が」となっている文を質問と捉え、文末が名詞の文を質問以外と捉えている。後者は、疑問符を使わずに判断している為、やむを得ないと考える。

表 1 分類誤りの例

#	データ	予測	正解
1	しかし、ラスト、怒涛の展開！なんで、倉持はベルト締め直すかなあ？	質問	質問以外
2	そんな物件をわざわざ選んで住みたいでしょうか。	質問	質問以外
3	一応、部屋の寸法を測ってからは買おうと思っているのですが。	質問	質問以外
4	「売るも自由」と言っても本当に売れます？	質問以外	質問
5	そんなにあったのかっていうより、あれでこんな高視聴率？	質問以外	質問

5. 文章のベクトル表現可視化

Fine Tuning 後のモデルにおいて、文章のベクトル表現がどのようになっているか可視化を試みた。Google 社が提供する Embedding Projector に、テストデータのうち 300 件を取り込み、PCA で計算して三次元にプロットした。

結果は図 4 の通り、大きく二つに分かれた。左側にプロットされたのが質問、右側が質問以外となっていた。中央付近にある 5 つの点は、数百文字を超える長文であった。質問か質問以外か分類するのが困難であったことがわかる。

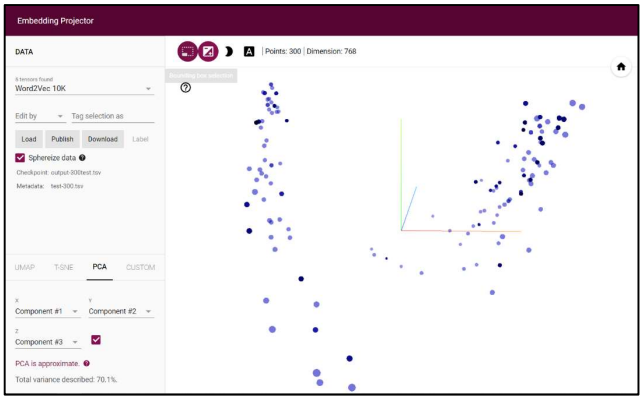

図 4 Embedding Projector の画面ショット

左下にある幾つかの点をクリックしたところ、図 5 の通り、選択疑問文が並んでいた。「芦屋と西宮」、「カツカレーと海鮮丼」のように、単語のベクトル表現としては遠いと思われるものが、文章のベクトル表現としては近いと、Fine Tuning 後の BERT が捉えていることがわかる。

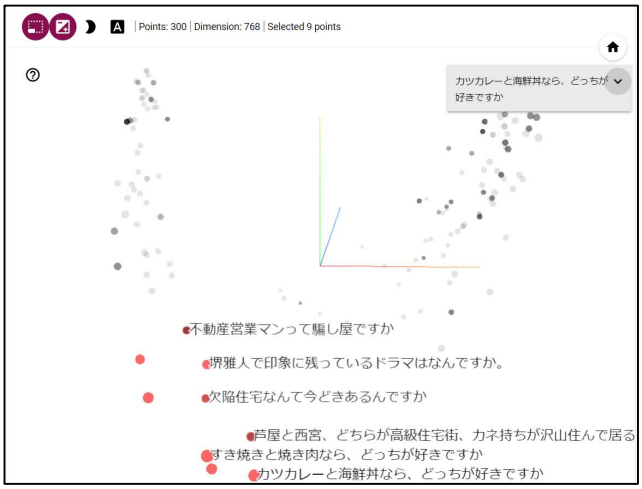

図 5 Embedding Projector の画面ショット

6. おわりに

今回は、質問と質問以外の二値分類を、BERT の日本語 Pre-Trained モデルを用いて構築した。しかし、可視化した結果からわかるように、BERT がオープン質問、選択質問など質問の形式を分類できる可能性を感じた。

参考文献

[1] 石垣 達也, 高村 大也, 奥村 学 “複数文質問を対象とした抽出型および生成型要約” 自然言語処理 2019 年 26 巻 1 号, p.37-58
[2] 山田 侑樹, 櫨山 淳雄, 小川 雄太郎 “OSS プロジェクトの Issue 議論内容に対する BERT および AutoML を用いた文章分類の提案” 人工知能学会第 34 回全国大会(2020)
[3] NICT BERT 日本語 Pre-trained モデル https://alaginrc.nict.go.jp/nict-bert/index.html