E-042

BERT の分散表現を用いた類似度に基づくデータセット拡張の検討

The investigation of similarity-based dataset expansion by variance representation of BERT

春日 優虎† 浦野 昌一†
Yuto Kasuga Shoichi Urano

1. はじめに

学習者の思考力・表現力を効率的に測るための手段として、入学試験等において記述式問題を導入するケースが増えつつあるが、導入に際しては採点コストの増大が課題として挙げられる。

例えば大学入学共通テストにおいても記述式問題を導入する計画が進められており、当初は増大する採点コストへの解決策として外部への採点業務委託を行うことが検討されていた。しかしながら、第三者への情報漏洩等への懸念の声が寄せられ、結果として現在に至るまで記述式問題の導入は見送られているという現状がある。

このように、入学試験などの高い機密性が求められる場面においては、記述式問題の導入によって増大する採点コストを抑えつつ、組織内部で採点業務を完結させる仕組み作りを行う必要がある。本研究ではこの課題を、BERT を用いた自動採点モデルを構築して解決することを目指した。

これまでの研究では、約 2000 件と比較的大規模なデータセットを用いて採点モデルのファインチューニングすることを検討し、結果として 90%以上の精度を得ることができた。しかしながら、実際の採点現場において同規模のデータセットを構築することは容易ではなく、実現可能性が低い。そこで、続く研究では、文章データの拡張手法である Back Translation（逆翻訳）を用いることで、小規模データセットからでも学習可能な採点モデルの構築を行なった。

本稿では、Back Translation による拡張前後の文章間の類似度を定量的に測定し、その結果をもとにデータセットを作成することで、より客観性と精度が高い採点モデルを構築することを考える。

2. BERT による記述式問題自動採点モデル

BERT[1]とは、Google が開発した大規模言語モデル(LLM)の一種であり、Transformer[2]の Encoder モデルの一種である。BERT は Transformer が持つ Attention 機構と呼ばれる仕組みにより文章データから文意を反映させたベクトル(分散表現)を得ることができる。本研究ではこの仕組みを利用し、答案データの文頭トークン[CLS]に対する分散表現を特徴量として入力して得点を出力する分類器を作成することで、自動採点モデルを構築した(図 1)。

本研究では、1 つの設問に対して複数の採点項目を設定し、それぞれの採点項目において個別に自動採点モデルを構築する。モデルに用いる分類器は、正解ラベルと不正解ラベルの 2 値分類器か、そこに部分正解ラベルを加えた 3 値分類器のいずれかとする(主には、採点基準が細分化された採点項目に対しては 3 値分類器を用い、部分正解ラベルに分類された答案に対しては採点者による手動採点が行われることを想定している)。

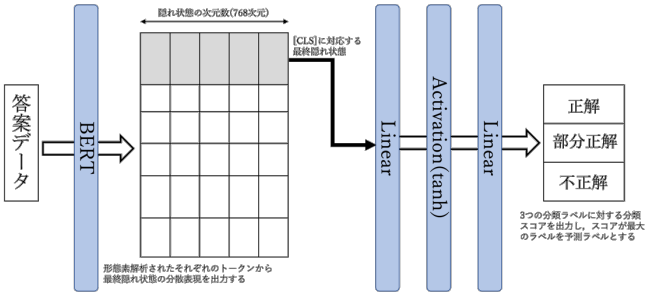

図 1 BERT を用いた自動採点モデルの概略図(3 値分類)

3. Back Translation によるデータセット拡張

ある言語で記述された文章を他の言語へ翻訳し、再び元の言語へ再翻訳すると、文意は同じであるが表現がわずかに異なる文章を得ることができる。Back Translation(逆翻訳)とは、この仕組みを利用した自然言語処理におけるデータセット拡張手法の一種である。

3.1 データセット拡張の有効性

本稿に先立ち、Back Translation によるデータセット拡張手法の有効性の検証を行なった。実際の記述問題における 1 つの採点項目に対する自動採点モデル(正解・部分正解・不正解の 3 値分類モデル)を、オリジナルデータセット 200 件を用いてファインチューニングした場合と、オリジナルデータセット 200 件を Back Translation によって 1000 件に拡張したものを用いてファインチューニングした場合における精度を比較した。

シミュレーションの結果、データセット拡張を行わなかった場合には正解ラベルや不正解ラベルに属する答案のほとんどが部分正解に分類されてしまい、Recall が非常に低くなった。その反面で、Back Translation により拡張を行うことで、大きな偏りを生じさせることなく Recall に向上が見られたため、Back Translation によるデータセット拡張手法には一定の有効性があると考えられる。

3.2 拡張前後の文章の類似度検証

本稿ではさらに精度を向上させるために、Back Translation 前後の文章の類似度を定量的に測定することでデータセットを構築できるようにすることを目指す。類似度の測定には(1)式に示す cos 類似度を用いた。

$$cossim(\boldsymbol{u},\boldsymbol{v}) = \frac{\boldsymbol{u}\cdot\boldsymbol{v}}{|\boldsymbol{u}||\boldsymbol{v}|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} {u_i}^2}\sqrt{\sum_{i=1}^{n} {v_i}^2}} \quad (1)$$

† 明治大学大学院 先端数理科学研究科 ネットワークデザイン専攻
Meiji University Graduate School of Advanced Mathematical Sciences Network Design Program

ここで式(1)において、$\boldsymbol{u}$および$\boldsymbol{v}$はそれぞれ次元数がnのベクトルであるとする。

cos 類似度は 2 つのベクトルのなす角の余弦値を用いたものであり、−1 から 1 までの値を取る。0 から 1 に近づくほど 2 つのベクトルは同じような性質を持つと判断することができ、逆に 0 から−1 に近づくほど 2 つのベクトルは反対の性質を持つと判断することができる。

4. シミュレーション

本稿では、Back Translation 前後の文章の類似性に基づくデータセット形成の有効性を検証するため、主観的に英語との相性が良いと判断した 4 ヶ国語を用いて拡張した学習データを用いてモデルの学習を行なった場合(ケース 1)と、客観的な類似度に基づいて拡張した学習データを用いてモデルの学習を行なった場合(ケース 2)との間で、テストデータに対するモデルの精度の比較を行う。

4.1 使用データセット

理化学研究所が提供する記述問題データセット「Y14_1-2_1_3」[3]を使用した。このデータセットは代々木ゼミナールが 2014 年に実施した模擬試験に対する高校生の解答 2000 件にアノテーションを施したものである。1 つの設問に対して 4 つの採点項目(A〜D)が設けられており、それぞれに対する点数がアノテーションされているが、本稿では採点項目 D のみを扱う。採点項目 D に対しては 0 点〜6 点の点数がつけられているが、これを 0 点、2 点〜5 点、6 点の 3 つに分け、それぞれのラベルを「不正解」、「部分正解」、「正解」とする。また、Back Translation によるデータセット拡張を行うために、日本語の文章で記述されている答案データを英語に翻訳したものを使用する。

4.2 類似度に基づくデータセット拡張

ケース 1 では、オリジナルデータセットから 200 件を無作為に抽出し、フランス語・イタリア語・ドイツ語・スペイン語の 4 ヶ国語においてそれぞれ Back Translation を行い、元データと合わせて 1000 件までデータセット拡張を行う。

ケース 2 では、DeepL API に対応している 13 ヶ国語のそれぞれにおいて Back Translation を行い、その中から元の文章との cos 類似度が高いものから上位 1000 件のみを採用した(図 2)。なお、cos 類似度の測定には、拡張前後の文章データの先頭トークン[CLS]に対する分散表現ベクトルを使用した。

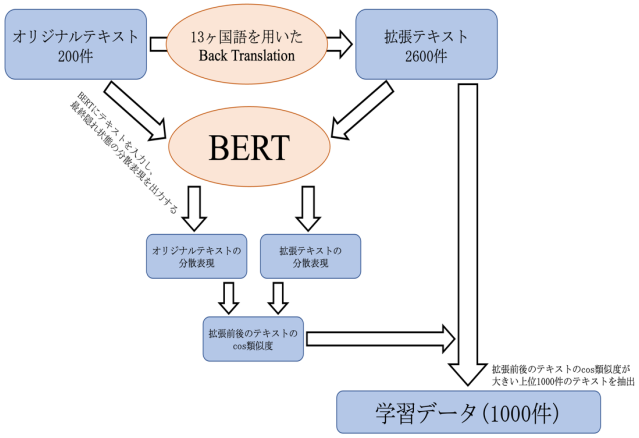

図 2　cos 類似度に基づくデータセットの選択

表 1　逆翻訳の有無によるモデルの精度比較

	データ件数	Accuracy	Precision	Recall
ケース1	特定の 4 ヶ国語を用いた拡張データ 1000件を使用	0.852	0.845	0.779
ケース2	類似度に基づいた拡張データ 1000件を使用	0.829	0.799	0.743

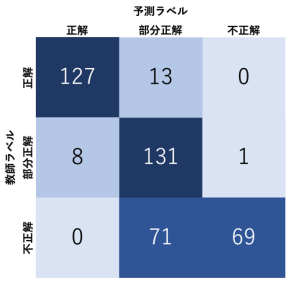

図 3　ケース 1 の混同行列

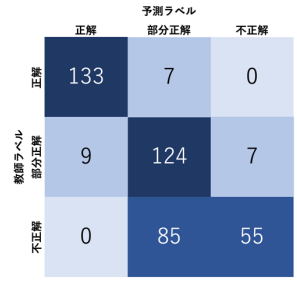

図 4　ケース 2 の混同行列

4.3 ファインチューニングと精度検証

モデルのファインチューニングにおいては、拡張後データセット 1000 件のうち、800 件を訓練用データ、200 件を検証用データとし、学習率 1.0×10^{-5}、最大エポック数 10 回で学習を行い、検証用データに対する損失値が最小のエポックにおけるパラメータを採用する。

最後に、このパラメータにおいて、オリジナルデータ 420 件からなるテスト用データを用いて精度の検証を行う。

4.4 シミュレーション結果・考察

シミュレーションの結果を表 1 および図 3、図 4 にまとめた。表 1 から、直感的に選んだ 4 ヶ国語を用いた Back Translation によるデータセット拡張が、定量的な類似度に基づくデータセット拡張よりも優れていることがわかった。その原因として主に以下の 2 つの可能性が考えられる。

1 つは、類似度が高い文章を優先して選んだことにより、モデルの自由度が小さくなった可能性である。翻訳精度が極めて高い文章が重複してデータセットに含まれることで、過学習が生じた可能性が高い。また、ケース 2 では 13 ヶ国語を用いているため、同じテキストから拡張されたデータが複数選ばれ、不均衡な学習データになった可能性も考えられる。

5. おわりに

本稿では、cos 類似度を用いた定量的な評価に基づく小規模データセット拡張手法についてシミュレーションを行った。結果として、類似度を測らずに構築したデータセットを用いて学習したモデルの方が良い精度が得られることがわかった。しかしながら、考察で述べた通り、本稿におけるデータセット構築手法では過学習や不均衡なデータセットが形成される恐れがあるため、今後も cos 類似度の指標は用いつつ、これらの課題の解消に努めていきたい。

参考文献

[1] Jacob Devlin, Ming-WeiChang, Kenton Lee, Kristina Toutanova, “*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*”, in Proc. NAACL-HLT 2019, pp. 4171–4186 (2019)

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, “*Attention is All you Need*”, arXiv, 1706.03762 (2017)

[3] 理化学研究所, “理研記述問題採点データセット”, 国立情報学研究所情報学研究データレポジトリ (2020)