

センター英語試験の不要文除去問題に対する BERTの適用方法の検討

An Approach for Applying BERT to Sentence Elimination Problem in English Exam

成松宏美^{*1}
Hiromi Narimatsu

杉山弘晃^{*1}
Hiroaki Sugiyama

菊井玄一郎^{*2}
Genichiro Kikui

平博順^{*3}
Hirotoshi Taira

的場成紀^{*3}
Seiki Matoba

東中竜一郎^{*1}
Ryuichiro Higashinaka

^{*1}NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

^{*2}岡山県立大学
Okayama Prefectural University

^{*2}大阪工業大学
Osaka Institute of Technology

We have been working on the English problems in the “Can a Robot Get into the University of Tokyo?” project. This paper focuses on the sentence elimination problem by applying BERT, which has obtained the state-of-the-art results in a number of machine comprehension tasks. We show how we apply BERT and report the improvements made over baselines. Finally, we show our error analysis and the problems that still need to be solved.

1. はじめに

我々は「ロボットは東大に入れるか」プロジェクト [新井 18] において、引き続き英語（特に、センター試験の英語問題）に取り組んでいる。本稿は、センター試験の英語問題で出題される不要文除去問題に対して、近年多くの機械読解タスクにおいて State-of-the-art (SOTA) を達成している汎用言語表現モデル BERT [Devlin 18] を用いた解法について述べる。

不要文除去問題は、文章中に 1 つの不要な文が含まれており、それを取り除くことで全体のまとまりが良くなるような文の一つを選ぶという問題である。図 1 に示す例では、(1) より前の文脈により、(1) 以降に良い靴選びのポイントが提示されると推測できる。ここで、(1)(2)(4) はそのポイントが提示されているものの、(3) はブランドの革靴の話をしており、主題が異なる。よって、(3) が不要文であり、これを選べば正解となる。我々はこの問題に対して、様々な手法を検討してきたが Word2vec [Mikolov 13] を用いて選択肢間の距離を測るシンプルな手法がもっとも良いスコアであったことを報告した [東中 17]。

不要文除去問題は、文同士の類似性だけでなく、文書としての自然さの評価が必要な点で、近年取り組まれている機械読解タスクにはあまり見られない問題である。近年、機械読解タスクにおいて注目を集めている OpenAI GPT [Radford 18] や BERT は、Transformer と呼ばれる自己注意機構を備えたニューラルネットワークを大規模なテキストコーパスを用いて事前学習し、個別の問題に対して転移学習することで、様々なタスクにおいて SOTA を達成している。転移学習により様々なタスクに適用できるため、不要文除去問題においても精度向上が期待できる。

本稿では、BERT を不要文除去問題に適用する方法を検討し、Word2vec に比べ、有意に正解率が良くなったことを示す。また、エラー分析により、BERT により解けるようになった

Wearing proper shoes can reduce problems with your feet. Here are some important points to think about in order to choose the right shoes.

(1) Make sure the insole, the inner bottom part of the shoe, is made of material which absorbs the impact on your foot when walking. (2) The upper part of the shoe should be made of breathable material such as leather or cloth. (3) Some brand-name leather shoes are famous because of their fashionable designs. (4) When you try on shoes, pay attention not only to their length but also to their depth and width. Wearing the right shoes lets you enjoy walking with fewer problems.

図 1: 不要文除去問題の例 (平成 29 年センター英語試験問題より引用。正解は (3)。)

問題とそうでない問題がどのようなものかを示す。

2. BERT による不要文除去問題の解法

ここでは、問題解答に用いる Bidirectional Encoder Representations from Transformers (BERT) [Devlin 18] について説明するとともに、BERT を不要文除去問題に適用する方法を述べる。転移学習を行う際の、学習に用いるデータの作成方法 (2.2) および BERT の入力形式への問題の変換方法 (2.3) と、転移学習を行わずに BERT の後続文予測モデルを用いる解法 (2.4) を説明する。

2.1 Bidirectional Encoder Representations from Transformers (BERT)

BERT は、図 2 (左) に示す Transformer モデル [Alec 18] を大規模なテキストコーパスで事前学習し、個別の問題に対して転移学習する手法である [Devlin 18]。Transformer は、位置情報 (Position embedding) 付きのテキストを入力として、「自分と関係する周辺情報を集約する」機能を持つ自己注意機

連絡先: 成松宏美, NTT コミュニケーション科学基礎研究所,
〒 619-0237 京都府相楽郡精華町光台 2-4, 0774-93-5311,
hiromi.narimatsu.eg@hco.ntt.co.jp

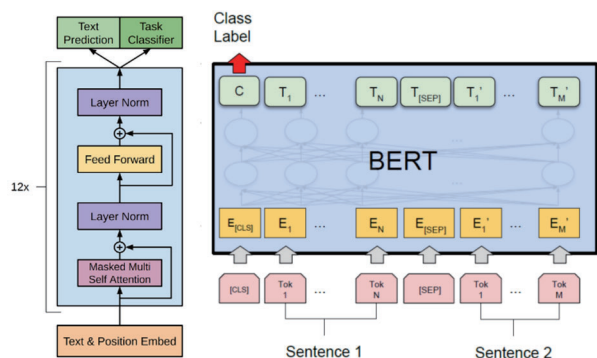


図 2: Transformer モデルの構造 (右) [Alec 18] と個別タスクに対する転移学習時の入出力・モデル構造 (左) [Devlin 18]

構を繰り返し適用することで、タスクに適した特徴ベクトルを計算するモデルである。BERT では Transformer の事前学習にアノテーションが不要である。事前学習のための汎用的なタスクとして、双方向の言語モデルタスクと 2 文の結束性を判定するタスクの 2 種類が採用されている。双方向言語モデルタスクは、文中のマスクされた単語を推定するタスクであり、対象単語の前後の文脈情報を利用する言語モデルを学習することができる。2 文の結束性タスクは、与えられた 2 文が連続する文が無関係な文かを推定するタスクである。文という、単語よりも大きい単位でのつながりの良さを学習できると考えられる。なお BERT では、この 2 文を異なるものとして陽に表現するため、入力情報に文のまとまりを表す segmentID およびそのベクトル表現 (Segment embeddings) を追加し、直接的に文のまとまりを与える工夫がなされている。

2.2 転移学習に用いる擬似問題の作成

BERT の転移学習は、比較的少量のデータでも実行可能ではあるものの、数百では十分な量とは言えない。著者らが保持する不要文除去問題は 249 問であり、転移学習には不十分な分量だと考えられる。そのため、本研究では、既存の長文テキストの一部に不要な文を追加することで擬似的に不要文除去問題を作成し、転移学習に利用するデータの量を増やすことで、解答精度の向上を試みる。

利用するテキストとして、本研究では RACE データセット [Lai 17] の本文部分を用いる。擬似問題を 1 問作成する場合、この本文から連続する 7 文を抜き出して正しい文章とし、7 文以外の本文の箇所からランダムに抜き出した 1 文を不要文とする。得られた不要文を、先頭・末尾以外の箇所に挿入し、図 1 のように、不要な文を 1 文含む擬似問題とする。上記の連続する 7 文を抽出するウィンドウをスライドさせていくことで、RACE データセットの 1 つの問題から、おおよそ 10 問程度の擬似問題が作成できる。最終的に作成できた問題数は 80 万問程度であった。

2.3 4 択の不要文除去問題への適用

本研究では、ある選択肢の要否を判断する 2 値分類器として BERT を学習し、各選択肢について個別に推定された不要らしさを 4 つの選択肢間で比較し、不要らしさが最大であったものを解答として出力するというアプローチを採用する。

また、BERT への入力形式は、選択肢を置く位置や segmentID の値、選択肢近傍の見る範囲など、数種類のパターンが考えられる。以下に、採用した 4 つの入力形式を述べる。

doc-opt 選択肢 (opt) を除く本文 (doc) を文頭側にまとめて並べ、セパレータを挟んで opt を文末側に配置する方法 (図 3)。本文側の segmentID を 0, opt 側を 1 とする。BERT の事前学習の一つである、2 文間の結束性判定を利用したものであり、opt が doc と無関係な文であった場合に、不要であると考えられる。

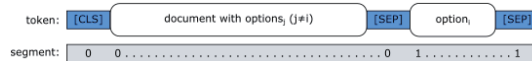


図 3: doc-opt

3opt-opt 4 つの選択肢のうち、対象とする選択肢以外 (3opt) をまとめて文頭に置き、対象とする選択肢 opt を文末に配置する方法 (図 4)。doc-opt 同様、3opt の segmentID を 0, opt の segmentID を 1 として、2 文間の結束性判定を利用して解答する。doc-opt に比べて見る範囲が狭いため、判定に情報が欠落する可能性がある一方、学習を効率的に行える可能性がある。

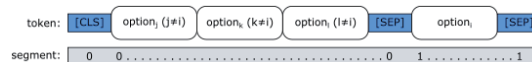


図 4: 3opt-opt

prevN-opt-nextN 対象とする選択肢 (opt) を中心として、前後 N 文ずつを抽出し、出現順通りに並べて配置する方法 (図 5)。opt の segmentID を 1, それ以外を 0 とする。N を無限に大きくした場合は、doc-opt における opt の配置を文中の出現箇所としたものに対応する。position encoding による、出現位置の情報を利用することで、より出現位置に敏感なモデルになると考えられる。



図 5: prevN-opt-nextN

prevN-nextN prevN-opt-nextN のうち、opt を取り除いてその前後のみを利用する方法。opt が含まれない状態で判定するため、opt とその他の箇所との意味的な距離を判定に利用することができず、必要な文が抜けた場合の不自然さ、および正しく不要な文が抜けた場合の自然さ、を利用して判定する必要がある。そのため、opt を利用する他のモデルとは解答傾向が異なることが期待される。

2.4 BERT の後続文予測を用いた手法 (転移学習なし)

BERT は事前学習における目的関数として、穴埋め (cloze test) および後続文予測 (next sentence prediction) の正解率を目的関数としている。事前学習は大量のコーパスを使い大きな計算コストをかけて行っていることから、特に学習データが少ない場合に、対象とする問題を事前学習の目的関数に類似した問題に帰着させることができれば転移学習なしである程度の精度が得られることが期待できる。そこで本研究では、前節までで述べた疑似負例を用いて転移学習を行う方法に加え、事前学習自体を活かし転移学習なしで問題を解く方法を提案する。

不要文除去問題を「各選択肢がその直前の文脈に後続するかどうか判定する問題」と考えると事前学習における後続文

A : <選択肢の左 k 文> [SEP] <選択肢> <選択肢の右 m 文>
 B : <選択肢の左 k 文> [SEP] <選択肢の右 m+1 文>

図 6: 後続文予測への 2 つの入力形式

予測のタスクと同等とみなすことができる。この考えのもと、我々は BERT の事前学習モデルのみによる後続文予測問題として解答を試みる。各選択肢についてそれぞれ独立にその選択肢が直前の文に後続しうるかどうかを判別（二値分類）し、接続しないものを除去すべき選択肢として選ぶのが最も単純な実装であるが、この場合、選択肢が一つのみ選ばれるとは限らない。二値分類の前提となる尤度を用いて尤度最低の（すなわち最も後続性の低い）選択肢を選ぶ方法が考えられるが確率計算のベースが異なるためか予備実験では精度が低かった。

我々は、各選択肢について「その選択肢を除去しない場合」と「その選択肢を除去して次の文に移移する場合」の接続性の良さを求め、この差を「当該選択肢を除去すべきスコア」と考えた。すなわち、一つの選択肢に対して図 6 のような A, B という入力を BERT の後続文予測モデルに与え、出力の尤度値（対数 odds）の差をこの選択肢のスコアとする。なお k, m は実験的に $k = 2, m = 2$ と定めた。各選択についてこのスコアを求め、スコアが最大のものを解答とする。

3. 評価

3.1 実験設定

2014~2019 年のセンター本試験および追試験、予備校の模試に含まれる不要文除去問題 129 問をテストセット、独自に作成した不要文除去問題 120 問を開発セットとして、評価に用いる。独自に開発した問題は、平均的な英語力を持つ人の正解率が 50% 程度になるように難易度を調整している。比較するモデルは、9 種である。ベースラインとして、3 つの選択肢と 1 つの選択肢の距離を Word2vec のコサイン類似度で算出し、距離の最も遠い選択肢を選んだ場合と比較する。prevN-opt-nextN の N は 1, 2, all とする。不要文除去問題は約 7 文程度から構成されていることから、 $N = 3$ とすることは、ほぼ全文使用に等しい。また、prevN-nextN については、文章としての自然さを評価するものであるため、segmentID を全て 0 にした場合も評価する。転移学習のパラメータは、バッチサイズ 32, 最大系列長 512, dropout は 0.1 固定, epoch 数は 4, 学習率は $5e^{-6}$ と $5e^{-7}$ の 2 種類とした。各手法において開発セットで最大正解率時のモデルを用いてテストセットの正解率を評価する。学習データ数は 670,540 で、正例・負例の割合は同じになるようにした。

3.2 結果

各手法の正解率を表 1 に示す。Word2vec で 0.457 だったのに対して、BERT で遷移学習した場合に prevN-nextN (seg0)（選択肢を除いた前文と後文を segmentID 0 で埋め込んだ場合）において最大の 0.612 のスコアが得られ、カイ二乗検定においても有意な向上が見られた ($p=0.0009$)。また、prevN-opt-nextN においては、前後 1 文だけを見るよりも、前後 2 文またそれ以上をみた方が選択肢の要否を正しく判断できることがわかった。これは、人間が問題を解く際にも前後 1 文から対象の文が必要か不要かを判断することが難しいことから、妥当な結果であると考えられる。

また、prevN-nextN については、segmentID を切り替えるよりも、全て 0 で固定した場合の方が高い正解率が得られた。文章としての自然さをそのまま評価する方が適していると考え

表 1: 各手法の正解率。

手法	正解率
(1) ベースライン (w2v)	0.457 (59/129)
(2) Doc-opt	0.543 (70/129)
(3) 3opt-opt	0.558 (72/129)
(4) prev1-opt-next1	0.372 (48/129)
(5) prev2-opt-next2	0.550 (71/129)
(6) prevN-opt-nextN	0.535 (69/129)
(7) prevN-nextN	0.550 (71/129)
(8) prevN-nextN (seg0)	0.628 (81/129)
(9) 転移学習なし	0.512 (66/129)

られる。

次に、手法毎に解けている問題がどのように異なるかを検証する。2 つの手法の正誤関係に対し、カイ二乗検定を行い独立性が棄却されれば、正誤の傾向に関係があることが示される。すなわち、同様の問題に正答する傾向もしくは異なる問題に正答する傾向が示される。さらに残差分析により異なる問題に対しての正答している数に対して特に有意な差が見られれば、異なる問題に正答していることが示せることを利用して、正答の傾向が近いかどうかを見る。本手法を用いて、2 種類の比較を行う。表 2 に Word2vec とそれと同様の特徴を学習すると考えられる Doc-opt および 3opt-opt との比較と、表 3 に BERT の転移学習を行う手法におけるトップモデル間の比較を示す。各値は p 値であり、*印は、多重検定前の有意水準を 0.05 としたとき、ホルム補正および残差分析を行った結果、一方が正答している箇所有意な差が見られた場合に付与した。Word2vec との比較においては、Doc-opt による手法は異なる問題に正答していることがわかった。これは Doc 内での文章の自然さおよび Doc と後続する opt との含意関係が学習された可能性が考えられる。また、転移学習を行う手法におけるトップモデル間の正誤比較より、3opt-opt と prevN-nextN (seg0) に有意な差が見られた。3opt-opt は選択肢間の類似度を、prevN-nextN (seg0) は文章としての自然さをというように異なる点を表現できている可能性がある。これらをうまくアンサンブル学習することができれば、さらに正解率が向上できる可能性があると考えられる。

表 2: W2V と類似モデル間の関係比較。

	(1)	(2)	(3)
(1) ベースライン (w2v)	—	*0.00793	0.04746
(2) Doc-opt	—	—	1.0
(3) 3opt-opt	—	—	—

表 3: BERT 学習モデル間の正誤の関係比較。

	(3)	(6)	(8)
(3) 3opt-opt	—	1.0	*0.02103
(6) prevN-opt-nextN	—	—	*0.02414
(8) prevN-nextN (seg0)	—	—	—

3.3 分析

BERT の適用によって Word2vec からどのような問題が解けるようになり、一方で依然正答できていないかについて、問題毎の正誤の傾向により分析する。ここでは、Word2vec と BERT の適用によって最高スコアが得られた prevN-nextN (seg0) とを比較し（表 1）、BERT でのみ正答した問題、両方で誤答した問題の例を用いて分析する。4 つの選択肢のスコアのうち、

Food can do more than fill our stomachs? it also satisfies feelings. If you try to satisfy those feelings with food when you are not hungry, this is known as emotional eating. There are some significant differences between emotional hunger and physical hunger. (1) Emotional and physical hunger are both signals of emptiness which you try to eliminate with food. (2) Emotional hunger comes on suddenly, while physical hunger occurs gradually. (3) Emotional hunger feels like it needs to be dealt with instantly with the food you want; physical hunger can wait. (4) Emotional eating can leave behind feelings of guilt although eating due to physical hunger does not. Emotional hunger cannot be fully satisfied with food. Although eating may feel good at that moment, the feeling that caused the hunger is still there.

図 7: Word2vec および BERT で誤った問題の例 (2016 年センター試験本試験より引用。正解は (1), BERT は (2) を選択)。

最大値が、他の 3 つの選択肢のスコアと離れているものは自信を持って選択したと考え、そのような問題を分析対象として選出した。尚、Word2vec と prevN-nextN (seg0) の正誤を比較すると両者とも正答は 36 問、BERT でのみ正答は 45 問、Word2vec でのみ正答は 23 問、両者とも誤答は 25 問であり、2 手法の組み合わせオラクルでは他の組み合わせと比較してもっとも高く、スコアは 0.806 (104/129) であった。

図 8 は、word2vec で誤り、BERT で正解した問題である。主題は缶切りの利点についてであり、(1) から (3) は共通してその主題をサポートしているものの、(4) についてはサポートしていない。よって、BERT によって、主題のすり替えによる不自然さをうまく判断できるようになった可能性がある。

続いて、図 7 は、両手法で誤った問題である。(1) より前の文にて Emotional hunger と physical hunger には重要な違いがいくつかあることが述べられているものの、後続する (1) では、共通点が述べられている。このように、論理的なつながりの判定が必要な問題においては、現在の手法では判断できないと考えられる。

One of the most important kitchen tools is the simple hand-operated can opener – the manual can opener. (1) Can openers are needed to open some canned foods, and nowadays many people have easy-to-use electric ones. (2) However, with a manual can opener, even when there is an electric power failure, you can still open cans. (3) Another advantage of a manual can opener is that it will last for years without any maintenance. (4) Recently, even some electric can openers with multiple functions have been getting cheaper. In any event, it is always a good idea to have a manual can opener in your kitchen.

図 8: Word2vec で誤り BERT で正答した問題の例 (2014 年センター試験追試験より引用。正解は (4))。

4. まとめと今後の課題

センター英語試験で出題される不要文除去問題に対し、近年あらゆるタスクで SOTA を出している BERT の適用方法について検討し、比較を行った。これまでの最高得点を得ることが

できた。これは適用の際に用いた擬似負例が有効であったと考えられる。また、選択肢の前の N 文と選択肢の後の N 文を単純に連結した埋め込み方法がもっとも正解率が高くなる (試験問題の正解率 0.628) ことがわかり、効果的な適用方法を明らかにした。また、従来の Word2vec 手法と比較して正誤の傾向を分析したところ、この埋め込み方法によって文章としての自然な流れを判断できるよう学習された可能性が高い。しかしながら、論理的なつながりや飛躍の判定が必要な問題の場合には、誤った選択肢を選ぶ傾向があることがわかった。今後はこの問題を解決するため、全体としてのつながりの良さと局所的なつながりの良さの両方を判断できるようなアンサンブル学習を検討する。合わせて、係り受け関係などとの併用により、論理的な構造が自然さに反映されるような工夫を検討していく。

謝辞

本研究を推進するにあたって、大学入試センター試験問題のデータをご提供下さった独立行政法人大学入試センターおよび株式会社ジェイシー教育研究所に感謝いたします。実験データをご提供くださいました学校法人高宮学園、株式会社ベネッセコーポレーションに感謝いたします。

参考文献

- [Alec 18] Alec, R., Karthik, N., Tim, S., and Sutskever, I.: Improving Language Understanding by Generative Pre-Training, *arXiv:1802.05365* (2018)
- [Devlin 18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv:1810.04805* (2018)
- [Lai 17] Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E.: RACE: Large-scale Reading Comprehension Dataset From Examinations, in *Proc. of EMNLP 2017*, pp. 785–794 (2017)
- [Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, pp. 3111–3119 (2013)
- [Radford 18] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I.: Improving language understanding by generative pre-training, URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language_understanding_paper.pdf (2018)
- [新井 18] 新井紀子, 東中竜一郎 F 人工知能プロジェクト「ロボットは東大に入れるか」: 第三次 AI ブームの到達点と限界 (2018)
- [東中 17] 東中 竜一郎, 杉山 弘晃, 成松 宏美, 磯崎 秀樹, 菊井 玄一郎, 堂坂 浩二, 平 博順, 南 泰浩, 大和 淳司 F「ロボットは東大に入れるか」プロジェクトにおける英語科目の到達点と今後の課題, 2017 年度人工知能学会全国大会予稿集, pp. 2H2–1 (2017)