XLNet を用いたセンター試験英語不要文除去問題の解答とその分析

坂本美帆　松崎拓也
東京理科大学　理学部第一部　応用数学科
1417055@ed.tus.ac.jp　matuzaki@rs.tus.ac.jp

1. はじめに

「ロボットは東大に入れるか」(以下、東ロボ) は、国立情報学研究所を中心とする研究プロジェクトで、人工知能がセンター試験や東京大学の 2 次試験をどこまで解けるのかを明らかにすることを目的とする。2020 年に東ロボ英語チームは BERT や XLNet の利用によってセンター試験英語に対する得点を大幅に向上させた[1]。特に不要文除去問題については 2017 年から 2019 年に出題された問題 15 問にすべて正解したと報告されている。不要文除去問題とは、与えられた 8 文程度の文章から、取り除くことで全体のまとまりが良くなる文を一つ選ぶという問題である。図 1 にセンター試験 2019 年本試験第 3 問 A 問 1 を例として示す。

本論文では、XLNet がどのような手がかりを用いて不要文除去問題を解いているのかを探るために、単語を削除する、あるいは未知語を表す特殊トークンに置き換える等の方法で入力を変更した際のモデルの挙動の変化を調べた。結果として、例えば、人が重要と判断したキーワードを削除しても、正解率が低下しないという結果を得たが、一方で、正解の尤度を最も低下させる単語を選んで特殊トークンに置き換えた場合、問題文のトピックに沿ったごく少数の語の置き換えによって正解の尤度が急激に減少するケースが観察された。

2. 背景：不要文除去問題の解法

この節では杉山ら[1]による不要文除去問題の解法の流れを説明する。まず 10〜20 文からなる文章を用いて、擬似問題を作成する。具体的には、文章から連続する 7 文を抜き出し、抜き出した 7 文以外の範囲からランダムに 1 文を選択し、不要文として 7 文中のランダムな位置に挿入する。そして、これら計 8 文から、挿入した不要文を含むように選択肢とする文を 4 つ選ぶ。

When flying across the United States, you may see giant arrows made of concrete on the ground. Although nowadays these arrows are basically places of curiosity, in the past, pilots absolutely needed them when flying from one side of the country to the other. ① The arrows were seen as being so successful that some people even suggested floating arrows on the Atlantic Ocean. ② Pilots used the arrows as guides on the flights between New York and San Francisco. ③ Every 16 kilometers, pilots would pass a 21-meter-long arrow that was painted bright yellow. ④ A rotating light in the middle and one light at each end made the arrow visible at night. Since the 1940s, other navigation methods have been introduced and the arrows are generally not used today.Pilots flying through mountainous areas in Montana, however, do still rely on some of them.

図 1　不要文除去問題の例　（正解は①）

擬似問題の元となる文章には、RACE データセット[2]の本文を用いている。RACE は中高生向けの英語の読解問題を WEB から収集して作成されたデータセットである。本文の長さが 10 文以上からなる文書を対象として、各文書から少なくとも 1 つ以上の擬似問題を作成し、XLNet の Fine-tune に用いる。

XLNet への入力には、選択肢を除いた計 7 文全てを用いる。入力形式として、最初に[CLS]、選んだ選択肢の位置に[SEP]、最後に[SEP]の特殊トークンを

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

置く。この形式の入力に対する XLNet の[CLS]トークンに対する出力ベクトルと、パラメータベクトルの内積をスコアとする。実際に試験問題を解く際も同様の入力形式によりスコアを算出し、4 つの選択肢のうち、スコアが最大となるものを解答とする。

3. Fine-tuning の方法

杉山らはモデルの Fine-tuning において以下述べる二値分類による訓練を行っているが、本研究では下記の多値分類による訓練も行い結果を比較した。

3.1 二値分類による訓練

まず、訓練データ中の各問題について、正解の選択肢の文（真の不要文）を抜いたテキスト（正例）と残りの 3 つの選択肢のうちランダムに一つを選んで抜いたテキスト（負例）の計 2 つを作り、これらを合わせたものを fine-tuning の訓練データとする。訓練時は正例・負例いずれかを入力し、スコアのロジスティック損失をロス関数として分類用のパラメータベクトルおよび XLNet 本体をチューニングする。

3.2 多値分類による訓練

訓練データの 4 つの選択肢の文をそれぞれ抜いた 4 通りのテキストを作る。XLNet に 4 通りのテキストを（別々に）入力し、それぞれのスコアを算出する。4 つのスコアに softmax を適用し、クロスエントロピーをロス関数として分類用のパラメータベクトルおよび XLNet 本体をチューニングする。

4. 実験設定及び正解率

事前訓練済みモデルとして XLNet 及び BERT の base モデルと large モデルを使用した。いずれも大文字と小文字を区別する cased モデルを用いた。

入力文の最大トークン数は 170 とし、これ以上は切り捨てた。この最大トークン数で、訓練データと開発データの 9 割及びテストデータの全てがカバーできている。XLNet 及び BERT の実装は transformers ライブラリ[3]のものを使用した。勾配法アルゴリズムは AdamW［4］を用い、学習率は10^{-5}とした。1 エポックごとに開発データに対するロスとスコアを計算し、ロスあるいはスコアがそれまでで最良ならばモデルを保存するという手続きを 8 エポック分行い、最後に保存されたモデルをテストに用いた。

擬似問題のうち訓練データとして 638, 940 問を使用し、開発データとして 34, 662 問を使用した。テストデータとして、センター試験の 2014 年～2020 年本試験・追試験から集めた 42 問及び参考書 2 冊[5, 6]から集めた 75 問を使用した。センター試験の問題の総単語数は 6043 語、参考書の問題の総単語数は 7274 語である。

表 1 に二値分類による訓練の結果を、表 2 に多値分類による訓練の結果を示す。最も正解率が高かったのは XLNet の large モデルを用い、多値分類の形で訓練したときのものだった。以降の分析ではこのモデルを使用する。

表 1　二値分類による訓練の結果

	開発データ（擬似問題）	センター試験	参考書
BERT,base	0. 76	0. 59	0. 44
BERT,large	0. 79	0. 69	0. 52
XLNet,base	0. 79	0. 59	0. 52
XLNet,large	0. 84	0. 78	0. 53

表 2　多値分類による訓練の結果

	開発データ（擬似問題）	センター試験	参考書
BERT,base	0. 80	0. 61	0. 65
BERT,large	0. 82	0. 78	0. 64
XLNet,base	0. 85	0. 69	0. 66
XLNet,large	0. 89	0. 80	0. 68

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 3　キーワードを削除した結果

	正解率	
	削除前	削除後
主キーワード（すべて）	0.80	0.88
不要文に含まれる	0.80	0.92
不要文に含まれない	0.82	0.82
副キーワード	0.80	0.80
主キーワードと副キーワード	0.80	0.76

5. 分析

5.1 特定の単語を削除した場合の変化

人が問題を解く上で重要な単語が、モデルでも重要な働きをしているか調べたい。そのために、センター試験の問題を解く上で手がかりとなりそうな単語を人手で（主観的に）決め、その単語を問題文から消去したときの正解率の変動を調べた。

以下、正解の選択肢を選ぶために必要と判断した単語を主キーワード、文章を理解する上で重要と判断した単語を副キーワードと呼ぶ。主キーワードは合計 365 個、副キーワードは合計 312 個あった。例を図 2 に示す。図 2 では正解の選択肢を下線で、主キーワードを赤文字、副キーワードは青文字で示す。

表 3 に結果を示す。元の正解率である 0.80 と比べると主キーワードだけを削除したときは正解率が向上している。不要文に主キーワードが含まれる場合、それを削除することで不要文が不自然あるいは非文法的になることが多い。その結果、不要文以外を取り除いたときの入力のスコアが低くなることで、正解しやすくなっている可能性がある。これは主キーワードがモデルが問題を解く手がかりになっているかどうかと無関係な変化である。しかし、不要文以外のみに主キーワードが含まれる問題 17 問に対する正答率も、主キーワード削除の前後で（向上することもないが）低下しなかった。そのため、不要文を除いた時の文章のまとまりの良さを評価する手がかりとして、モデルは人間が重要と考えるキーワードと異なる部分を使っている可能性がある。

In Japan, there are several ways of transporting goods. Each method has its own advantages and disadvantages. ① Transportation by air, though it can be expensive, is suitable for carrying goods which require speedy delivery. ② Buses can carry many passengers, and they are convenient for daily life. ③ Ships, on the other hand, can carry large quantities at low cost, but it takes much time for them to reach their destinations. Trains can stop only at stations, but their arrival times can easily be estimated. ④ Although trucks cannot carry much compared with trains, they are useful for carrying things from door to door. Such merits and demerits of each method of transportation should be taken into consideration, so the best way can be chosen, depending on the needs.

図 2　主キーワードを削除することで正解した例

5.2 特定の品詞を削除した場合の変化

問題を解く際にどの品詞が重要になっているかを知るために、問題文から特定の品詞の単語を全て削除したときに正解率がどのように変動するかを調べた。形態素解析には TreeTagger[7]を使用し、品詞セットは Penn Treebank のものを用いた。また、削除する品詞のグループとして、等位接続詞（CC）・副詞（RB）・名詞（NN、NNS、NP、NPS 等）・動詞（VB、VBD、VBG 等）・形容詞（JJ、JJR 等）を比較した。これらの品詞の単語を単に消去した場合を DEL、未知語を表す特殊トークンに置き換えた場合を UNK と呼ぶ。表 4 にセンター試験の問題に対する結果を、表 5 に参考書の問題に対する結果を示す。センター試験の問題と参考書の問題のどちらも名詞が最も多く、DEL、UNK いずれの方法で削除した場合も正解率が最も低下した。次に数が多い動詞に関しては、UNK の場合、参考書の問題に対しては正解率が 10 ポイントほど低下したものの、センター試験に対する正解率は逆に向上した。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 4　特定の品詞を削除した場合（センター試験）

品詞（単語の割合）	DEL	UNK
名詞(29.5%)	0.52	0.61
動詞(19.0%)	0.76	0.85
形容詞(9.3%)	0.73	0.73
副詞(4.7%)	0.78	0.83
等位接続詞(2.7%)	0.76	0.76

表 5　特定の品詞を削除した場合（参考書）

品詞（単語の割合）	DEL	UNK
名詞(28.3%)	0.44	0.44
動詞(19.5%)	0.50	0.57
形容詞(8.2%)	0.57	0.60
副詞(6.6%)	0.48	0.52
等位接続詞(2.5%)	0.66	0.66

6. 正解の尤度変化に基づく置き換え

モデルが手がかりとしている単語をさらに特定するために、未知語を表す特殊トークン[UNK]に置き換えたとき、正解の尤度が最も低くなる単語を選び、実際にその単語を[UNK]に置き換えるという操作を、尤度が 0.25 を下回るまで繰り返した。各ステップで選ばれる単語は、モデルが主要な手がかりとしている単語だと考えられる。

センター試験の問題の 1 つに対しこれを行った例を図 3 に示す。太字の単語が[UNK]に置き換えられた単語である。この例では、salting、pepper、ham、Many の順に置換された。このうち最初の 3 語は問題のテーマであるハムないし保存食に直接関係する単語と言える。図 1 に示した問題に対して同じことを行った場合、Pilots（不要文の次の文の最初の単語）、floating（不要文中の単語）の 2 語のみを[UNK]に置き換えた時点で正解の尤度が 0.25 を下回った。この例では、モデルの挙動は極めて少数の単語に依存しているといえる。

One of the most important features in the development of civilization was the preservation of food. Preserving pork legs as ham is one such example.Today, many countries in the world produce **ham**, but when and where did it begin?① **Many** credit the Chinese with being the first people to record **salting** raw pork, while others have cited the Gauls, ancient people who lived in western parts of Europe.② Another common seasoning is **pepper**, which works just as well in the preservation of food. ③ It seems almost certain that it was a well-established practice by the Roman period. ④ A famous politician in ancient Rome wrote extensively about the “salting of hams” as early as 160B.C. Regardless of the origin, preserved foods like ham helped human culture to evolve and are deeply rooted in history.

図 3　正解の尤度変化に基づく未知語への置き換え例（センター試験英語 2020 年本試験第 3 問 A 問 3）

7. おわりに

不要文除去問題においてXLNetが何を手がかりとしているのかを知るために、入力を変更した場合のモデルの挙動を調べた。人が重要と判断したキーワードを問題文から削除する実験では、正解率の低下がほぼ見られなかった。一方で、正解の尤度が最も低下するようにいくつかの単語を選び、未知語を表す特殊トークンに置き換える実験では、問題文のトピックに沿った少数の語の置き換えによって正解の尤度が急激に減少するケースが観察された。その他、品詞ごとに単語を削除する実験の結果から、名詞が手がかりとなっている度合いが大きいという知見を得たが、XLNet が何を手がかりにしているのか、完全に理解したといえるまでにはさらに研究が必要である。今後は、Saliency に基づく方法[10]など、別のアプローチからも手がかりを探していきたい。

謝辞

XLNet による不要文除去問題の解答手法の詳細を教えて頂いた NTT コミュニケーション科学基礎研究所の成松宏美氏に感謝いたします。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

［1］ 杉山弘晃, 成松宏美, 菊井玄一郎, 東中竜一郎, 堂坂浩二, 平博順, 南泰浩, 大和淳司. センター試験を対象とした高性能な英語ソルバーの実現. 言語処理学会第 26 回年次発表論文集, 2020.

［2］Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 785-794, 2017.

［3］Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38-45, 2020.

［4］Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, 2019.

［5］ 小林 功. マーク式基礎問題集英語[不要文選択・発言の主旨]. 河合出版, 2017.

［6］ 佐藤一行. センター試験英語 不要文削除 問題集. デザインエッグ社, 2016.

［7］ Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing, 1994.

［8］Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional trans- formers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the As- sociation for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019.

［9］ Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Advances in Neural Information Processing Systems 32, pp. 5753-5763, 2019.

［10］Shuoyang Ding, Hainan Xu, and Philipp Koehn. Saliency- driven word alignment interpretation for neural machine translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pp. 1-12, 2019.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).