2V-07

マスク言語モデルによる英文空所補充問題の解答能力に関する分析

田中 康介† 吉見 菜那† 梶原 智之† 内田 諭‡ 荒瀬 由紀*
†愛媛大学 ‡九州大学 *大阪大学

1 はじめに

自然言語処理をはじめとする人工知能技術の研究の進展によって、人間の様々な知的活動を計算機上で模倣できるようになりつつある。例えば、2011 年から取り組まれた「ロボットは東大に入れるか？」（東ロボ）プロジェクト[1]では、東京大学などの大学入試問題を題材とするベンチマークが行われ、人間の平均値を大きく上回る成果を得るなど注目を集めた。

東ロボプロジェクト終了後、自然言語処理の分野では BERT[2]などのマスク言語モデルが提案され、様々なタスクにおける活躍が報告されている。本研究では、BERT およびその後続のマスク言語モデル[3-6]を用いて、入試問題のうちの英文空所補充問題の解答能力について分析する。

本稿では、日本および中国の入試問題における評価実験を通して、モデルごとに得意な問題タイプが異なることや、他の自然言語処理タスクにおける性能と空所補充問題における性能が必ずしも相関しないことを明らかにする。

2 マスク言語モデルによる英文空所補充

英文空所補充問題とは、図 1 に示すとおり、一部の単語が空所として隠蔽された問題文と、空所に補充される単語候補の選択肢によって構成される。一般的には、正解選択肢が 1 つと不正解選択肢が 3 つの 4 択形式が採用されている。

本研究では、BERT[2]などのマスク言語モデルを用いて英文空所補充問題を解く。文中の空所部分を特殊トークン[MASK]に置換してマスク言語モデルに入力し、穴埋め確率が最大となる単語を出力する。本研究では、4 択の選択肢の中から穴埋め確率が最大の候補を出力する実際の設定（4 択）と、選択肢を考慮せずに語彙全体の中から穴埋め確率が最大の候補を出力するより難しい設定（全体）の両方で空所補充問題を解く。

An Analysis of the Ability to Answer English Fill-in-the-blank Questions Using a Masked Language Model
Kosuke Tanaka† (h520251x@mails.cc.ehime-u.ac.jp)
Nana Yoshimi† (yoshimi@ai.cs.ehime-u.ac.jp)
Tomoyuki Kajiwara† (kajiwara@cs.ehime-u.ac.jp)
Satoru Uchida‡ (uchida@flc.kyushu-u.ac.jp)
Yuki Arase* (arase@ist.osaka-u.ac.jp)
† Ehime University
‡ Kyushu University
* Osaka University

> The flight was delayed ___ 30 minutes.
> (a) by (b) of (c) on (d) up

図1：英文空所補充問題（金沢工業大学，2020）[1]

3 実験設定

3.1 データセット

英文空所補充問題のデータセットには、我々が人手で収集した日本の入試問題[7]および中国の入試問題[8]の 2 つを使用する。前者は、2017 年から 2021 年までの 5 年間に出題された日本の大学入試の問題の中から無作為に選択した 500 問であり、文法問題（66 問）・機能語問題（195 問）・文脈問題（153 問）・イディオム問題（86 問）の 4 つの問題タイプに人手で分類されている。後者は、CLOTH[2]と呼ばれる中国の入試問題に関するデータセットであり、高校入試の問題（2,341 問）および大学入試の問題（3,172 問）が含まれる。評価指標には正答率を用いる。

3.2 マスク言語モデル

本研究では、マスク言語モデルとして、BERT[3] [2] ・ RoBERTa[4] [3] ・ ELECTRA[5] [4] ・ DistilBERT[6][5]・ALBERT[7][6]の 5 つを使用した。RoBERTa および ELECTRA は BERT の改良版であり、前者は動的マスク処理、後者は Discriminator の導入などの工夫によって、多くの応用タスクにおいて BERT よりも高い性能を達成している。DistilBERT および ALBERT は BERT の軽量版であり、前者は知識蒸留、後者はパラメタ共有によって、品質を保持しつつ BERT よりもモデルサイズを削減している。各モデルは、Hugging Face Transformers にて公開されている Base サイズの英語の事前学習済みモデルを使用した。

前処理として、各モデル専用のサブワード分割器を用いてトークナイズし、512 トークンを超える長さの問題は除外した。なお、この前処理で除外された問題の数は、中国の高校入試のうち 6 問および中国の大学入試のうち 60 問である。

[1] https://jcshop.jp/SHOP/18149/list.html
[2] https://www.cs.cmu.edu/~glai1/data/cloth/
[3] https://huggingface.co/bert-base-uncased
[4] https://huggingface.co/roberta-base
[5] https://huggingface.co/google/electra-base-generator
[6] https://huggingface.co/distilbert-base-uncased
[7] https://huggingface.co/albert-base-v2

表 1：日本の大学入試問題での実験結果

	文法		機能語		文脈		イディオム	
	4 択	全体	4 択	全体	4 択	全体	4 択	全体
BERT	0.909	0.409	0.943	**0.764**	0.908	0.392	**0.988**	**0.826**
RoBERTa	**0.955**	**0.515**	**0.944**	0.712	**0.948**	**0.497**	0.965	0.825
ELECTRA	0.909	0.394	0.923	0.713	0.882	0.353	0.977	0.720
DistilBERT	0.924	0.348	0.892	0.605	0.902	0.431	0.906	0.593
ALBERT	0.954	0.394	0.908	0.682	0.928	0.392	0.895	0.581

4 実験結果

4.1 日本の入試問題での実験結果

表 1 に、日本の大学入試問題での実験結果を示す。4 択形式であれば、マスク言語モデルは問題タイプによらず、9 割を超える正答率を達成できた。一方で、選択肢が与えられない設定においては、機能語やイディオムの問題については 7 割程度の正解率を達成できるものの、文法や文脈の問題については 4 割程度まで大きく正答率を低下させてしまうことが明らかになった。

モデル別では、RoBERTa が全体に高い性能を示した。BERT は機能語およびイディオムの問題を得意とすることがわかった。ELECTRA は空所補充問題においては BERT を下回ることが明らかになった。これは、ELECTRA が Discriminator を含めて最適化するために、単語穴埋めのみに特化した事前訓練を行っていないことが原因であると考えられる。DistilBERT および ALBERT の軽量モデルは、機能語およびイディオムの問題に関しての性能低下が見られた。また、文法や文脈の問題に関しては他のモデルと同等の性能を示すものの、全体に選択肢を考慮しない場合に大きく性能を低下させる傾向がある。

4.2 中国の入試問題（CLOTH）での実験結果

表 2 に、中国の入試問題での実験結果を示す。全てのモデルが、大学入試よりも高校入試の問題において高い正答率を示している。このことから、人間にとって難しい問題はマスク言語モデルにとっても難しい問題であることがわかる。

多くのモデルが、4 択形式では 7 割程度、選択肢が与えられない設定では 3 割から 4 割程度の正答率を示した。なお、RoBERTa は一貫して顕著に高い正答率を達成することがわかった。

5 おわりに

本研究では、BERT などのマスク言語モデルを用いて、日本および中国の入試問題における英文空所補充問題の解答能力を分析した。実験の結果、現在のマスク言語モデルは、4 択の選択肢が与えられる設定では約 7 割、選択肢が与えられない設定では約 4 割の正答率を達成できることがわかった。問題タイプ別では、機能語およびイディオムに対しては BERT、文法および文脈に対しては RoBERTa が適することがわかった。

表 2：中国の入試問題（CLOTH）での実験結果

	高校入試		大学入試	
	4 択	全体	4 択	全体
BERT	0.769	0.462	0.734	0.333
RoBERTa	**0.871**	**0.572**	**0.826**	**0.419**
ELECTRA	0.735	0.409	0.694	0.279
DistilBERT	0.739	0.385	0.699	0.281
ALBERT	0.757	0.423	0.709	0.300

今後の課題として、マスク言語モデルをファインチューニングし、正答率の変化を調査する。

謝辞

本研究は JSPS 科研費（基盤研究 B，課題番号：JP21H03564，JP22H00677）の助成を受けたものです。

参考文献

[1] 新井紀子, 松崎拓也. ロボットは東大に入れるか？：国立情報科学研究所「人工知能」プロジェクト. 人工知能学会誌, Vol.27, No.5, pp.463-469, 2012.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. of NAACL, pp.4171-4186, 2019.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimizer BERT Pretraining Approach. arXiv:1907.11692, 2019.

[4] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proc. of ICLR, 2020.

[5] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In Proc. of EMC2-NIPS, 2019.

[6] Zhenzhong Lan, Mingda Chan, Sebastian Goodman, Kevin Gimple, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proc. of ICLR, 2020.

[7] 吉見菜那, 梶原智之, 内田諭, 荒瀬由紀, 二宮崇. 問題タイプを考慮した英単語穴埋め問題の不正解選択肢の自動生成. 言語処理学会第 29 回年次大会, 2023.

[8] Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. Large-scale Cloze Test Dataset Created by Teachers. In Proc. of EMNLP, pp.2344–2356, 2018.