JCommonsenseQA 2.0: 計算機と人の協働による常識推論データセットの改良

栗原健太郎 [1]　河原大輔 [1]　柴田知秀 [2]
[1] 早稲田大学理工学術院　[2] ヤフー株式会社
{kkurihara@akane., dkw@}waseda.jp　tomshiba@yahoo-corp.jp

概要

計算機モデルの言語理解能力のさらなる向上に向けて、ベンチマークを改良し、より高度な言語理解能力を測ることができるようにする必要がある。本研究では多肢選択式の常識推論データセット JCommonsenseQA [1] に焦点をあて、計算機と人の協働によってデータセットを改良する。このデータセットの計算機による精度はすでに 90%を越えており難易度が低いため、常識推論能力を測るには適切ではない。難易度が低い要因の一つとして、誤り選択肢群の中に正解とあまり関連がない選択肢が含まれていることが挙げられる。この問題に対処するため、まず正解と類似している誤り選択肢をテキスト生成モデルで自動生成し、次に生成された誤り選択肢候補の中からクラウドソーシングで適切な誤り選択肢を選択することによって、難易度の高いデータセットを構築する。実験の結果、構築したデータセットは元のデータセットよりも難易度が高くなっていることを確認した。

1　はじめに

高性能な言語理解モデルを開発するためには、言語理解の能力を様々な観点から評価し分析するためのベンチマーク (データセット群) が必要である。英語においては、GLUE (General Language Understanding Evaluation) [2] が構築、公開されている。GLUE である程度の高スコアを達成できる言語理解モデルが開発されると、より難易度の高いベンチマークとして SuperGLUE [3] などが構築され、ベンチマーク構築と言語理解モデル開発の好循環が生まれている。

このような英語における言語理解研究活性化の潮流に乗じて、世界中の各言語におけるベンチマーク構築が進んでいる。日本語については、言語理解ベンチマーク JGLUE [1] を我々が構築した。それに伴い、多くの計算機モデルが JGLUE 上で評価されている。英語の GLUE と同様に、多くのタスクで計算機モデルが人間を越える、もしくは、人間に近い精度を達成している。

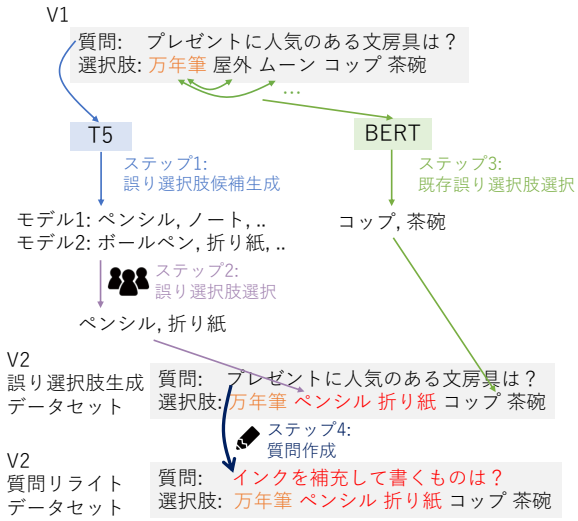

図 1　JCommonsenseQA の改良フロー

日本語においても、計算機モデルの言語理解能力のさらなる向上に向けて、ベンチマークを改良し、より高度な言語理解能力を測ることができるようにする必要がある。本研究では、言語理解に直接的な関係があると考えられる質問応答 (QA) データセットである JCommonsenseQA (JGLUE に含まれているデータセットのうちの一つ) にフォーカスし、より難易度が高いデータセットに改良する。

JCommonsenseQA は、CommonsenseQA [4] の日本語版データセットであり、常識推論能力を評価するための 5 択の QA で構成されている。人間の精度は 98.6%であるが、汎用言語モデル RoBERTa [5] (large) は 90.7%の精度であり、すでに十分に高い精度を達成している。モデルにとって解きやすくなっている要因の一つとして、誤り選択肢群に正解とほとんど関連がないものが存在することが挙げられる。図 1

の上の例では、正解選択肢の「万年筆」と「屋外」や「ムーン」はほとんど関連がない。このような問題はモデルによって正解が簡単に弁別できてしまうため、常識推論能力を測るには適切ではないと考えられる。

本論文では、計算機と人の協働によって JCommonsenseQA を改良する手法を提案する。具体的には、まず正解と類似している誤り選択肢をテキスト生成モデルで自動生成し、次に自動生成された誤り選択肢候補の中からクラウドソーシングで適切な選択肢を選択する。実験の結果、構築したデータセットは以前構築したデータセットよりも難易度が高くなっていることを確認した。

2 関連研究

英語における SQuAD [6] や CommonsenseQA などの QA データセットにおいて、高性能な言語理解モデルの開発とともに、より難易度の高いデータセットを構築する流れが生まれている。

SQuAD は段落、質問、答えの 3 つ組から成る QA データセットで、答えとなるスパンを段落から抽出するタスクである。 SQuAD 1.1 においては段落内に答えが必ず存在するという仕様であった。後に構築された SQuAD 2.0 [7] では、段落内に答えが必ずしも存在しないという仕様に変更されたことで、SQuAD 1.1 より難易度の高いデータセットとなっている。また、異なるタイプの難易度の高い QA データセットとして、マルチホップ推論の能力を測る HotpotQA [8] や、算術演算の能力を測る DROP [9] などが提案されている。

敵対的なデータセットの構築による難易度向上の流れも存在する。Jia ら [10] は SQuAD について、質問の答えにはなっていないものの関連している文を段落の最後に追加することで、敵対的な SQuAD を構築している。多肢選択式 QA の CommonsenseQA についても、ゲーミフィケーションを用いた model-in-the-loop を実施することによって敵対的でより難易度の高い CommonsenseQA 2.0 [11] が構築されている。しかし、選択肢は yes / no の 2 択形式に簡略化されている。

T5 [12] や GPT-2,3 [13, 14] などのテキスト生成モデルはデータセット構築にも利用されている。Liu ら [15] は、自然言語推論データセット MultiNLI [16] を基に敵対的データを獲得し、GPT-3 を用いてデータ拡張することによって WANLI を構築している。

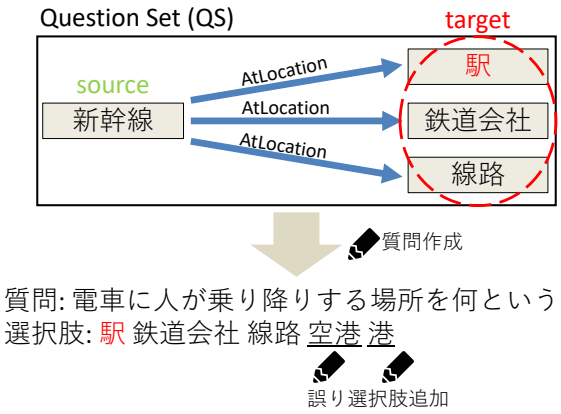

図 2 JCommonsenseQA v1 の構築フロー

3 JCommonsenseQA の構築方法と問題点

提案手法の説明をする前に、本節では、以前提案した JCommonsenseQA の構築方法とその問題点について説明する。以降、以前提案したバージョンを v1、本論文で提案するバージョンを v2 と呼ぶ。

JCommonsenseQA は CommonsenseQA の日本語版データセットであり、常識推論能力を評価するための 5 択の QA で構成されている。JCommonsenseQA は、知識ベース ConceptNet [17] をシードとし、クラウドソーシングを用いて構築している。ConceptNet は、2 つの概念 (concept) と、その間の関係 (relation) を表す 3 つ組からなる多言語知識ベースである。3 つ組は方向性を持ち、例えば新幹線が駅に存在するという関係は (新幹線, AtLocation, 駅) のように (source concept, relation, target concept) として表される。逆方向の関係性を持つ知識は (source concept, relation^{-1}, target concept) と表す[1)]。

構築方法 JCommonsenseQA v1 の構築フローを図 2 に示す。まず、source concept とそれに対して同じ関係を持つ 3 つの target concept の集合 (以下 Question Set, QS と呼ぶ) を収集する。次に各 QS 内の 3 つの target concept について、それぞれの target concept のみが正解となる質問の作文タスクをクラウドソーシングを用いて実施する。最後にそれぞれの質問に対して、誤り選択肢を 2 つ追加するタスクをクラウドソーシングを用いて実施する。

問題点 JCommonsenseQA v1 には、誤り選択肢に正解とほとんど関連がないものが含まれているという問題がある。この問題が生じる原因として、

1) 例えば、駅に新幹線が存在するという関係は (駅, AtLocation^{-1}, 新幹線) と表す。

ConceptNet を用いた選択肢の自動獲得が挙げられる。5 つの選択肢のうち 3 つは、ConceptNet から収集した QS 内の 3 つの target concept に由来しているが、これらの target concept 間に関連性がほとんどない場合がある。例えば、「綺麗」と “HasProperty^{-1}” の関係を持つ target concept として「万年筆」「屋外」「ムーン」の 3 つが抽出される[2]が、これらはほとんど関連していない。

4 JCommonsenseQA の改良

JCommonsenseQA v1 の問題に対処するため、テキスト生成モデルとクラウドソーシングによるフィルタリングを用いて、正解と類似した誤り選択肢を新たに導入する。提案手法は、図 1 に示すとおり、4 つのステップからなる。なお、各ステップにおいて事例がフィルタリングされることで最終的に獲得できる事例数が少なくなってしまうことから、v1 と同じ構築手法を用いて事前に v1 の拡張を行う。各ステップの詳細を以下に示す。

ステップ 1 テキスト生成モデル T5 を用いて、正解と類似した誤り選択肢候補を生成する。T5 は質問を入力し、正解を出力するようにファインチューニングする。この際、2 種類のテキスト生成モデルを用いる[3]ことによって、多様な誤り選択肢候補を生成する。なお、正解により類似した誤り選択肢候補を得るため、v1 の全件を用いてクローズドな訓練を行う。

2 つのモデルそれぞれから 20 件の生成結果を得る。この生成結果には長い語句や不自然なものが含まれている。これらに対処するため、ConceptNet にエントリが存在しない生成結果を除去する。また、生成結果に正解自体が含まれていれば除去する。この結果、各モデルから 5 件の誤り選択肢候補が得られた事例を次のステップに渡す。例えば図 1 の例では、1 つのモデルから得られた誤り選択肢候補は「ペンシル」「ノート」「鉛筆」「手提げ袋」「消しゴム」である。

ステップ 2 生成された誤り選択肢候補には、正解の同義語などの別解が含まれる可能性が高い。それらを除去するため、クラウドソーシング[4]を用いてフィルタリングを行う。クラウドワーカに、質問と、生成した誤り選択肢候補 5 件に正解を加えた 6 つの選択肢を提示し、正解となりうるものを全て選んでもらう。5 人のクラウドワーカから回答を収集し、1 人も正解として選択しなかった選択肢のうち、生成結果における確率がもっとも高いものを誤り選択肢として採用する。上記の処理を 2 つのモデルの誤り選択肢候補それぞれで実行し、2 つの誤り選択肢を得る。いずれかのモデルの誤り選択肢候補において、5 件中の全ての候補が正解とみなされた場合、その事例は除去し、次のステップに渡さない。

ステップ 3 5 択問題を作成するために、2 つの新しい誤り選択肢以外に誤り選択肢が 2 つ必要である。v1 の誤り選択肢において、正解とほとんど関連のない誤り選択肢が存在する一方で、多くの場合、正解と関連している誤り選択肢も存在している。そのため、v1 の誤り選択肢から正解に類似したものを 2 つ選択する。これには事前学習モデル BERT [18][5]に基づく類似度を用いる。まず、正解を含む 5 つの選択肢のそれぞれを BERT の Embedding 層に入力し、5 つのベクトルを取得する[6]。次に、正解とのコサイン類似度が高い誤り選択肢を 2 つ選択する。

ステップ 2 と 3 によって得られた 4 つの誤り選択肢と正解を用いて選択肢を再構成する。得られた選択肢と v1 の質問をペアにし、人間にとって解答可能かどうかをクラウドソーシングで検証する。10 人のクラウドワーカに回答してもらい、7 人以上が正解した事例のみを採用する。この結果得られるデータセットを **v2-誤り選択肢生成データセット**と呼ぶ。

ステップ 4 選択肢間の類似性が上がったことにより、正解と誤り選択肢間の弁別性が低下する恐れがある。例えば図 1 の例では、「ペンシル」は正解の「万年筆」に近く、人間・計算機の双方が誤答する可能性がある。弁別性を上げるために、新しい選択肢を基に質問の再作成をクラウドソーシングで行う。正解の選択肢のみが正解になるようにクラウドワーカに質問を作成してもらう。作成された質問は、v1 と同様の方法でフィルタリングを行う[7]。

v2-誤り選択肢生成データセットと同様に、人間にとって解答可能かどうかをクラウドソーシングで

2） この source concept と関係からは、「虹」「夕焼け」「夜空」のような良い 3 つ組も抽出される。
3） Hugging Face Hub にある “google/mt5-large” と “sonoisa/t5-base-japanese” の 2 つの事前学習モデルを用いる。
4） Yahoo!クラウドソーシング (https://crowdsourcing.yahoo.co.jp/) を用いた。
5） Hugging Face Hub にある “cl-tohoku/bert-base-japanese-v2” を用いる。
6） 選択肢が複数のサブワードからなる場合はそれらのベクトルを平均する。
7） 付録 A に質問のフィルタリング方法を示す。

表 1　各種モデルの評価結果

モデル	v1		v2 誤り選択肢生成		v2 質問リライト	
	dev	test	dev	test	dev	test
人間	0.986	0.988	0.998	0.997	0.992	0.996
東北大 $\text{BERT}_{\text{BASE}}$	0.808	0.782	0.586	0.571	0.699	0.678
東北大 $\text{BERT}_{\text{LARGE}}$	0.816	0.822	0.648	0.617	0.746	0.736
早稲田大 $\text{RoBERTa}_{\text{BASE}}$	0.840	0.849	0.588	0.551	0.701	0.672
早稲田大 $\text{RoBERTa}_{\text{LARGE}}$	**0.907**	**0.901**	**0.816**	**0.807**	**0.856**	**0.865**

検証する。解答不可能なものを除去した結果得られるデータセットを **v2-質問リライトデータセット**と呼ぶ。

以上の手順により、拡張 v1 の 27,400 件から 13,117 件からなる v2-誤り選択肢生成データセットと、10,524 件からなる v2-質問リライトデータセットを構築した。2 つのデータセットに共通の事例を抽出し、8,899 件からなる 2 種類の実験用データセットを構築した。これは次節の比較実験で使用する。

5　事前学習モデルによる評価

本研究で構築した v2 データセットが以前構築した v1 よりも難易度が上がっているかを事前学習モデルを用いて評価した。また、v2-誤り選択肢生成データセットと v2-質問リライトデータセットの比較も行った。

5.1　実験設定

広く用いられている事前学習モデルをファインチューニングし、精度を算出した。実験に用いた事前学習モデルの詳細を付録 B に示す。モデルのファインチューニングでは、v1 と同様に、質問と選択肢を連結した多肢選択式問題を解く[8)]。実験に用いたハイパーパラメータを付録 C に示す。また、v1 構築時と同様に、クラウドソーシングを用いて人間による精度を算出し、人間のスコアとモデルのスコアも比較する。

5.2　実験結果・考察

結果を表 1 に示す。まず、v2 の難易度が v1 よりも高くなっていることがわかる。また、質問リライトデータセットの難易度は誤り選択肢生成データセットよりも下がっている。これは、正解と誤り選択肢間の弁別性が上がり、計算機にとっては解きやすい問題となっていることがわかる。例えば図 1 で、誤り選択肢生成データセットでは、システムの出力は「ペンシル」となり誤っているが、質問リライトデータセットではシステムの出力は「万年筆」となって正解できている。また、人間のスコアは全てのデータセットにおいてほぼ同じ高いスコアであることから、v2 が v1 に比べて問題の質が低下しているわけではないことがわかる。

質問:　つゆにつけてズルズルと音を立てて食べる麺類は？
選択肢: コシヒカリ ご飯 担々麺 **そば** ラーメン

質問:　修学旅行でもよく行く赤い色の電波塔は？
選択肢: **東京タワー** 浅草寺 お台場 東京ドーム スカイツリー

図 3　早稲田大 $\text{RoBERTa}_{\text{LARGE}}$ の誤り例（**太字**が正解、下線が不正解の出力）

v1 と v2 の精度を比べると多くのモデルでかなり精度が低下しているが、最も精度の高かった早稲田大 $\text{RoBERTa}_{\text{LARGE}}$ は非常に強力であり、数ポイントしか精度が低下していないことがわかる。その強力なモデルでの誤り例を図 3 に示す。このような問題に正解するにはどのようにして常識的な知識を獲得すればよいのか、今後の事前学習モデルの改良を考える上での手がかりとなればよいと考えている。

6　おわりに

本論文では、計算機と人の協働によって常識推論データセット JCommonsenseQA を改良する手法を提案した。実験の結果、v2 は v1 よりも難易度の高いデータセットであることを確認できた。本データセットの構築によって、日本語において初めてベンチマーク構築と言語理解モデルの改良のサイクルを一周回せたと考えている。v2 は今後、v1 と同じサイト (https://github.com/yahoojapan/JGLUE) で公開する予定である。

今後、さらに難易度の高いデータセットの検討を進めるとともに、別の種類の質問応答データセットの構築にも取り組む予定である。

8)　Hugging Face によって提供されている Transformers ライブラリ (https://github.com/huggingface/transformers) を用いて行った。

謝辞

本研究はヤフー株式会社と早稲田大学の共同研究により実施した。

参考文献

[1] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the 13th LREC**, pp. 2957–2966, Marseille, France, June 2022.

[2] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018.

[3] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in NeurIIPS**, Vol. 32, 2019.

[4] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In **Proceedings of the 2019 Conference of the NAACL-HLT**, pp. 4149–4158, Minneapolis, Minnesota, June 2019.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv, 2019. abs/1907.11692.

[6] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **Proceedings of the 2016 Conference on EMNLP**, pp. 2383–2392, Austin, Texas, November 2016.

[7] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In **Proceedings of the 56th ACL**, pp. 784–789, Melbourne, Australia, July 2018.

[8] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In **Proceedings of the 2018 Conference on EMNLP**, pp. 2369–2380, Brussels, Belgium, October-November 2018.

[9] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In **Proceedings of the 2019 Conference of the NAACL-HLT**, pp. 2368–2378, Minneapolis, Minnesota, June 2019.

[10] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In **Proceedings of the 2017 Conference on EMNLP**, pp. 2021–2031, Copenhagen, Denmark, September 2017.

[11] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In **Thirty-fifth Conference on NeurIIPS Datasets and Benchmarks Track (Round 1)**, 2021.

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **JMLR**, Vol. 21, No. 140, pp. 1–67, 2020.

[13] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in NeurIIPS**, Vol. 33, pp. 1877–1901, 2020.

[15] Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation. arXiv. abs/2201.05955.

[16] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In **Proceedings of the 2018 Conference of the NAACL-HLT**, pp. 1112–1122, New Orleans, Louisiana, June 2018.

[17] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**, AAAI'17, p. 4444–4451, 2017.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the NAACL-HLT**, pp. 4171–4186, Minneapolis, Minnesota, June 2019.

A　JCommonsenseQA 構築における質問のフィルタリング方法

クラウドソーシングによる質問の作成後、正解の文字数やキーワードの指定などでヒントを与えられた難易度の低い質問、及び不適切な形式の質問を除去するために、以下のいずれかに該当する質問を持つ問題を除去している。

- 質問に選択肢の言葉が含まれる
- 例えば「海の対義語を漢字 1 文字で？」など、質問に「〇文字」という文字列が含まれる (〇は数字)
- 質問の文末が「？」ではない

B　事前学習モデルの詳細

実験に用いた事前学習モデルの詳細を表 2 に示す。

モデル名	基本単位	事前学習テキスト
東北大 $\text{BERT}_{\text{BASE}}$ (cl-tohoku/bert-base-japanese-v2)	サブワード (MeCab + BPE)	Wikipedia
早稲田大 $\text{RoBERTa}_{\text{BASE}}$ (nlp-waseda/roberta-base-japanese)	サブワード (Juman++ + Unigram LM)	Wikipedia + CC

表 2　実験に用いた事前学習モデルの詳細。モデル名の丸括弧内は Hugging Face Hub での名称を示す。どちらのモデルも LARGE サイズも使用する。事前学習テキストの CC は Common Crawl を表す。

C　ハイパーパラメータ

実験に用いたハイパーパラメータを表 3 に示す。

表 3　実験に用いたハイパーパラメータ

名前	値
learning rate	{5e-5, 3e-5, 2e-5}
epoch	{3, 4}
warmup ratio	0.1
max seq length	64