Decoder ベースの大規模言語モデルに基づくテキスト生成の自動評価指標

笠原智仁 [1]　河原大輔 [1]
山崎天 [2]　新里顕大 [2]　佐藤敏紀 [2]
[1] 早稲田大学理工学術院　[2] LINE 株式会社
{tomo_k@ruri.,dkw@}waseda.jp
{takato.yamazaki,kenta.shinzato,toshinori.sato}@linecorp.com

概要

テキスト生成の自動評価はタスクの精度を向上させる上で欠かせないものである。本研究では、Decoder ベースの言語モデルの大規模化が進む動向を踏まえ、それらに基づくテキスト生成の自動評価指標を提案する。翻訳評価と意味的類似度計算の 2 種類の日本語のタスクで実験を行い、Encoder ベースの言語モデルと比較して、同等かそれ以上の精度を出すことが可能であることを示す。

1　はじめに

ニューラルネットワークによるテキスト生成モデルは機械翻訳、対話システム、文章要約など、様々なタスクで用いられる。しかし、モデルからの出力はオープンエンドであり、正解も 1 つとは限らないため、生成結果の評価は難しい。人手による評価は精度が高く、よく用いられるが、時間的、金銭的コストが高いという問題がある。そのため、テキスト生成モデルを迅速に開発するためには自動評価が必要不可欠となる。

テキスト生成の自動評価手法として、かつては BLEU [1] や ROUGE [2] などのような、生成テキストと正解とされるテキストとの間の表層的な単語の重複による評価が主流であった。近年では、BERT [3] や BART [4] などの自己教師有り学習モデルの発展に伴い、それらを活用したより精度の高い自動評価手法が提案されている [5, 6, 7, 8, 9, 10]。これらの手法では主に Transformer [11] アーキテクチャの Encoder ベースのモデルや Encoder-Decoder ベースのモデルが用いられるが、Decoder ベースのモデルを利用した手法の研究はなされていない。

しかしながら、自己教師有り学習モデルは Decoder ベースのモデルの大規模化が顕著であり、GPT3 [12]、Megatron-Turing [13]、PaLM [14] などが開発されている。以降、Transformer の Decoder ベースの自己教師有り学習された大規模言語モデルを LLM と呼ぶ。一方で、Encoder ベースのモデルは Decoder ベースのモデルと比較するとそこまで大規模化が進んでいない。

これらを踏まえ、本論文では LLM を用いたテキスト生成の自動評価手法を提案する。翻訳評価と意味的類似度計算 (Semantic Textual Similarity, STS) の 2 種類の日本語のタスクで実験を行い、既存のテキスト生成自動評価手法と同等かそれ以上の精度を出すことが可能であることを示す。LLM を 3 種類の Tuning 手法でそれぞれ学習し、Fine-Tuning よりも低コストな学習手法である LoRA-Tuning [15] において最も高い精度を出すことが可能であることを示す。また、本手法ではテキストの埋め込みも得られるため、テキスト間の意味的比較なども可能であり、応用範囲が広い。

2　関連研究

2.1　テキスト生成の自動評価指標

テキスト生成を自動評価する際に必要となるものは、モデルによって生成されたテキストと正解とされるテキストである。古典的な自動評価指標である、BLEU や ROUGE、METEOR [16]、CIDEr [17] などではこれら 2 つのテキスト間で N-gram がどれほど重複するかに基づいて評価する。これらの手法では N-gram が完全に一致していなければスコアが上がらないため、類義語が含まれている場合でもスコアが上がらないという欠点がある。編集距離に基づいて評価を行う TER [18] なども同様の欠点を持つ。

METEOR については類義語辞書を利用することでこの欠点を克服することを目指したが、文脈を考慮した類義語判定までは行うことができない。

自己教師有り学習モデルの埋め込みを利用することで、文脈を考慮した上で類義語を似ていると判断することが可能となる。BERTScore [5] は生成テキストと正解テキストをそれぞれ BERT によって埋め込み、その類似度でスコアを計算する手法である。BARTScore [6] では Encoder に 1 つ目のテキストを、Decoder に 2 つ目のテキストを入力し、2 つ目のテキストの生起確率に基づいてスコアを計算する。

テキストペアとその類似度ラベルのデータセットを用いて自己教師有り学習モデルを Fine-Tuning することで精度を上げる手法もある。翻訳評価のデータセットで学習したモデルとして BLEURT [7]、COMET [8] などが、STS データセットで学習したモデルとして Sentence-BERT [9] などが挙げられる。SimCSE [10] のように自然言語推論データセットを用いて Contrastive Learning を行うことにより文の埋め込みを学習し、それをテキストペアの類似度計算に利用する手法も存在する。これらの自己教師有り学習を利用する手法は Encoder モデルが用いられることが多く、Decoder モデルを使用する手法の研究はなされていない。

2.2 自己教師有り学習モデルのチューニング手法

自己教師有り学習モデルをタスクへ適応させる手法として最も主流なものは、モデルのパラメータ全てをチューニングする Fine-Tuning である。しかし、大規模なモデルの Fine-Tuning はコストが高いため、一切パラメータをチューニングせずに複数の例 (Prompt) だけをモデルに入力してタスクを解く Few-Shot Learning [12] や、少数のパラメータのみをチューニングする Prompt-Tuning や LoRA-Tuning などの手法が提案されている。Prompt-Tuning は学習によって Prompt を最適化する手法であり、離散的な語彙の中から最適な単語を選択する手法 [19] と、連続的な埋め込みベクトルを用意してそれを最適化する手法 [20, 21, 22, 23, 24] が存在する。LoRA-Tuning は Transformer アーキテクチャの各層に含まれる重み行列に対して階数分解を行ったパラメータを追加し、そのパラメータのみを学習する手法である。

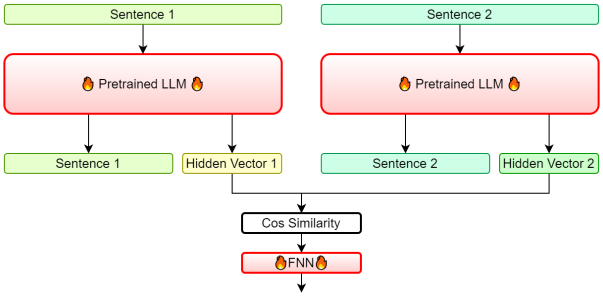

図 1 提案モデルのアーキテクチャと入出力関係

3 提案手法

本論文では LLM を用いたテキスト生成の自動評価指標を提案する。具体的には、テキストペアとその類似度ラベルのデータセットを用いて LLM を LoRA-Tuning することによって自動評価システムを構築する。

3.1 アーキテクチャと入出力関係

提案モデルのアーキテクチャと入出力関係を図 1 に示す。テキストのペアをモデルに入力し、その類似度を出力する。類似度の算出は以下の手順で行う。

1. テキストペアをそれぞれ LLM に入力
2. EOS トークンの 1 つ前の文末のトークンに対応する埋め込みをそれぞれ得る
3. 2 つの埋め込み間の cos 類似度を計算
4. 1 層の FNN 層に cos 類似度を入力し、その出力をテキストペアの類似度とする

なお、1 層の FNN 層を通す理由は cos 類似度の値をデータセットのラベル分布に変換するためである。また、予備実験の結果から EOS トークンではなく文末のトークンの埋め込みを利用することとした。本手法では手順 2 で得られた埋め込みをテキストの埋め込みベクトルとして利用することも可能である。

3.2 学習方法

データセットの正解ラベルはあらかじめ 0〜1 の間に正規化する。3.1 節の手順通りにテキストペアの類似度を算出し、その値とラベルとの間の平均二乗誤差に基づいて、LoRA-Tuning のために新たにモデルに追加したパラメータと FNN のパラメータのみを更新する。なお、FNN の初期値は weight を 1、bias を 0 に設定する。

表 1　自動評価指標による評価と正解ラベルとの相関係数 (WMT20 en-ja)

手法	モデル	Pearson	Spearman	Kendall
BERTScore	waseda RoBERTa-large	0.484	0.459	0.319
BLEURT-20	RemBERT	0.462	0.452	0.315
BERT Fine-Tuning	waseda RoBERTa-large	0.561	0.566	0.396
LLM LoRA-Tuning	rinna GPT2-xsmall	0.528	0.496	0.342
	waseda GPT2-small	0.583	0.554	0.387
	rinna GPT2-small	0.573	0.541	0.378
	rinna GPT2-medium	0.583	0.568	0.396
	HyperCLOVA 6.9B	**0.593**	**0.579**	**0.404**
LLM Prompt-Tuning	rinna GPT2-xsmall	0.459	0.438	0.302
	waseda GPT2-small	0.522	0.502	0.347
	rinna GPT2-small	0.519	0.484	0.336
	rinna GPT2-medium	0.502	0.450	0.312
	HyperCLOVA 6.9B	0.515	0.475	0.329
LLM Fine-Tuning	rinna GPT2-xsmall	0.530	0.507	0.349
	waseda GPT2-small	0.552	0.541	0.377
	rinna GPT2-small	0.556	0.526	0.366
	rinna GPT2-medium	0.581	0.560	0.392
	HyperCLOVA 6.9B	-	-	-

4　実験

実験に使用する GPU は 40GB の GPU メモリを搭載している NVIDIA A100 SXM4 である。

4.1　データセット

実験に使用するデータセットは WMT20 [25] の英語から日本語への翻訳タスク (WMT20 en-ja) のデータセットと、日本語言語理解ベンチマーク JGLUE [26] に含まれる JSTS である。WMT20 のデータセットには人手による翻訳文、機械翻訳モデルによる翻訳文とその評価ラベル (Direct Assessment) が含まれる。JSTS は日本語の STS データセットであり、文ペアとその類似度ラベルからなる。なお、WMT20 についてはデータセットが Train、Valid、Test にあらかじめ分けられていないため、ランダムに 8:1:1 の割合で分割した。

4.2　実験設定

実験には、日本語のコーパスで事前学習がなされた GPT2-xsmall (37M)[1]、small (110M)[2][3]、medium (336M)[4] と、LINE 社が開発した HyperCLOVA [27] と呼ばれる GPT3 ライクなモデルのパラメータ数 6.9B のモデルを用いる。4.1 節の 2 種類のデータセットに対してそれぞれモデルを学習させる。

LoRA-Tuning の他に Prompt-Tuning と Fine-Tuning も比較対象として実験する。Prompt-Tuning ではテキストをトークン列に変換する際、文末のトークンの後ろに新たに追加した特殊トークンを配置し、そのトークンの埋め込みと FNN のパラメータのみを最適化する。Fine-Tuning では図 1 のアーキテクチャを用いた上で、モデルのすべてのパラメータと FNN のパラメータを更新する。なお、HyperCLOVA の Fine-Tuning については GPU メモリの不足により実験することができなかった。

また、ベースラインとして、RoBERTa-large (337M) [28][5] の Fine-Tuning、BERTScore[6]、BLEURT[7] と比較する。BERTScore では Train データを用いて埋め込みを得るための最適な出力層を選択した。BLEURT では追加の学習はしておらず、日本語を含む多言語で学習された BLEURT-20 [29] を利用した。学習時のハイパーパラメータを付録 A に示す。

4.3　実験結果・議論

WMT20 en-ja と JSTS における、自動評価指標による評価と正解ラベルとの相関係数を表 1、表 2 に示す。どちらのデータセットにおいても 3 種類の Tuning 手法の全てにおいてモデルサイズが大きくなるほど精度が上がることが分かる。WMT20 en-ja によって LoRA-Tuning した GPT2-xsmall

1) https://huggingface.co/rinna/japanese-gpt2-xsmall
2) https://huggingface.co/nlp-waseda/gpt2-small-japanese
3) https://huggingface.co/rinna/japanese-gpt2-small
4) https://huggingface.co/rinna/japanese-gpt2-medium
5) https://huggingface.co/nlp-waseda/roberta-large-japanese
6) https://github.com/Tiiiger/bert_score
7) https://github.com/google-research/bleurt

表 2　自動評価指標による評価と正解ラベルとの相関係数 (JSTS)

手法	モデル	Pearson	Spearman	Kendall
Human		0.909	0.872	-
BERTScore	waseda RoBERTa-large	0.721	0.744	0.558
BLEURT-20	RemBERT	0.771	0.752	0.569
BERT Fine-Tuning	waseda RoBERTa-large	**0.923**	**0.889**	**0.729**
LLM LoRA-Tuning	rinna GPT2-xsmall	0.823	0.782	0.600
	waseda GPT2-small	0.852	0.810	0.629
	rinna GPT2-small	0.865	0.823	0.644
	rinna GPT2-medium	0.889	0.850	0.677
	HyperCLOVA 6.9B	0.909	0.872	0.702
LLM Prompt-Tuning	rinna GPT2-xsmall	0.833	0.778	0.597
	waseda GPT2-small	0.861	0.807	0.629
	rinna GPT2-small	0.862	0.810	0.631
	rinna GPT2-medium	0.874	0.823	0.646
	HyperCLOVA 6.9B	0.898	0.848	0.674
LLM Fine-Tuning	rinna GPT2-xsmall	0.868	0.817	0.640
	waseda GPT2-small	0.883	0.837	0.660
	rinna GPT2-small	0.887	0.839	0.664
	rinna GPT2-medium	0.905	0.865	0.696
	HyperCLOVA 6.9B	-	-	-

表 3　LoRA-Tuning した rinna GPT2-xsmall と HyperCLOVA 6.9B による自動評価例 (WMT20 en-ja)

機械翻訳	人間による翻訳	ラベル	xsmall	6.9B
マリアデイエスはその目的は ご利用いただけるようインフラを 利用してもらうことを目的としています。	マリア・デ・ヘスス大臣は、インフラを うまく活用し、一般市民が利用できるように することが目的だと語った。	0.52	0.74	0.51
「シャックはまだリハビリ中だ」 とクロップは言った。	「シャチはまだだ、現在リハビリ中」 と述べた。	0.54	0.74	0.67
これまでに 4 カ所の介護施設に音楽家を 招いてもてなしたり、ビンゴティーを 開いたりしてきた。	4 つの各ケアホームに音楽家を連れて行き、 ビンゴ夕食会を開催して 楽しんでもらいました。	0.71	0.58	0.71
問題は、政策論争から、移民、 自動車の効率化、住宅に関する法的課題 にまで及んでいる。	争点は政策論争から移民や自動車の燃料効率、 住宅に関する訴訟まで多岐にわたる	0.92	0.64	0.79

と HyperCLOVA による自動評価例を表 3 に示す。GPT2-xsmall に対して HyperCLOVA による自動評価の方が、正解ラベルとの相関が強いことが分かる。

WMT20 en-ja では Prompt-Tuning、Fine-Tuning、LoRA-Tuning の順に精度が上がっていく傾向が見られ、HyperCLOVA を LoRA-Tuning したモデルが最も精度が高くなった。一方で JSTS では、Prompt-Tuning、LoRA-Tuning、Fine-Tuning の順に精度が上がっていく傾向が見られたが、RoBERTa-large を Fine-Tuning したモデルが最も精度が高くなった。ただし、HyperCLOVA を LoRA-Tuning したモデルも人間 (Human) と同等の精度になっている。テキストペアの類似度に関するタスクに真の正解は無いため、Human の精度に近い値が出ていればそれ以上の精度向上に意味は無いと考える。すなわち、JSTS においてはモデルにとってタスクが容易すぎるために、手法の比較が正しく行えていない可能性がある。

3 種類の Tuning 手法の Epoch 数を揃えた上での学習時間の比は Prompt:LoRA:Fine = 0.89:1:1.24 となり、GPU メモリ使用率の比は Prompt:LoRA:Fine = 0.96:1:1.33 となった。なお、Prompt-Tuning は他の Tuning 手法と比較して学習が収束しにくい傾向があるため、収束に必要な Epoch 数は大きくなる。

5　おわりに

本論文では LLM を用いたテキスト生成の自動評価手法を提案した。LLM のモデルサイズが大きい場合には、Encoder モデルを用いた自動評価指標よりも高い精度を出すことが可能である。3 種類の Tuning 手法を比較し、Fine-Tuning よりもコストの低い LoRA-Tuning において、Fine-Tuning と同等かそれ以上の精度を達成できることを示した。今後、LLM のモデルサイズが大きくなるにつれて、それらを用いた本手法の精度も向上していくと考えられる。

謝辞

本研究は LINE 株式会社との共同研究の助成を受けて行った。

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **ACL2002**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[2] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT2019**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **ACL2020**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. **ICLR2020**, 2020.

[6] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. **NeurIPS2021**, Vol. 34, pp. 27263–27277, 2021.

[7] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In **ACL2020**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[8] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **EMNLP2020**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[9] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **EMNLP-IJCNLP2019**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **EMNLP2021**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **NeurIPS2017**, 2017.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **NeurIPS2020**, Vol. 33, pp. 1877–1901, 2020.

[13] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv, 2022. abs/2201.11990.

[14] Aakanksha Chowdhery, Sharan Narang, and Jacob Devlin et al. Palm: Scaling language modeling with pathways. arXiv, 2022. abs/2204.02311.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv, 2021. abs/2106.09685.

[16] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[17] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 4566–4575, 2015.

[18] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In **AMTA2006**, pp. 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas.

[19] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In **EMNLP2020**, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics.

[20] Guanghui Qin and Jason Eisner. Learning how to ask: Querying LMs with mixtures of soft prompts. In **NAACL-HLT2021**, pp. 5203–5212, Online, June 2021. Association for Computational Linguistics.

[21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In **ACL-IJCNLP2021**, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics.

[22] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In **EMNLP2021**, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[23] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. arXiv, 2021. abs/2103.10385.

[24] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. SPoT: Better frozen model adaptation through soft prompt transfer. In **ACL2022**, pp. 5039–5059, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[25] Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. Results of the WMT20 metrics shared task. In **Proceedings of the Fifth Conference on Machine Translation**, pp. 688–725, Online, November 2020. Association for Computational Linguistics.

[26] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **LREC2022**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[27] Boseop Kim, HyoungSeok Kim, and Sang-Woo Lee et al. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In **EMNLP2021**, pp. 3405–3424, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[28] 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.

[29] Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for MT. In **EMNLP2021**, pp. 751–762, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

A　ハイパーパラメータ

モデルの学習時に設定したハイパーパラメータを表 4 に示す。

表 4　モデルの学習時に設定したハイパーパラメータ

ハイパーパラメータ	RoBERTa Fine-Tuning	LLM LoRA-Tuning	LLM Prefix-Tuning	LLM Fine-Tuning
Learning Rate	2e-5	1e-5, 2e-4	1e-2, 1e-1	5e-5
Epoch Num	10	10	30	10
LoRA Dim	-	4	-	-
LoRA Alpha	-	32	-	-
LoRA Dropout	-	0.1	-	-