大規模言語モデルの文生成確率を用いた教師なし品質推定

樽本 空宙　梶原 智之　二宮 崇
愛媛大学大学院理工学研究科
{tarumoto@ai.cs.,kajiwara@cs.,ninomiya.takashi.mk@}ehime-u.ac.jp

概要

本研究では，言語生成タスクにおける参照なし自動評価の改善のために，大規模言語モデルの文生成確率に基づく教師なし品質推定の手法を提案する．生成文の品質推定は主に機械翻訳タスクを対象に取り組まれており，先行研究では対訳コーパスで訓練した Transformer や mBART に基づく品質推定が提案されている．先行研究の系列変換モデルよりも大規模なデータで事前訓練した大規模言語モデルは，高性能な品質推定と様々なタスクへの応用が期待できる．機械翻訳およびテキスト平易化における品質推定に関する評価実験の結果，提案手法は機械翻訳では多資源言語対において既存手法の性能を上回り，テキスト平易化では一部の教師あり品質推定の性能を上回ることを確認した．

1　はじめに

機械翻訳やテキスト平易化などの言語生成タスクでは，出力文の品質を自動評価するために多くの評価指標が提案されてきた．研究開発の現場における生成文の自動評価は，BLEU [1] や SARI [2]，COMET [3] など，参照文との表層的または意味的な類似度に基づく評価が主流である．しかし，実際にユーザが言語生成システムを使用する際には，参照文を用意できない場合が多く，これらの参照あり自動評価を用いることは困難である．このような背景から，参照文を用いずに生成文の品質を自動評価する品質推定 [4] の研究が盛んに取り組まれている．

品質推定 (QE: Quality Estimation) は主に機械翻訳を対象に研究されてきた．機械翻訳に関する国際会議 WMT の品質推定コンペティションでは，これまで多くの教師あり品質推定モデル [5, 6] が提案されてきた．しかし，機械翻訳の教師あり品質推定モデルは「原言語文，目的言語文，人手評価値」の 3 つ組を用いて訓練するため，人手評価には原言語と目的言語の両方に精通したアノテータが必要であり，アノテーションコストが非常に高い．そのため，教師あり品質推定モデルの多くは，充分な量の訓練データを利用可能なタスクや言語でしか適用できない．

この問題に対処するために，参照文や人手評価値を用いず，入出力の文対から出力文の品質を評価する教師なし品質推定が研究されている．教師なし品質推定の先行研究は，Encoder に基づく品質推定 [7, 8]，Encoder-Decoder に基づく品質推定 [9, 10]，Decoder に基づく品質推定 [11] に大別できる．

本研究では，大規模言語モデル (LLM: Large Language Model) の文生成確率を用いた Decoder に基づく品質推定の手法を提案する．具体的には，入力文を含むタスクの指示（プロンプト）を LLM に入力し，forced-decoding によって計算した出力文の生成確率に基づき，品質スコアを推定する．WMT20 QE タスク [12] における機械翻訳の品質推定に関する実験の結果，多資源言語対において教師なし品質推定の既存手法を上回る性能を達成した．また，他のタスクへの適応能力を検証するために，テキスト平易化タスクの品質推定 [13] においても実験した結果，一部の教師あり手法を上回る性能を示した．

2　関連研究

Encoder QE　入力文と出力文をそれぞれ文符号化器に入力し，得られた分散表現の間の余弦類似度に基づいて品質スコアを推定する．先行研究では，WMT20 QE タスクにおいて，LaBSE [14] や DREAM [7]，MEAT [8] が検証されており，特に少資源言語対において高い性能が報告されている．

Encoder-Decoder QE　対訳コーパスを用いて訓練した系列変換モデルを用いて，入力文から出力文への forced-decoding によって計算した出力文の生成確率に基づいて，品質スコアを推定する．先行研究では，WMT20 QE タスクにおいて，Transformer [15] に基づく手法 [9] や mBART [16] に基づく手法 [10] が検証されており，後者について多資源および中資源の言語対において高い性能が報告されている．

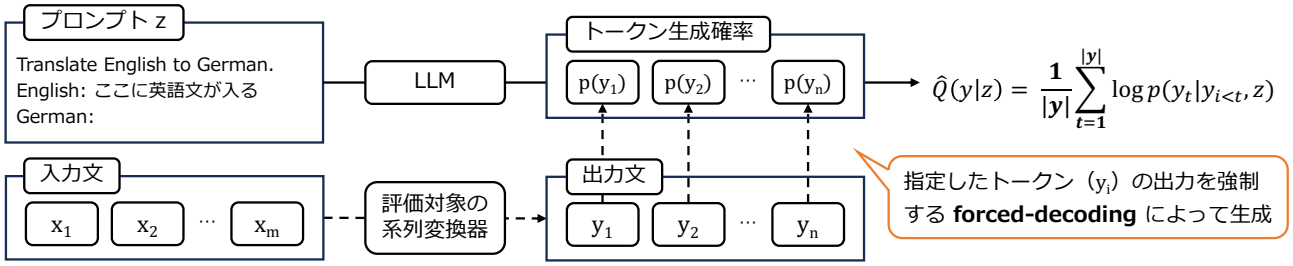

図 1　提案手法（英独の機械翻訳における例）

Decoder QE　このアプローチは，LLM に基づく品質推定である．入出力文および出力文の品質評価を指示するプロンプトを LLM に与え，LLM に品質スコアを推定させる．品質評価のためのプロンプト [11] には，出力文の品質を絶対評価するアプローチや，2 種類の出力文を比較する相対評価のアプローチがある．また，品質推定には用いられていないが，参照文に基づく自動評価の枠組みで，LLM の文生成確率に基づいて品質スコアを推定する手法 [17] も提案されている．

3　提案手法

本研究では，文生成確率に基づく教師なし品質推定や自動評価 [9, 10, 17] を発展させ，大規模言語モデルの文生成確率に基づく教師なし品質推定の手法を提案する．具体的には，図 1 に示すように，入力文を含む目的タスクのプロンプトを LLM に与え，LLM が出力文（評価対象文）を生成する確率を計算する．出力文の生成確率は，LLM が推定するトークン生成確率によらず，指定したトークンの生成を強制する forced-decoding を用いて計算し，式 (1) のように品質スコアを推定する．

$$\hat{Q}(y|z) = \frac{1}{|y|}\sum_{t=1}^{|y|} \log p(y_t|y_{i<t}, z) \qquad (1)$$

ここで，z は入力文を含む目的タスクのプロンプト，$y_{i<t}$ はタイムステップ t より前の生成トークンである．ただし，出力文が長くなるほど，この対数確率が低くなってしまう．これを防ぐために，文レベルの対数確率をトークン数 $|y|$ で割って正規化する．

4　評価実験

機械翻訳およびテキスト平易化の品質推定において，提案手法の有効性を検証する．

4.1　機械翻訳における品質推定

機械翻訳の品質推定は，WMT20 QE タスク[1] [12] において評価実験を行う．WMT20 QE タスクは，原言語および目的言語の文対に対して，品質推定モデルが推定した品質スコアとアノテータが付与した品質スコアのピアソン相関を評価するタスクである．

4.1.1　実験設定

データセット　WMT20 QE データ[1]は，英語からドイツ語 (en-de) および英語から中国語 (en-zh) の多資源言語対，ルーマニア語から英語 (ro-en) とエストニア語から英語 (et-en) の中資源言語対，ネパール語から英語 (ne-en) とシンハラ語から英語 (si-en) の少資源言語対の 6 言語対から構成される．各言語対において機械翻訳の原言語文，目的言語文，人手評価値の 3 つ組が 1,000 件ずつ提供されている．なお，目的言語文は fairseq ツールキット[2] [18] によって訓練された Transformer [15] で生成されている．

モデル　提案手法の LLM には 4 種類の LLaMA[3)4)5)6)] [19]（8B モデルおよび 70B モデル，そしてそれぞれの指示チューニングモデル）を

1) https://www.statmt.org/wmt20/quality-estimation-task.html
2) https://github.com/facebookresearch/fairseq
3) https://huggingface.co/meta-llama/Meta-Llama-3-8B
4) https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
5) https://huggingface.co/meta-llama/Meta-Llama-3-70B
6) https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

表 1 機械翻訳の品質推定（WMT20 QE タスク）におけるピアソン相関係数

		多資源		中資源		少資源		
		en-de	en-zh	ro-en	et-en	ne-en	si-en	Avg.
Encoder QE	LaBSE	0.084	0.036	0.705	0.550	0.545	0.455	0.396
	DREAM	0.196	0.197	0.724	0.578	**0.636**	0.568	0.483
	MEAT	0.215	0.222	0.717	0.587	0.634	**0.571**	0.491
Encoder-Decoder QE	mBART+M2M	0.278	0.317	**0.819**	**0.703**	0.538	0.474	**0.522**
Decoder QE	Llama-3-8B	0.355	**0.390**	0.697	0.666	0.539	0.437	0.514
	Llama-3-8B-Instruct	**0.447**	0.316	0.719	0.651	0.503	0.408	0.507
	Llama-3-70B	0.216	0.328	0.497	0.542	0.412	0.358	0.392
	Llama-3-70B-Instruct	0.357	0.331	0.517	0.530	0.368	0.355	0.410
	Llama-3-8B (5-shot)	0.240	0.311	0.499	0.527	0.440	0.346	0.394
	Llama-3-8B-Instruct (5-shot)	0.340	0.310	0.608	0.572	0.406	0.300	0.423
	Llama-3-8B (LoRA)	0.304	0.207	0.632	0.645	0.502	0.421	0.452
	Llama-3-8B-Instruct (LoRA)	0.352	0.177	0.725	0.666	0.538	0.440	0.483

使用し，図 1 の左上に示すプロンプトを LLM に与えた．また，LLM による翻訳能力を引き出すために，8B モデルについて，few-shot 文脈内学習 [20] および LoRA チューニング [21] を行った．few-shot 文脈内学習には，WMT20 QE タスクで利用可能な対訳コーパスのうち，対象の言語対における先頭 5 文対を用いる 5-shot 設定を使用した．LoRA チューニングには，同じく対訳コーパスのうち，原言語および目的言語の文長が 100 トークン未満となる文対を言語対ごとに 1 万文対ずつ使用した．LoRA チューニングのハイパーパラメータは，バッチサイズを 8 文対，学習率を 5×10^{-5} とし，ランクを $r = 32$，スケーリング係数を $\alpha = 8$，ドロップアウト率を 0.1 に設定した．また，検証用データにおける交差エントロピー損失が 3 エポック連続で改善しない場合に訓練を終了する early stopping を適用した．

比較手法　提案手法の性能を，2 節で Encoder QE および Encoder-Decoder QE として紹介した，既存の教師なし品質推定手法たちと比較した．Encoder QE には LaBSE[7] [14]，DREAM[8] [7] と MEAT[9] [8] を用い，Encoder-Decoder QE には mBART+M2M[10] [22, 10] を用いた．

7） https://huggingface.co/sentence-transformers/LaBSE
8） https://github.com/nattaptiy/qe_disentangled
9） https://github.com/EhimeNLP/MEAT
10） https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

4.1.2 実験結果

表 1 に機械翻訳の品質推定に関する実験結果を示す．分析のために，モデルの違い，モデルサイズの違い，few-shot 文脈内学習の有無，再訓練の有無の 4 観点から性能を比較した．

モデルの違いによる性能比較　提案手法である Decoder QE (Llama-3-8B，Llama-3-8B-Instruct) と先行研究である Encoder QE および Encoder-Decoder QE との性能差に注目する．提案手法は，多資源言語対である en-de および en-zh において最高性能を達成した．また，多資源言語対では Decoder QE，中資源言語対では Encoder-Decoder QE，少資源言語対では Encoder QE が高い性能を示す傾向が見られた．

モデルサイズの違いによる性能比較　モデルサイズの違いによる性能差に注目し，Llama-3-8B と Llama-3-70B，Llama-3-8B-Instruct と Llama-3-70B-Instruct の性能をそれぞれ比較する．指示チューニング [23] の有無に関わらず，Llama-3-70B-Instruct の en-zh 言語対を除くすべての言語対において，パラメータ数の少ないモデルの方が高い性能を示した．

few-shot 文脈内学習の有無による性能比較　few-shot 文脈内学習の有無が性能に与える影響を確認するために，Llama-3-8B と Llama-3-8B (5-shot)，Llama-3-8B-Instruct と Llama-3-8B-Instruct (5-shot) の性能を比較する．指示チューニングの有無に関わらず，すべての言語対において 0-shot 設定の方が高い性能を示した．

表 2　LoRA チューニングによって解消された生成エラーの例

モデル	生成文
Llama-3-8B-Instruct	**Deutsch:** Hier sind die vier obersten Ansätze, die sie verwenden.
Llama-3-8B-Instruct (LoRA)	Die folgenden vier Methoden werden häufig eingesetzt.gegen Lee Grace ab.

再訓練の有無による性能比較　再訓練の有無が性能に与える影響を確認するために，Llama-3-8B と Llama-3-8B (LoRA)，Llama-3-8B-Instruct と Llama-3-8B-Instruct (LoRA) の性能を比較する．Llama-3-8B-Instruct に LoRA チューニングすることで多資源言語対の性能は低下したものの，中資源言語対および少資源言語対では性能向上が見られた．

この要因を分析するために，forced-decoding の代わりに greedy search を用いて LLM の出力を観察した．表 2 に例示するように，再訓練していない Llama-3-8B-Instruct では生成の先頭に期待しない生成が含まれていた．LoRA チューニングによって期待している出力が得られるようになったために性能が向上したと考えられる．

4.2　テキスト平易化における品質推定

テキスト平易化の品質推定は，Simplicity-DA データセット[11] [13] を用いて実験した．本タスクは，難解文および平易文の文対に対して，品質推定モデルが推定した品質スコアとアノテータが付与した品質スコアのピアソン相関を評価するタスクである．

4.2.1　実験設定

データセット　Simplicity-DA は，英語の難解文と，テキスト平易化モデルから得られた平易文の対に対して，流暢性・同義性・平易性の 3 観点からアノテータが品質スコアを付与したものである．比較手法の訓練・検証・評価データと，提案手法の評価データは先行研究 [24] と同一のデータを使用した．

比較手法　比較手法には，機械学習に基づく手法である Kajiwara-17 [25] および Martin-18 [26]，深層学習に基づく手法である Hironaka-24 [24] を用いた．ただし，これらの手法は Simplicity-DA の一部で訓練された教師あり手法であることに注意されたい．

モデル　提案手法の LLM には，機械翻訳と同様に 4 種類の LLaMA モデルを使用した．LLM に入力するプロンプトは，図 2 を使用した．

11）https://github.com/feralvam/metaeval-simplification

```
Simplify the following complex sentence.
complex sentence: ここに難解文が入る
simplex sentence:
```

図 2　テキスト平易化における品質推定のプロンプト

表 3　テキスト平易化における品質推定の結果

	流暢性	同義性	平易性
Kajiwara-17	0.405	0.670	0.373
Martin-18	0.462	0.680	0.320
Hironaka-24	**0.750**	**0.770**	**0.622**
Llama-3-8B	0.414	0.574	0.404
Llama-3-8B-Instruct	0.509	0.609	0.476
Llama-3-70B	0.410	0.552	0.398
Llama-3-70B-Instruct	0.417	0.536	0.419

4.2.2　実験結果

実験結果を表 3 に示す．提案手法は，Hironaka-24 には及ばなかったものの，流暢性と平易性については Kajiwara-17 と Martin-18 の性能を上回った．これらの比較手法が教師あり品質推定である一方，提案手法が教師なし品質推定であることを踏まえると，品質推定の訓練データを使用できない状況において，提案手法が有用であると考えられる．

5　おわりに

本研究では，言語生成タスクにおける参照なし自動評価の性能改善のために，LLM に基づく教師なし品質推定モデルを提案した．提案手法では，タスクの遂行を指示するプロンプトを入力文とともに LLM に与え，forced-decoding によって計算した出力文の生成確率に基づいて，品質スコアを推定する．

機械翻訳の教師なし品質推定における評価実験の結果，提案手法は多資源言語対において最高性能を達成した．また，テキスト平易化の品質推定における評価実験では，最先端の教師あり品質推定モデルの性能には及ばなかったものの，その他の教師あり品質推定モデルの性能を上回った．これらの結果から，訓練データを充分に用意できないタスクや言語において，提案手法の活用が期待できる．

謝辞

本研究は，JSPS 科研費（基盤研究 B，課題番号：JP23K24907）の助成を受けたものです．

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 **Proc. of ACL**, pp. 311–318, 2002.

[2] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. **TACL**, pp. 401–415, 2016.

[3] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In **Proc. of EMNLP**, pp. 2685–2702, 2020.

[4] Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. **Quality Estimation for Machine Translation**, Vol. 11. Synthesis Lectures on Human Language Technologies, 2018.

[5] Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation. **TALLIP**, Vol. 17, No. 1, p. 22, 2017.

[6] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In **Proc. of COLING**, pp. 5070–5081, 2020.

[7] Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. In **Proc. of EMNLP**, pp. 7764–7774, 2021.

[8] Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation. In **Proc. of COLING**, pp. 5240–5245, 2022.

[9] Brian Thompson and Matt Post. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In **Proc. of EMNLP**, pp. 90–121, 2020.

[10] 西原哲郎, 岩本裕司, 吉仲真人, 梶原智之, 荒瀬由紀, 二宮崇. 多言語雑音除去自己符号化器による教師なし品質推定. 自然言語処理, Vol. 29, No. 2, pp. 669–687, 2022.

[11] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. Human-like Summarization Evaluation with ChatGPT. **arXiv:2304.02254**, 2023.

[12] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. Findings of the WMT 2020 Shared Task on Quality Estimation. In **Proc. of WMT**, pp. 743–764, 2020.

[13] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. **CL**, Vol. 47, No. 4, pp. 861–889, 2021.

[14] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In **Proc. of ACL**, pp. 878–891, 2022.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Proc. of NIPS**, pp. 5998–6008, 2017.

[16] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. **TACL**, Vol. 8, pp. 726–742, 2020.

[17] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as You Desire. In **Proc. of NAACL**, pp. 6556–6576, 2024.

[18] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In **Proc. of NAACL**, pp. 48–53, 2019.

[19] Llama Team. The Llama 3 Herd of Models. **arXiv:2407.21783**, 2024.

[20] 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 **Proc. of NIPS**, pp. 1877–1901, 2020.

[21] 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. In **Proc. of ICLR**, 2022.

[22] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. **arXiv:2407.21783**, 2020.

[23] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned Language Models Are Zero-Shot Learners. In **Proc. of ICLR**, 2022.

[24] Yuki Hironaka, Tomoyuki Kajiwara, and Takashi Ninomiya. Transfer Fine-tuning for Quality Estimation of Text Simplification. In **Proc. of LREC-COLING**, pp. 16738–16744, 2024.

[25] Tomoyuki Kajiwara and Atsushi Fujita. Semantic Features Based on Word Alignments for Estimating Quality of Text Simplification. In **Proc. of IJCNLP**, pp. 109–115, 2017.

[26] Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. Reference-less Quality Estimation of Text Simplification Systems. In **Proc. of INLG**, pp. 29–38, 2018.