質問応答を用いた翻訳評価における GPT-3 の利用

鈴木淳平 [1] 菅原朔 [2] 相澤彰子 [1,2]
[1] 東京大学 [2] 国立情報学研究所
junpppsuzuki@gmail.com {saku,aizawa}@nii.ac.jp

概要

機械翻訳の評価は、人手で行うとコストが高いため、機械を用いて自動で行えることが望ましい。自動評価の一つに、質問応答の技術を利用して、参照訳に対して生成した質問に、機械翻訳結果を文脈として用いて正しく回答できるかに着目する枠組みがある。本研究では、既存研究の精度の低い質問生成のステップを、GPT-3 を few-shot 学習することで改善できることを示す。また、回答比較のステップも、既存研究では表層的な一致度のみ考慮していたが、GPT-3 を用いて柔軟な比較を行うことで、人手評価との相関だけでなく、難しいとされている数字や日付の翻訳ミスの評価も向上することを示す。

1 はじめに

機械翻訳の評価は、プロの翻訳家による人手評価が最も正確に行えるが、時間的、金銭的なコストの観点から機械によって自動で行えることが望ましい。人手の評価でも、機械による自動評価でも、与えられた参照訳との比較で評価するのが一般的である。計算コストの低さから、最も標準的に使われる自動評価手法は、BLEU [1] に代表される、参照訳と機械翻訳結果の表層的な一致度を測る手法である。これらの手法は、表層的には異なるが意味的には等しい場合を見逃してしまう。

そこで近年は、BERT [2] 等の大規模言語モデルによる分散表現を用いることで、表層によらず意味的な一致度を捉える自動評価手法が多く提案されている (BERTScore [3]、 BLEURT [4]、 COMET [5])。これらの手法は、プロの人手評価との相関が高いことが示されている [6, 7, 8]。一方、分散表現による評価手法は、埋め込み空間において近くに位置することが原因で、固有名詞や数字等の深刻な誤訳を見逃してしまうことが報告されている [7, 9]。

他方、機械による翻訳が原文の情報を十分に表現できているなら、参照訳に対する質問を生成した場合に、翻訳結果を文脈として用いても正しく回答できるはずである。この仮説に基づいて、質問応答の技術を用いて機械翻訳を評価する MTEQA (Machine Translation Evaluation with Question Answering) [10] が提案されている。図 1 に評価過程を示す。まず参照訳内の重要な情報について質問と正解を生成し、それらに機械翻訳結果を文脈として用いて QA モデルに回答させ、最後に参照訳からの正解と翻訳結果からの回答を比較することで評価する。この手法の利点としては、評価手順で使われた質問、正解、回答を実際に見ることで、なぜその評価値になったかを容易に知ることができる解釈性を持つ。しかし既存研究において、ある程度の人手評価との相関は示されていたが、質問生成の質が低く、回答比較の部分も表層一致しか着目できていない。

そこで本研究では、質問生成の部分を正解の抽出を含めて GPT-3 [11] で行うことで、適切な質問の割合が増え、人手評価との相関が向上することを示す。さらに、回答比較の部分も GPT-3 を用いて、意味を考慮して柔軟に比較する (例: “15” に対して “fifteen” を正解にする) ことで人手評価との相関だけでなく、上で挙げた分散表現による評価手法が苦手とする誤訳への評価性能も向上することを示す。

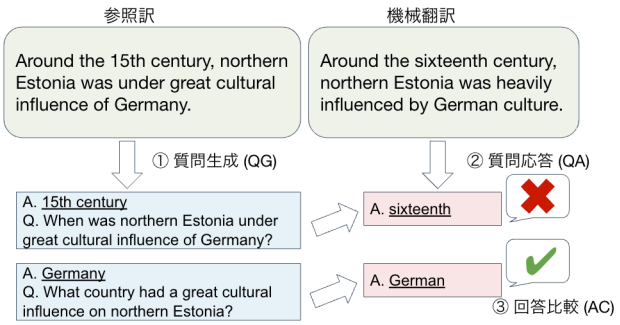

図 1 MTEQA の評価の流れ。1. 参照訳から正解と対応する質問を生成、2. 翻訳結果を文脈として用いて質問に回答、3. 参照訳での正解と翻訳結果の回答を比較。

表 1 GPT-3 による質問と正解の生成のためのプロンプト内の具体例。

Context: The league narrowed the bids to three sites: New Orleans' Mercedes-Benz Superdome, Miami's Sun Life Stadium, and the San Francisco Bay Area's Levi's Stadium.
Q&A:
Answer: Levi's Stadium
Question: What is located in the San Francisco Bay Area?
Answer: New Orleans
Question: Where is Mercedes-Benz Superdome located?
Answer: ...

表 2 GPT-3 による柔軟な回答比較のためのプロンプト内の具体例。

Context: It certainly paid off, since he played splendidly in a different role from normal, and was able to score 15 points.
Question: In what role did he score 15 points?
Correct Answer: different role
Student's Answer: different
Output: Correct. In this context, "different" is accepted.
.....
Student's Answer: role
Output: Incorrect. The answer should include "different".
.....
Student's Answer: unusual role
Output: Correct. In this context, "unusual" is accepted.

2 MTEQA-GPT-3

MTEQA は、図 1 が示すように、質問生成、質問応答、回答比較の三つのステップからなる。

2.1 質問生成 (Question Generation; QG)

最初のステップでは、与えられた比較対象の参照訳から、重要な情報について質問と正解を生成する。既存の MTEQA [10] では、まず参照訳に対して品詞の系列ラベリングを行い、名詞等を回答候補として抽出し、文脈と回答候補から質問を生成するように訓練された T5 モデル [12] に入力する。

本研究では、より質問と回答が一貫しておりかつ文脈から回答可能な質問の割合を増やすために GPT-3 を few-shot 学習した。表 1 に few-shot プロンプト内の具体的な例を示す。回答抽出と質問生成は分けずに 1 つの参照訳を与え、固有名詞や数字等について複数の回答と質問のペアを生成するように学習させた。ここで、質問生成の質が向上していることを確かめるために、WMT21(Workshop on Statistical Machine Translation)[13] の中国語から英語への翻訳タスクからランダムに 10 個参照訳を選んで、両者で質問生成を行った。質問と正解が一貫しており、かつ生成元の参照訳から回答可能なものの個数を数えた。結果として、適切な質問と回答の割合が、既存の T5 モデルが 0.603 (47 / 78 件)、提案手法の GPT-3 が 0.974 (38 / 39 件) であった。質問生成の質が改善されたことが分かる。以降、"MTEQA-gpt3-qg-" で始まる手法が提案手法である。

2.2 質問応答 (Question Answering; QA)

次に、2.1 節で生成された質問に、機械翻訳結果を文脈として抽出型の機械読解モデルに回答させる。MTEQA と提案手法共に、SQuAD-v1 [14] で訓練した T5 モデルを用いる。テストデータでの F1 スコアが 90.27 であった。

2.3 回答比較 (Answer Comparison; AC)

最後に、2.1 節で生成された正解と 2.2 節で抜き出された回答との比較を行う。既存研究においては、完全一致 (EM)、 単語レベルの F1 スコア (F1)、BLEU、 chrF [15] の 4 つが採用されていた。しかしこれらは表層的な一致度によることから、正確に比較できないことが多い。例えば、正解が "December 29, 2019" の場合、"Dec 29th, 2019" よりも "November 29, 2019" の方が高得点となってしまう。

そこで本研究では、回答比較の段階でも GPT-3 を利用することで、意味的な一致度を考慮した柔軟な採点を実現する。2.1 節と同様に、GPT-3 を回答比較用に few-shot 学習する。表 2 は実際のプロンプト内の具体例を示す。参照訳と質問に対して、"Correct Answer" として正解を、"Student's Answer" として抜き出された回答を与え、"Output" に比較結果を出力させる。例えば表の三つ目の例では、正解が "different role" に対して、この文脈においては "unusual role" も正解であることを指示している。今回は、出力の最初が "Correct" であれば 1 点、"Incorrect" であれば 0 点とした。実際に用いたプロンプトの詳細は付録 B を参照。

3 メタ評価実験

本章では、自動評価手法のメタ評価を行う。自動評価手法の良さは通常、どれだけ人手評価との相関が高いかという観点で行われる。3.1 節で MQM という人手評価について、3.2 節で比較のための既存手法について、3.3 節で MQM との相関に関する実験について、3.4 節で評価の難しい誤訳に関する評価の実験について、3.5 節でエラー分析を述べる。

3.1 人手評価

従来の WMT の翻訳タスクでは、人手評価として DA (Direct Assessment) [16] を用いてきた。しかし、DA ではプロの翻訳家の訳が機械翻訳より低い順位になったり、翻訳調と呼ばれる、翻訳特有の癖がある訳の方が好まれてしまうなどの問題が指摘され [8]、WMT21 [6] や WMT22 [7] では、相関すべきスコアとして MQM (Multidimentional Quality Metrics) が使われることになった。MQM では、プロの翻訳家が、参照訳と前後の文脈を見ながら、与えられた翻訳の誤り全てに、深刻度とカテゴリーと共にアノテートする減点方式で行う。本研究でも 3.3 節において、MQM との相関に基づいて比較を行う。

3.2 Baselines

提案手法との比較のためのベースラインとして、BLEU、chrF、BLEURT-20、COMET-22、MTEQA を用いる。BLEURT-20 と COMET-22 は、WTM21 と WMT22 でそれぞれ最も良い自動評価手法のうちの一つであった。付録 A にそれぞれの手法の詳細を記す。また、BERTScore [3] と BLERUT-20 は、翻訳の自動評価手法であるが、回答の比較部分にも使えるため、これらも利用する。

3.3 人手評価と自動評価手法の相関

自動評価手法がどれくらい MQM スコアと相関しているか評価する。評価には、WMT22 の Metrics Shared Task を利用し、言語対は中国語から英語である。テストデータとして、1875 個の中国語の原文があり、それぞれに参照訳が 2 つ用意されており (reference A、 reference B)、18 種類のシステムの翻訳結果が与えられる。システムレベルの相関を測定するので、各自動評価手法がつけたスコアの平均をシステムごとのスコアとみなし、同様に計算された MQM スコアとの相関係数を求める (Pearson、Spearman)。参照訳が二つ存在するので、一方を参照訳として用いた場合は、もう一方は追加のシステムとみなして相関を計算する。

表 3 に結果を示す。太字の値は、その列において、他の任意の手法に統計的に上回られていないことを示す。 4 列全てにおいて最も相関が高かったのは、COMET-22 と MTEQA-gpt3-qg-gpt3-ac であった。

また、MTEQA 同士を同じ回答比較方法で比べると (MTEQA-EM と MTEQA-gpt3-qg-EM など)、全ての組み合わせにおいて、相関係数の値が向上している。つまり、2.2 節で示した質問生成の質の向上が、自動評価手法全体の改善に繋がっている。

さらに、MTEQA-gpt3-qg-gpt3-ac は MTEQA-gpt3-qg-{EM、 F1、 BLEU、 chrF、bertscore、bleurt} に対し、相関係数の値が有意に向上している。GPT-3 を用いた回答比較は、表層一致や分散表現による比較よりも正確に回答の比較が可能なことが示された。

表 3 WMT22 Metrics Shared Task の中国語から英語への翻訳での MQM スコアとのシステムレベルの相関。P は Pearson、 S は Spearman を表す。太字の値は、その列の任意の値に対して有意に低くない場合を示している。

Metric	reference A		reference B	
	P	S	P	S
BLEU	0.579	0.454	0.599	0.496
chrF	0.647	0.507	0.667	0.539
BLEURT-20	0.909	**0.868**	0.878	**0.836**
COMET-22	**0.947**	**0.886**	**0.940**	**0.893**
MTEQA-EM	0.783	0.410	0.539	0.561
MTEQA-F1	0.815	0.429	0.546	0.496
MTEQA-BLEU	0.821	0.410	0.543	0.511
MTEQA-chrF	0.845	0.448	0.568	0.611
MTEQA-gpt3-qg-EM	0.802	0.564	0.836	0.596
MTEQA-gpt3-qg-F1	0.837	0.554	0.849	0.561
MTEQA-gpt3-qg-BLEU	0.833	0.568	0.849	0.586
MTEQA-gpt3-qg-chrF	0.883	0.593	0.897	0.614
MTEQA-gpt3-qg-bertscore	0.891	**0.825**	0.912	0.768
MTEQA-gpt3-qg-bleurt	0.927	0.793	**0.951**	**0.818**
MTEQA-gpt3-qg-gpt3-ac-all	**0.977**	**0.921**	**0.974**	**0.843**

3.4 数値・日付・固有名詞に着目した評価

本節では、自動評価手法にとって正しく評価することが難しい誤訳の評価性能を測定する。本研究では、WMT22 の challenge sets subtask で提案された中国語から英語でのテストデータ [17] を用いる。各データは、原文、参照訳、ある難しい現象に対して正しい訳、誤訳を含んだ訳からなる。表 4 に具体例を示す。Reference 内の “GDP” に対し、正しい訳では “gross domestic product”、誤った訳では “GPP” と訳されていて、自動評価手法は正しい訳により高い点を与えることが求められる。今回は、3 種類の現象を扱う。Number は、数字に関する誤りが 355 個、D/T は日付に関する誤りが 140 個、NE は固有名詞や専門用語に関する誤りが 110 個含まれる。メタ評価は、以下の式で表される Kendall’s tau-like correlation [6] に従う。ここで、“Concordant” は、正しい訳の方が正しくない訳よりも高い点数がついた個数、“Discordant” はそうでない場合の個数である。

表 4 Challenge set の具体例。

Source	“到 2020 年之前，我 们将努力使 单位国内生 产总值二氧化 碳排放量比 2005 年...
Reference	We will endeavour to cut carbon dioxide emissions per unit of **GDP** by...
Correct	By 2020, we will strive to make carbon dioxide emissions per unit of **gross domestic product**...
Incorrect	By 2020, we will strive to make carbon dioxide emissions per unit of **GPP**...

$$\tau = \frac{Concordant - Discordant}{Concordant + Discordant} \quad (1)$$

結果を表 5 に示す。Overall としては、BLEURT-20 と MTEQA-gpt3-qg-gpt3-ac が最も高い相関を示した。COMET-22 は、人手評価との全体的な相関では高い性能を示すが、今回の Challenge set では上記の 2 手法に及ばなかった。その他の手法では全て負の相関となり、ランダムに点数をつけた場合よりも性能が悪いという結果になった。

カテゴリ別に見ると、BLEURT-20 は D/T と NE に強く、MTEQA-gpt3-qg-gpt3-ac は Number に強い。COMET-22 は Number や NE の誤りに弱く、分散表現ベースの自動評価手法が数字や固有名詞の誤りを見逃してしまうという過去の研究の観察と一致する [7, 9]。数字同士は Embedding 空間で近い位置に埋め込まれるので、分散表現ベースの手法が見逃してしまうという説明が考えられる。

また、MTEQA-gpt3-qg-gpt3-ac はその他の MTEQA と比較して大幅に性能が向上しており、GPT-3 による柔軟な回答比較が寄与していると考えられる。本データセットでは、表 4 の具体例のように、表層的には不正解訳内の誤訳の方が参照訳に近い場合が多々ある。よって、EM、 F1、 BLEU、 chrF 等の表層ベースの手法はうまくいかないが、gpt3-ac では意味的な一致度を考慮できるので正しく判定できていると考えられる。さらに、分散表現ベースの BERTScore と BLEURT よりも高い相関を示した。

3.5 エラー分析

3.4 節の実験で、MTEQA-gpt3-qg-gpt3-ac が評価に失敗した例について原因を分析する。各カテゴリについて 30 個ずつランダムに選び、失敗の原因を Data (誤訳が含まれていない)、QG、QA、AC に分類する。結果を表 6 に示す。全体では、誤訳の含まれる箇所について質問ができていない QG のミスが最多であった。またカテゴリごとでは、Number では、

表 5 Challenge set における Kendall’s tau-like correlation。Numer は数字、D/T は日時、NE は固有名詞に関する誤り。

Metric	Overall	Number	D/T	NE
SentBLEU	-0.729	-0.735	-0.743	-0.691
chrF	-0.382	-0.301	-0.300	-0.745
BLEURT-20	**0.491**	0.476	**0.629**	**0.364**
COMET-22	0.296	0.206	0.500	0.327
MTEQA-EM	-0.841	-0.891	-0.914	-0.873
MTEQA-F1	-0.580	-0.493	-0.714	-0.691
MTEQA-BLEU	-0.580	-0.487	-0.771	-0.636
MTEQA-chrF	-0.329	-0.251	-0.243	-0.564
MTEQA-gpt3-qg-EM	-0.838	-0.808	-0.971	-0.764
MTEQA-gpt3-qg-F1	-0.712	-0.690	-0.829	-0.636
MTEQA-gpt3-qg-BLEU	-0.726	-0.707	-0.843	-0.636
MTEQA-gpt3-qg-chrF	-0.474	-0.544	-0.200	-0.600
MTEQA-gpt3-qg-bertcore	-0.312	-0.177	-0.471	-0.545
MTEQA-gpt3-qg-bleurt	-0.177	-0.054	-0.357	-0.345
MTEQA-gpt3-qg-gpt3-ac	**0.501**	**0.662**	0.471	0.018

表 6 Challenge set における MTEQA-gpt3-qg-gpt3-ac-all のエラー分析。Data はテストデータ、QG は質問生成、QA は質問回答、AC は回答比較のエラー。

Category	Overall	Number	D/T	NE
Data	7	4	1	2
QG	38	7	15	16
QA	18	8	5	5
AC	27	11	9	7
Overall	90	30	30	30

AC のミスが多く、“54 million” に対して “54,000,000” を Incorrect と判定するようなミスが見られた。D/T と NE では QG のエラーが多いので、プロンプトを工夫して質問数を増やすことで改善したい。

4 おわりに

本研究では、質問応答を利用した機械翻訳の評価手法である MTEQA の質問生成部分に GPT-3 を用いることで、より質の高い質問と正解が生成でき、MQM スコアとの相関を改善させた。また、回答比較部分でも、GPT-3 を用いて柔軟な採点を行うことで、MQM スコアとの相関で BLEURT を超え、COMET と同等の性能を記録した。さらに、評価するのが難しい数字や日付等の誤訳をより正確に評価できることを示した。一方、誤訳の部分について質問できていない故にミスを見逃すケースが多くあったので、質問数を増やして参照訳内の情報を網羅できるようにしたい。また、回答比較においても、現状は Correct か Incorrect の二択で判定するが、より柔軟に部分点を与えられるように改良をしたい。

謝辞

本研究は Google Research Grant の支援を受けたものです。

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 **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[2] 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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In **International Conference on Learning Representations**, 2020.

[4] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[5] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[6] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In **Proceedings of the Sixth Conference on Machine Translation**, pp. 733–774, Online, November 2021. Association for Computational Linguistics.

[7] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In **Proceedings of the Seventh Conference on Machine Translation**. Association for Computational Linguistics, 2022.

[8] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1460–1474, 2021.

[9] Katsuhito Sudoh, Kosuke Takahashi, and Satoshi Nakamura. Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors. In **Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)**, pp. 46–55, Online, April 2021. Association for Computational Linguistics.

[10] Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. Just ask! evaluating machine translation by asking and answering questions. In **Proceedings of the Sixth Conference on Machine Translation**, pp. 495–506, Online, November 2021. Association for Computational Linguistics.

[11] 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 Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[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. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[13] Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In **Proceedings of the Sixth Conference on Machine Translation**, pp. 1–88, Online, November 2021. Association for Computational Linguistics.

[14] 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 Empirical Methods in Natural Language Processing**, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[15] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[16] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In **Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse**, pp. 33–41, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[17] Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. Exploring robustness of machine translation metrics: A study of twenty-eight automatic metrics in the WMT22 metric task. In **Proceedings of the Seventh Conference on Machine Translation**, 2022.

[18] Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the WMT17 metrics shared task. In **Proceedings of the Second Conference on Machine Translation**, pp. 489–513, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[19] Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. In **International Conference on Learning Representations**, 2021.

[20] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, , and André F. T. Martins. COMET-22: Unbabel-ist 2022 submission for the metrics shared task. In **Proceedings of the Seventh Conference on Machine Translation**, 2022.

表 7　回答比較のプロンプトの全例 1。

Context: Herro missed 15 games in February and March due to an ankle injury, and the season began a long hiatus after his first game back against the Hornets on March 12th.
Question: What date did Herro's first game back after the injury take place?
Correct Answer: March 12th
Student's Answer: March 12
Output: Correct. "March 12" is accepted.
.....
Student's Answer: 12th
Output: Incorrect. The answer should include "March".
.....
Student's Answer: April 12th
Output: Incorrect. In this context, the answer should be not "April" but "March".

A　Baseline 手法の詳細

- BLEU [1] は、機械翻訳と参照訳の間の n-gram precision と短い翻訳に対するペナルティで計算される。3.4 節では、一文単位での評価のためのスムージング処理が加わった手法である SentBLEU [18] を用いる。
- chrF [15] は、機械翻訳と参照訳の間の文字レベルの n-gram precision である。
- BLEURT-20 [4] は、RemBERT [19] を機械翻訳と参照訳を入力として DA スコアを予測するように、過去の WMT のデータで訓練したものである。WMT21 において最高性能の自動評価手法のうちの一つである。
- COMET-22 [20] は、多言語モデルを用いて原文、機械翻訳、参照訳をそれぞれ埋め込み、それらのベクトルの組み合わせから、DA スコアを予測するように訓練したものと、MQM スコアを予測するように訓練したもののアンサンブルモデルである。WMT22 において最高性能の自動評価手法のうちの一つである。
- BERTScore[3] は、参照訳と機械翻訳のそれぞれのトークンに対する埋め込み表現を BERT を用いて計算し、参照訳内の各トークンに対して機械翻訳内で最も似ているトークンのアラインメントをとることで計算する。

B　AC のプロンプト詳細

Answer Comparison 用のプロンプトには、日付、固有名詞、数字、形容詞、副詞の比較を学習させる例がそれぞれ 3 つずつ含まれる。表 7 は日付の例、表 8 は固有名詞の例、表 9 は数字、形容詞、副詞の実際の例を示す。これらは、GPT-3 による質問生成でよく質問の対象になっていたことから選ばれた。

表 8　回答比較のプロンプトの全例 2。

Context: The data shows that, up to now, Brazil ranks second in the world in terms of the number of confirmed cases and deaths behind the United States.
Question: Which country is first in the world in terms of the number of confirmed cases and deaths?
Correct Answer: United States
Student's Answer: the United States of America
Output: Correct. The student's answers has the same meaning as the correct answer.
.....
Student's Answer: UK
Output: Incorrect. In this context, the answer should be "United States".
.....
Student's Answer: U.S.
Output: Correct. In this context, "U.S." is accepted.

表 9　回答比較のプロンプトの全例 3。

Context: It certainly paid off, since he played splendidly in a different role from normal, and was able to score 15 points.
Question: How many points did the player score?
Correct Answer: 15 points
Student's Answer: 15
Output: Correct. In this context, "15" is accepted.
.....
Student's Answer: 150 points
Output: Incorrect. In this context, "150" doesn't have the same meaning as "15".
.....
Student's Answer: fifteen
Output: Correct. In this context, "fifteen" is accepted.
.....
Question: In what role did he score 15 points?
Correct Answer: different role
Student's Answer: different
Output: Correct. In this context, "different" is accepted.
.....
Student's Answer: role
Output: Incorrect. The answer should include "different".
.....
Student's Answer: unusual role
Output: Correct. In this context, "unusual" is accepted.
.....
Question: How did he play in a different role?
Correct Answer: splendidly
Student's Answer: excellently
Output: Correct. In this context, "excellently" is accepted.
.....
Student's Answer: different
Output: Incorrect. In this context, "different" doesn't have the same meaning as "splendidly".
.....
Student's Answer: magnificently
Output: Correct. In this context, "magnificently" is accepted.