自然言語生成におけるタスク横断自動評価のメタ分析

星野　翔[1]　張　培楠[1]
[1] 株式会社サイバーエージェント
{hoshino_sho,zhang_peinan}@cyberagent.co.jp

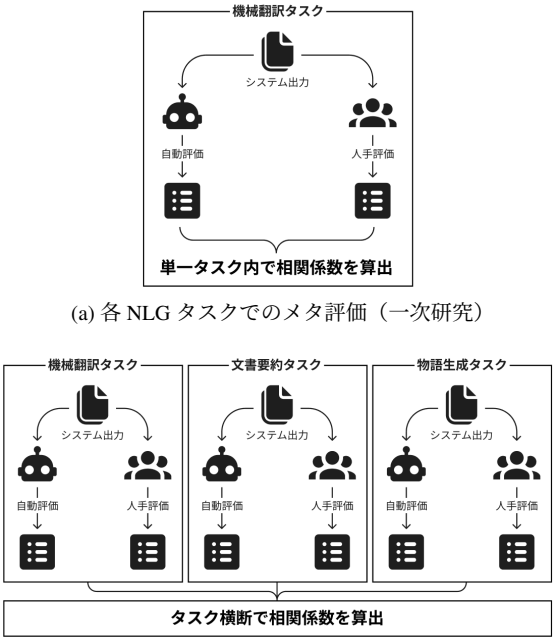
概要

本研究は、自然言語生成における自動評価尺度の信頼性をタスク横断の観点でのメタ分析で明らかにする。具体的には、機械翻訳・文書要約・物語生成のメタ評価データセットにてタスク横断で用いられる自動評価手法 4 つを調査した。従来研究のメタ評価との違いは、データセット間の比較が可能な形でシステム出力の自動・人手評価結果の相関係数を分析した点である。分析結果からは相関係数がタスク横断で著しく低下する現象が観察され、自動評価尺度のタスク横断での信頼性に留意すべきことが示唆された。従来研究の比較検討から新たな知見が得られたことでメタ分析の有用性が示された。

1　はじめに

自然言語生成 (natural language generation; NLG) の自動評価尺度[1)]はタスク横断でも用いられる。例えば、代表的な手法 BLEU [2] は元々機械翻訳 [3, 4] のために提案されたが、いまや文書要約 [5] や物語生成 [6] など他タスクに応用されている。しかし、これら手法がタスク横断の自動評価において信頼に足るか十分検討されていない。その結果、比較的新しいタスクの一つで自動評価手順が確立されていない雑談対話 [7] にて、BLEU がベンチマーク指標として流用され、未だに使われ続けている問題がある。

この問題に対し、従来研究の NLG **メタ評価** [4, 5, 6] では、機械翻訳や文書要約などの各タスク内でシステム出力の自動・人手評価結果の相関係数を算出し、自動評価手法の信頼性を議論している（図 1a）。また Marie らの研究 [8] は、機械翻訳において共通する自動評価手順の問題点を二次的な文献調査の形で報告している。しかし従来研究の議論は各タスク内に閉じているため、俯瞰的なタスク横断の観点での信頼性は議論されていない。



(a) 各 NLG タスクでのメタ評価（一次研究）

(b) NLG タスク横断でのメタ分析（二次研究）

図 1　メタ評価とメタ分析での研究方法の違い。

そこで本研究は、NLG タスク横断での自動評価尺度の信頼性を**メタ分析**により明らかにする（図 1b）。メタ分析はサーベイ論文と似た二次研究の一種であり、一次研究（原著論文）の統計的な分析を目的とする。原則的に新たなデータ・実験を用いず、従来研究の比較のみから新たな解釈や知見を導き出す形で一次研究を俯瞰的に再検討する。

本研究では機械翻訳・文書要約・物語生成のメタ評価データセットにてタスク横断で用いられる自動評価手法 4 つの信頼性を調査した。具体的には、メタ評価での知見も踏まえつつ、自動評価と人手評価のシステム順位相関係数をタスク間で比較可能な形で報告し、自動評価尺度の信頼性を議論する。

調査の結果、BLEU など自動評価手法の信頼性がタスク横断で著しく低下する現象や、タスク横断の観点でより信頼性が高い手法の存在などが観察された。総じて、自動評価尺度のタスク横断での信頼性

1)　NLG の自動評価尺度は質的評価尺度 quality metrics と多様性評価尺度 diversity metrics に大きく二分され、両者ともに広く用いられるが、多様性評価尺度についてはメタ評価方法 [1] が定まっていないため本研究の対象外とする。

に留意すべきことが示唆される。一次研究の比較検討から新たな知見が得られたことで、メタ分析の有用性が示された。

2 調査方法

本研究は NLG タスクの機械翻訳・文書要約・物語生成を対象として、データセット 4 つをメタ評価の再現実験とメタ分析の二段階で調査する（表 1）。まず予備検証として、従来研究であるメタ評価の再現実験を一般公開データセットを使用し改めて実施する。この再現実験で本研究の分析手順の正しさを確認し、また既存データセットにおける再現手順の曖昧性を明らかにする。次にメタ分析で、既存データセットをタスク横断の比較が可能な形で統計的に分析し、新たにタスク横断での議論を可能とする。

データセット　本研究には下記 3 タスクのメタ評価データセット 4 つを用いた[2)]。これらのタスクは NLG タスク三種の類型化 [9] にそれぞれ対応する。またシステム出力の自動評価結果に加えて人手評価結果も含んだデータセットが一般公開されており、メタ評価の再現性が期待できるため使用した。

- **機械翻訳**は、入力文をある言語から別の言語へ過不足なく「変換」するタスクである。WMT20・21 metrics shared task データセット [10, 4] のうち新聞ドメインの英→独翻訳のみを対象に、Freitag ら [3] に従い Google 提供の multidimensional quality metrics (MQM) に基づく人手評価結果を使用した。他の言語対や WMT21 の TED ドメイン [4] は用いなかった。
- **文書要約**は、入力文書の内容を保持しながらより短い文書へと「圧縮」するタスクである。新聞ドメインの CNN / Daily Mail [11] に基づく SummEval データセット [5] を対象に、専門家の人手評価結果を使用した。
- **物語生成**は、入力文に後続する内容のより長く一貫性がある文書を「創作」するタスクである。匿名掲示板を収集した WritingPrompts [12] に基づく HANNA データセット [6] を対象に、クラウドワーカーの人手評価結果を使用した。

統計処理　本研究では、タスク毎に自動評価結果と人手評価結果の相関係数をシステム単位で算出しタスク横断での信頼性を調査した。統計処理にはメタ評価での知見 [3] の多くを取り入れた。まず人手評価にはクラウドワーカーより信頼が置ける専門家の評価結果を可能な限り採用した。また相関係数には標本数の少なさを考慮し Pearson [13] の相関係数ではなく Kendall [14] の順位相関係数を値域 $[-1, 1]$ で用いた。さらに人間（参照文）はシステムに数えず、外れ値扱いのシステムは集計対象から除いた。

自動評価手法　自動評価手法には、参照文との単語 n-gram 一致率に基づく BLEU [2]、ROUGE [15] の F 値、文字 n-gram 一致率に基づく chrF [16]、事前学習済みモデル BERT [17] に基づく BERTScore [18] の F 値の計 4 手法を使用した。これら手法はタスク横断の質的評価尺度として用いられており、文書単位での単一参照文との比較結果を分析対象とした。

表 1　使用した NLG メタ評価データセットの内訳。システム数には人間を含めず、外れ値も除いた。

名称	タスク	システム数	自動評価手法数
WMT20 news	機械翻訳	7	29
WMT21 news	機械翻訳	13	24
SummEval	文書要約	17	21
HANNA	物語生成	10	72

3 メタ評価の再現実験

表 2 に従来研究である NLG メタ評価の再現実験結果を示す。なおこの再現実験に限り WMT21 では Pearson の相関係数、HANNA では ROUGE の F 値ではなく recall 値を用いた。また WMT20・21 では ROUGE を、WMT20 では BERTScore を自動評価に用いていないため欠損値扱いの空欄 (–) とした。

これらの値は WMT21 論文 [4] Table 23、SummEval 論文 [5] Table 2、HANNA 論文 [6] Figure 4 とそれぞれ完全に一致し、従来研究を再現できたことで分析手順の正しさを確認できた。例外的に WMT20 では今回参照した Freitag らの論文 [3] Figure 7 に測定値の記載がなく値の一致まで確認できなかった。そのため WMT20 でのメタ分析手順はより新しく値の一致まで確認できた WMT21 に準じた。

既存データセットの再現性には同様の難があり、再現手順の曖昧性に起因する論文中の値との不一致が GitHub Issues で報告されている。また報告に基づき WMT21 論文が改訂された事例も存在する[3)]。

2)　雑談対話 [7] のメタ評価データセットも存在するが、他データセットと異なる設定となるため活用できなかった。

3)　https://github.com/google-research/mt-metrics-eval/issues/1
https://github.com/Yale-LILY/SummEval/issues/17
https://github.com/dig-team/hanna-benchmark-asg/issues/1
https://github.com/google/wmt-mqm-human-evaluation/issues/9

表 2　NLG メタ評価の再現実験結果。データセット毎に異なる設定であり直接比較可能な値ではない。Pearson の相関係数または Kendall の順位相関係数での値で、桁数を丸めると各論文中の報告値と一致する。

データセット	人手評価観点	BLEU-4	ROUGE-1	ROUGE-2	chrF	BERTScore
		機械翻訳タスク				
WMT20 news [3]	Google MQM	0.7143	–	–	0.8095	–
WMT21 news [4]	Google MQM	0.9371	–	–	0.8456	0.9300
		文書要約タスク				
SummEval [5]	Coherence	0.1176	0.2500	0.1618	0.3971	0.2059
	Consistency	0.0735	0.5294	0.5882	0.5294	0.0441
	Fluency	0.3321	0.5240	0.4797	0.4649	0.2435
	Relevance	0.2206	0.4118	0.2941	0.5882	0.4265
	平均値	0.1860	0.4288	0.3810	0.4949	0.2300
		物語生成タスク				
HANNA [6]	Relevance	0.5556	0.5111	0.4667	0.6000	0.5111
	Coherence	0.3333	0.3778	0.4222	0.4667	0.5556
	Empathy	0.4222	0.4667	0.6000	0.4667	0.7333
	Surprise	0.4222	0.4667	0.4222	0.5556	0.5556
	Engagement	0.3333	0.3778	0.4222	0.4667	0.5556
	Complexity	0.5394	0.5843	0.3596	0.6742	0.4944
	平均値	0.4343	0.4641	0.4488	0.5383	0.5676

表 3　NLG タスク横断での自動評価尺度のメタ分析結果。Kendall の順位相関係数での最高値を**太字**で表す。

データセット	人手評価観点	BLEU-4	ROUGE-1	ROUGE-2	chrF	BERTScore
		機械翻訳タスク				
WMT20 news	Google MQM	0.7143	–	–	**0.8095**	–
WMT21 news	Google MQM	**0.8718**	–	–	0.7436	**0.8718**
		文書要約タスク				
SummEval	平均値	0.1860	0.4288	0.3810	**0.4949**	0.2300
		物語生成タスク				
HANNA	平均値	0.4343	0.4491	0.3226	0.5383	**0.5676**

4　タスク横断でのメタ分析

表 3 に NLG タスク横断での自動評価尺度のメタ分析結果を示す。すなわち表 2 と表 3 の違いが図 1 に示したメタ評価とメタ分析の研究方法の違いに対応する。なおメタ分析では報告値を直接比較可能とするため SummEval と HANNA で全ての人手評価観点の代表値として平均値を用いた。

この分析結果から、自動評価尺度をタスク横断で用いた場合に信頼性が著しく低下する現象が観察された。例えば、機械翻訳タスクの WMT21 では BLEU と人手評価結果の相関係数が約 0.9 と非常に強い相関を示しているが、物語生成タスクの HANNA では約 0.4 に、文書要約タスクの SummEval では約 0.2 と弱い相関にまで低下した。他手法の chrF・BERTScore にても同様の現象が観察されたが、BLEU での低下が最も顕著だった。

さらに BLEU と chrF の手法間での比較において、WMT21 を例外として chrF が BLEU を相関係数で上回っていた。例えば、SummEval と HANNA の両方で chrF は約 0.5 と比較的強い相関を示しており、それぞれ BLEU を上回る。つまりタスク横断の観点では BLEU より chrF の方が信頼性が高いと言える。他方で、事前学習済みモデルに基づく BERTScore とその他の単語 n-gram 一致率に基づく手法間での比較において、手法の性質に起因する傾向の違いは特に観察されなかった。

これらの観察から、総じて、自動評価尺度のタスク横断での信頼性に留意すべきことが示唆される。従来研究であるメタ評価の分析からこの新たな知見が得られたことで、一次研究を比較検討するメタ分析の有用性が示された。ただしメタ分析（二次研究）の調査結果はメタ評価（一次研究）の質と量に左右されやすく、手法間の優劣など確固たる結論付けにはデータ量が不十分な調査だったと考える。

5 関連研究

Deng ら [9] の研究では、NLG タスクを文書要約など「圧縮」タスク・機械翻訳など「変換」タスク・物語生成など「創作」タスクの三種に類型化しており、本研究でもこの定義を採用した (§2)。また各タスクで重視する評価観点を使い分けることで従来手法と遜色ない性能でありながらタスク横断的に使用できる自動評価手法を提案している。

Pillutla ら [19] の研究では、物語生成や雑談対話など正解が一意に定まらない「創作」タスク、いわゆる open-ended タスク向けにタスク横断の自動評価手法 MAUVE を提案している。一方で本研究が対象とした文書要約や機械翻訳など、正解を参照文として定められる「圧縮」・「変換」タスク、通称 directed generation タスク [20] での性能は調査されていない。

Tevet と Berant [1] の研究では、多様性評価尺度のメタ評価において表層多様性 form diversity と内容多様性 content diversity の 2 つの人手評価観点が内在することを指摘し、多様性を 1 つの人手評価観点とみなす従来研究の曖昧性とその結果生じやすい人手評価結果不一致の問題を明らかにした。また多様性の人手評価結果がより一致しやすい新たな評価手順を提案している。本研究ではこれらの議論を踏まえて多様性評価尺度を分析の対象外とした (§1)。

6 おわりに

本研究は、機械翻訳・文書要約・物語生成にてタスク横断で用いられる自動評価手法 4 つをメタ分析として調査した。具体的には、メタ評価データセットに含まれる自動・人手評価結果のシステム順位相関係数を比較可能な形で統計的に分析した。分析結果から、タスク横断では相関係数が著しく低下する現象が観察され、自動評価尺度のタスク横断での信頼性に留意すべきことが示唆された。一次研究の比較検討から新たな知見が得られたことで、メタ分析の有用性が示された。雑談対話など他タスクへの応用や自動評価手法の拡充は今後の課題としたい。

参考文献

[1] Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 326–346, 2021.

[2] 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, 2002.

[3] 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.

[4] 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, 2021.

[5] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 391–409, 2021.

[6] Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 5794–5836, 2022.

[7] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2122–2132, 2016.

[8] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics**, pp. 7297–7306, 2021.

[9] Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7580–7605, 2021.

[10] 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, 2020.

[11] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In **Proceedings of the 20th Conference on Computational Natural Language Learning**, pp. 280–290, 2016.

[12] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics**, pp. 889–898, 2018.

[13] Karl Pearson. Note on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, Vol. 58, No. vii, pp. 240–242, 1895.

[14] Maurice George Kendall. A new measure of rank correlation. **Biometrika**, Vol. 30, No. 1/2, pp. 81–93, 1938.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Proceedings of the ACL Workshop: Text Summarization Branches Out**, pp. 74–81, 2004.

[16] 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, 2015.

[17] 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**, pp. 4171–4186, 2019.

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

[19] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. **Advances in Neural Information Processing Systems**, Vol. 34, pp. 4816–4828, 2021.

[20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In **Proceedings of the Eighth International Conference on Learning Representations**, 2020.