機械翻訳研究の科学的信頼性：品質評価方法の分析

Benjamin Marie[†]・藤田 篤[†]・Raphael Rubino[†]

本稿では、2021 年 8 月に開催された ACL-IJCNLP 2021 において Outstanding Paper Award を受賞した我々の論文 “Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers” (Marie et al. 2021) の内容の一部を、日本語で紹介する。

1 動機と方法

技術の性能向上や応用・実用の可否などを検討する上で、当該技術の評価は不可欠である。機械翻訳 (MT) 技術の研究・開発においては、MT システムが生成する訳出物（以下、MT 訳）の品質の評価を通じた間接的な評価が行われてきた。BLEU スコア (Papineni et al. 2002) の出現以降、人間の翻訳者によって別途作成された（複数の）参照訳との比較に基づく自動評価が、安価・高速・客観的な評価方法として MT に関する研究を加速させた。近年の MT 技術に関する研究論文（以下、MT 論文）は典型的に、(1) MT の品質を向上させる手法を提案し、(2) 当該手法による MT 訳の品質を自動評価尺度を用いて評価し、(3) それが既存の手法で得られる MT 訳の品質よりも高いことをもって提案手法の優位性を示す、という論理展開で構成される。しかしながら、上記の (2) や (3) の方法が適切でない論文が査読を経て出版されていたり、論文の査読過程で査読者から不適切な提案・要求がなされたりする場合がある。これらから、論文の著者や査読者の間で、評価方法の是非についての認識が一貫していないことが示唆される。

そこで我々は、MT 技術の品質評価に関するガイドラインの構築を目指し、MT のコミュニティによる査読を経て出版された MT 論文において用いられた品質評価方法を分析した。まず、ACL 主催の ACL, NAACL, EACL, EMNLP, CoNLL, AACL において 2010～2020 年に出版された MT 論文を、ACL Anthology¹ から収集した。そして、2 種類以上の MT 訳の品質を比較している 769 件について、次の 6 種類の情報を付与した²。

- A1. 品質評価に用いているすべての自動評価尺度（亜種は一括）
- A2. 人手評価を行っているか否か
- A3. 自動評価尺度のスコアの比較時に統計的仮説検定を行っているか否か

[†] 国立研究開発法人情報通信研究機構 (NICT)

¹ <https://aclanthology.org/>

² データを無償で公開している。 https://github.com/benjamin-marie/meta_evaluation_mt/ なお A2～A6 については、“Yes” と判断された場合でも、論文中のすべての実験において成り立つとは限らないことに注意されたい。

図 1 各自動評価尺度（5 件以上の論文で用いられた尺度に限る）および人手評価を含む MT 論文の割合。“Other” はそれら以外のすべての自動評価尺度すべてを，“Human” は人手評価を指す。

- A4. 先行研究で報告されている自動評価尺度のスコアをコピーして示しているか否か
 A5. 自動評価のツールとして SacreBLEU (Post 2018) を用いているか否か
 A6. 比較している複数の MT システムが同一のデータを用いていることを確認しているか否か

2 憂慮すべき傾向

769 件の MT 論文に対して付与した上述の 6 種類の情報を分析することで、科学的に信頼できない結論を導きかねない、4 つの傾向が明らかになった。

2.1 BLEU スコアへの過度な偏向

A1 と A2 の情報を整理した結果を図 1 に示す。ほぼすべて (98.8%) の MT 論文が BLEU スコアを用いていた。BLEU スコアの限界についてはよく知られており (Callison-Burch et al. 2006; Reiter 2018; Mathur et al. 2020), 以前は TER (Snover et al. 2006) や METEOR (Banerjee and Lavie 2005) などの他の自動評価尺度も併用されていた。しかしその割合は高くはなく、BLEU スコアのみを用いた MT 論文の割合は全体の 74.3%, 2019~2020 年に限れば 82.1% に上る。また、2010 年以降少なくとも 108 種類の自動評価尺度（亜種は一括、今回の分析対象の中では 29 種類）が提案され、人間による評価との相関が BLEU スコアよりも高いことをもって「優れている」という主張がなされてきた。それにもかかわらず、提案者以外によって評価に用いられたものは RIBES (Isozaki et al. 2010) と chrF (Popović 2015) の 2 種類のみであった。

このように、多くの MT 研究者が BLEU スコア以外の自動評価尺度を無視している。科学的な合意形成はなされていないにもかかわらず BLEU スコアの差について語る際 “small”, “large”, (統計的な意味に限らず) “significant” などの表現が頻繁に用いられていることから、BLEU スコアの差は解釈しやすい、先行研究との比較が容易である、などと認知されていることが推察さ

図 2 自動評価尺度のスコアの比較時に統計的仮説検定を行っている MT 論文の割合.

れる. 論文の査読過程で, 論文の主旨とは関係ない文脈でも BLEU スコアを示すことや, BLEU スコア以外の自動評価尺度についての言及を削除することが求められる場合もある.

すでに多くの研究者によって指摘されているように, BLEU スコアだけでなく, より適切な自動評価尺度を併用すべきである.

2.2 消えゆく統計的仮説検定

統計的仮説検定は, 実験結果が偶然ではないことを確認するための標準的な作法である. MT の研究においては, 主に自動評価尺度のスコアの差が有意であることを確認する手段 (Koehn 2004; Riezler and Maxwell 2005) として統計的仮説検定が用いられてきた. しかしながら, A3 の情報 (図 2) を見ると, 自動評価尺度のスコアの比較時に統計的仮説検定を行っている MT 論文の割合は, 最も多い年でも 65%に過ぎない. さらに, 2016 年以降, その割合が急激に下がっている. つまり, 近年多くの MT 論文が自動評価尺度のスコアの有意差を確認することなく結論を出しており, 論文の査読過程でもそれを問題視していない, ということである.

もちろん, 統計的仮説検定も万能ではない (Wasserstein et al. 2019). それでも, 検定を行わないよりは信頼できるし, 無視することを正当化する理由にはならない.

2.3 安直にコピーされる自動評価尺度のスコア

A4 と A5 の情報を整理した結果を図 3 に示す. 先行研究で報告されている自動評価尺度のスコア (大抵は BLEU スコア) をコピーするという行為が 2015 年以降急増しており, その割合は 2019~2020 年には 40%近くに上る. しかしながら, これらのスコアが本当に比較可能であるか否かは必ずしも確認されていない. BLEU スコアは, 他の多くの自動評価尺度のスコアと同様に, MT 訳と参照訳の前処理の方法などによって大きく異なりうる. Post (2018) は, 多くの MT 論文が前処理に関する詳細を報告しておらず, 自動評価尺度のスコアも適切に比較できていないことを指摘し, SacreBLEU を提案した. しかしながら, SacreBLEU は 2020 年でも 40%弱の MT 論文でしか用いられていない. SacreBLEU の “signature” を報告していない論文も多い.

先行研究で報告されている自動評価尺度のスコアをコピーする場合, 報告されていない尺度

図 3 先行研究における自動評価尺度のスコアをコピーしている MT 論文 (“Copied scores”), SacreBLEU を用いている MT 論文 (“SacreBLEU”), 自動評価尺度のスコアをコピーしつつも SacreBLEU を用いずに評価を行っている MT 論文 (“Copied w/o SacreBLEU”) の割合。

図 4 同一ではないデータを用いて訓練された MT システムを比較している論文の割合。

は比較できないため BLEU スコアへの偏向が加速してしまう (2.1 節) し, MT 訳がないため統計的仮説検定も行えない (2.2 節)。また, SacreBLEU の “signature” が同一であればスコアは比較可能であると言われているが, MT システムの訓練に用いられたデータの非同一性 (2.4 節) に対しては無力である。自動評価尺度のスコアのコピーは避けるべきである。

2.4 同一でないデータに基づく公平でない比較

科学研究における作法の一つに, 評価対象以外の変数を固定して行う対照実験がある。MT システムのアーキテクチャや訓練手法を提案する際は, 評価実験に用いるデータおよびその前処理 (フィルタリング, 分かち書きのツールやパラメタ, 大文字小文字の処理など) を比較手法間で統一する必要がある。そうでなければ, 自動評価尺度のスコアの差異が手法の違いとデータの違いのどちらに依拠するのかを見分けられない。

先行研究で報告されている自動評価尺度のスコアをコピーしている MT 論文に対して A6 の情報を付与したところ, データの同一性を確認しているものは 3 件のみであった。不公平な比較を行っている論文の割合 (対象を限定しているため, この値は下限値) を図 4 に示す。2019 ~ 2020 年の MT 論文のうち 38.5% が, 同一でないデータを用いて訓練された異なるシステムを比較していた。すなわち, これらの MT 論文における結論は, 評価に瑕疵があるため科学的には裏付けられていない。

同一のデータに基づく公平な比較を行うには, 結局は自動評価尺度のスコアをコピーすることを止め, 我々自身が比較手法を再現するのが確実である。その際の手間や計算コストは, 前処理済のデータや手法の実装を公開・共有することによって軽減できる。また, 論文の査読過程で対照実験としての適切さを厳格に審査することも不可欠である。

図 5 我々のガイドラインに照らしたときの得点の分布と平均値.

3 MT 研究の評価手法のメタ評価ガイドライン

上述の 4 つの問題は，先行研究でも指摘されており，かつ容易に回避できるものである．しかしながら，多くの MT 論文が 1 つ以上の問題を含んでおり，複数の問題を含む MT 論文の割合も年々増加している．

現状よりもいくらか科学的で建設的な議論を行うための一助となることを期待し，我々は，上述の 4 つの問題をふまえた簡単なガイドライン（4 点満点の採点手法）を作成した．

- (1) MT の評価の際に，人間による評価との相関が BLEU スコアよりも高い自動評価尺度を用いているか？あるいは人手評価を行っているか？
- (2) 自動評価尺度のスコアの有意差を，統計的仮説検定を用いて確認しているか？
- (3) 自動評価尺度のスコアを他の論文などからのコピーせず，自ら算出しているか？やむを得ずコピーした場合，比較可能であることを担保しているか？
- (4) 手法の優位性を検証する場合，同一のデータを用いた適切な対照実験を行っているか？（特定のデータや前処理に特化した手法についてはこの限りではないので，是とする）

このガイドラインは，自動評価尺度のスコアに基づく評価を行う MT 論文向けのものである．MT 論文の著者自身，また学会などのプログラム委員によって，論文の信頼性評価に利用できる．本研究で分析した全論文の得点分布を図 5 に示す．MT 論文の信頼性は年々劣化しているものの，2020 年に 4 点満点の論文が 7 件あったことは興味深い．

4 今後の展望

本研究で分析した点以外にも，評価に用いられる言語対の偏り（特に英語を含む場合が極端に多い），ベースラインの適切さ (Denkowski and Neubig 2017) など，MT の評価における問題

は多々存在する。今後はこれらについての分析も予定している。また、ACL 以外の機械学習や人工知能に関する会議で発表された MT 論文における傾向も調査したい。末筆になるが、MT 以外の自然言語処理の各分野においても、同様のメタ分析が行われることを期待している。

謝 辞

本稿で報告した内容の一部は、科研費若手研究『Neural Machine Translation for User-Generated Content』（課題番号：20K19879）ならびに科研費基盤研究 (S)『翻訳規範とコンピテンスの可操作化を通じた翻訳プロセス・モデルと統合環境の構築』（課題番号：19H05660）の支援を受けた。

参考文献

- Banerjee, S. and Lavie, A. (2005). “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.
- Callison-Burch, C., Osborne, M., and Koehn, P. (2006). “Re-evaluating the Role of BLEU in Machine Translation Research.” In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 249–256.
- Denkowski, M. and Neubig, G. (2017). “Stronger Baselines for Trustable Results in Neural Machine Translation.” In *Proceedings of the 1st Workshop on Neural Machine Translation*, pp. 18–27.
- Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). “Automatic Evaluation of Translation Quality for Distant Language Pairs.” In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 944–952.
- Koehn, P. (2004). “Statistical Significance Tests for Machine Translation Evaluation.” In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 388–395.
- Marie, B., Fujita, A., and Rubino, R. (2021). “Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 756 Papers.” In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 7297–7306.
- Mathur, N., Baldwin, T., and Cohn, T. (2020). “Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics.” In *Proceedings of the 58th*

- Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4984–4997.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). “BLEU: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318.
- Popović, M. (2015). “chrF: Character n -gram F-score for Automatic MT Evaluation.” In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT)*, pp. 392–395.
- Post, M. (2018). “A Call for Clarity in Reporting BLEU Scores.” In *Proceedings of the 3rd Conference on Machine Translation (WMT): Research Papers*, pp. 186–191.
- Reiter, E. (2018). “A Structured Review of the Validity of BLEU.” *Computational Linguistics*, **44** (3), pp. 393–401.
- Riezler, S. and Maxwell, J. T. (2005). “On Some Pitfalls in Automatic Evaluation and Significance Testing for MT.” In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 57–64.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). “A Study of Translation Edit Rate with Targeted Human Annotation.” In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 223–231.
- Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). “Moving to a World Beyond “ $p < 0.05$ ”.” *The American Statistician*, **73** (sup1), pp. 1–19.

略歴

Benjamin Marie: 2016 年フランス・Paris-Sud University で Ph.D. を取得。同年より NICT 先進的音声翻訳研究開発推進センター研究員。主に機械翻訳を含む多言語処理の研究に従事。

藤田 篤: 2005 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。博士 (工学)。2014 年より NICT 先進的音声翻訳研究開発推進センター主任研究員。主に言い換え表現の生成と認識、機械翻訳の研究に従事。

Raphael Rubino: 2011 年フランス・Avignon University で Ph.D. を取得。2019 年より NICT 先進的音声翻訳研究開発推進センター研究員、2021 年より同主任研究員。主に機械翻訳、テキスト品質評価、機械学習の研究に従事。