Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers

Benjamin Marie **Atsushi Fujita** **Raphael Rubino**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{bmarie,atsushi.fujita,raphael.rubino}@nict.go.jp

Abstract

This paper presents the first large-scale meta-evaluation of machine translation (MT). We annotated MT evaluations conducted in 769 research papers published from 2010 to 2020. Our study shows that practices for automatic MT evaluation have dramatically changed during the past decade and follow concerning trends. An increasing number of MT evaluations exclusively rely on differences between BLEU scores to draw conclusions, without performing any kind of statistical significance testing nor human evaluation, while at least 108 metrics claiming to be better than BLEU have been proposed. MT evaluations in recent papers tend to copy and compare automatic metric scores from previous work to claim the superiority of a method or an algorithm without confirming neither exactly the same training, validating, and testing data have been used nor the metric scores are comparable. Furthermore, tools for reporting standardized metric scores are still far from being widely adopted by the MT community. After showing how the accumulation of these pitfalls leads to dubious evaluation, we propose a guideline to encourage better automatic MT evaluation along with a simple meta-evaluation scoring method to assess its credibility.

1 Introduction

New research publications in machine translation (MT) regularly introduce new methods and algorithms to improve the translation quality of MT systems. In the literature, translation quality is usually evaluated with automatic metrics such as BLEU (Papineni et al., 2002) and, more rarely, by humans. To assess whether an MT system performs better than another MT system, their scores given by an automatic metric are directly compared. While such comparisons between MT systems are exhibited in the large majority of MT papers, there are no well-defined guideline nor clear prerequisites under which a comparison between MT systems is considered valid. Consequently, we assume that evaluation in MT is conducted with different degrees of thoroughness across papers and that evaluation practices have evolved over the years. What could be considered, by the research community, as a good evaluation methodology ten years ago may not be considered good today, and vice versa. This evolution has not been studied and whether MT evaluation has become better, or worse, is debatable.

On the other hand, several requirements for MT evaluation have been well-identified. For instance, the limitations of BLEU are well-known (Callison-Burch et al., 2006; Reiter, 2018; Mathur et al., 2020) and the necessity to report automatic metric scores through standardized tools, such as SacreBLEU, has been recognized (Post, 2018). Moreover, a trustworthy evaluation may adopt statistical significance testing (Koehn, 2004) and strong baselines (Denkowski and Neubig, 2017). However, to what extent these requirements have been met in MT publications is unclear.

In this paper, we propose the first large-scale meta-evaluation of MT in which we manually annotated 769 research papers published from 2010 to 2020. Our study shows that evaluation in MT has dramatically changed since 2010. An increasing number of publications exclusively rely on BLEU scores to draw their conclusions. The large majority of publications do not perform statistical significance testing, especially since 2016. Moreover, an increasing number of papers copy and compare BLEU scores published by previous work while tools to report standardized metric scores are still far from being extensively adopted by the MT community. We also show that compared systems are often trained, validated, or even evaluated, on data that are not exactly the same. After demonstrating

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 7297–7306
August 1–6, 2021.

how the accumulation of these pitfalls leads to dubious evaluation, we propose a general guideline for automatic evaluation in MT and a simple scoring method to meta-evaluate an MT paper. We believe that the adoption of these tools by authors or reviewers have the potential to reverse the concerning trends observed in this meta-evaluation.

2 A Survey on MT Evaluation

We manually annotated the MT evaluation in research papers published from 2010 to 2020 at *ACL conferences.[1] To identify MT papers, we searched the ACL Anthology website[2] for the terms "MT" or "translation" in their titles[3] and analyzed among them the 769 papers that make comparisons of translation quality between at least two MT systems. For each year between 2010 and 2020, we respectively annotated the following numbers of papers: 53, 40, 59, 80, 67, 45, 51, 62, 94, 115, and 103.

We annotated each paper as follows:

A1. All the automatic metrics used to evaluate the translation quality of MT systems. We did not list variants of the same metric: e.g., chrF3 and chrF++ are labeled chrF (Popović, 2015). Moreover, we did not consider metrics which only target specific aspects of the translation quality, such as pronoun translation and rare word translation.

A2. Whether a human evaluation of the translation quality has been conducted: yes or no. If the human evaluation only targets specific types of errors and did not evaluate the translation quality of the entire text, we answered "no."[4]

A3. Whether any kind of statistical significance testing of the difference between automatic metric scores has been performed: yes or no. Potentially, some papers did perform significance testing without mentioning it, but due to the lack of evidences such papers have been annotated with "no" for this question.

A4. Whether it makes comparisons with automatic metric scores directly copied from previous work to support its conclusion: yes or no. Most papers copying scores (mostly BLEU) clearly mention it. If there is no evidence that the scores have been copied, we annotated these papers with "no" for this question.

A5. Whether SacreBLEU has been used: yes or no. If there is no mention or reference to "SacreBLEU," we assume that it has not been used. Note that "yes" does not mean that the paper used SacreBLEU for all the MT systems evaluated.

A6. If previous work has not been reproduced but copied, whether it has been confirmed that all the compared MT systems used exactly the same pre-processed training, validating, and testing data: yes or no.

Except for A6, the annotation was straightforward since most papers present a dedicated section for experimental settings with most of the information we searched for. Answering A6 required to check the data exploited in the previous work used for comparison. Note that answering "yes" to the questions from A2 to A6 may only be true for at least one of the comparisons between MT systems, while we did not evaluate how well it applies. For instance, answering "yes" to A5 only means that at least one of the systems has been evaluated with SacreBLEU but not that the SacreBLEU signature has been reported nor that SacreBLEU scores have been correctly compared with other BLEU scores also computed with SacreBLEU.

Our annotations are available as a supplemental material of this paper. To keep track of the evolution of MT evaluation, we will periodically update the annotations and will make it available online.[5]

3 Pitfalls and Concerning Trends

This section discusses the four main pitfalls identified in our meta-evaluation of MT: the exclusive use of BLEU, the absence of statistical significance testing, the comparison of incomparable results from previous work, and the reliance on comparison between MT systems that do not exploit exactly the same data. We report on how often they affected MT papers and recent trends. Based on previous

[1] We considered only *ACL main conferences, namely ACL, NAACL, EACL, EMNLP, CoNLL, and AACL, as they are the primary venues for publishing MT papers.

[2] www.aclweb.org/anthology/

[3] There are potentially MT papers falling outside these search criteria but we considered the 769 papers we obtained to be representative enough for the purpose of this study.

[4] Note that we only check here whether the automatic evaluation is supported by a human evaluation. Previous work already studied pitfalls in human evaluation (Läubli et al., 2020).

[5] The up-to-date version can be found here: github.com/benjamin-marie/meta_evaluation_mt.

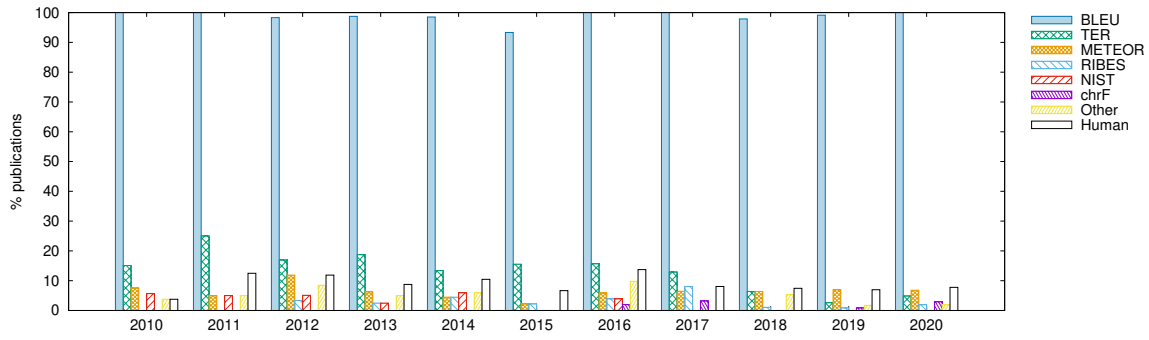

Figure 1: Percentage of papers using each evaluation metric per year. Metrics displayed are used in more than five papers. "Other" denotes all other automatic metrics. "Human" denotes that a human evaluation has been conducted.

work and supporting experiments, we show how each of these problems and their accumulation lead to scientifically dubious MT evaluation.

3.1 The 99% BLEU

Automatic metrics for evaluating translation quality have numerous advantages over a human evaluation. They are very fast and virtually free to run provided that a reference translation is already available. Their scores are also reproducible. As such, automatic metrics remained at the center of MT evaluation for the past two decades. New metrics that better correlate with human judgments are regularly introduced. We propose in this section to analyze the use of automatic metrics in MT research, relying on our annotations for A1 and A2.

This is probably the most expected finding in our study: the overwhelming majority of MT publications uses BLEU. Precisely, 98.8% of the annotated papers report on BLEU scores. As shown in Figure 1, the ratio of papers using BLEU remained stable over the years. On the other hand, BLEU scores used to be more often supported by scores from other metrics, such as TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005), than they are now. The large majority of papers, 74.3%, only used BLEU scores to evaluate MT systems, i.e., without the support of any other metrics nor human evaluation. It increases to 82.1% if we consider only the years 2019 and 2020.

This tendency looks surprising considering that no less than 108 new metrics[6] have been proposed in the last decade. They have been shown to better correlate with human judgments than BLEU. Some are even easier to use and more reproducible by being tokenization agnostic, such as chrF. We counted 29 metrics proposed at *ACL conferences since 2010 while the remaining metrics were proposed at the WMT Metrics Shared Tasks. 89% of these 108 new metrics have never been used in an *ACL publication on MT (except in the papers proposing the metrics). Among these metrics, only RIBES (Isozaki et al., 2010) and chrF have been used in more than two MT research paper.

When properly used, BLEU is a valid metric for evaluating translation quality of MT systems (Callison-Burch et al., 2006; Reiter, 2018). Nonetheless, we argue that better metrics proposed by the research community should be used to improve MT evaluation. To illustrate how wrong an evaluation can become by only relying on one metric, we computed with BLEU and chrF scores[7] of WMT20 submissions to the news translation shared task[8] (Barrault et al., 2020) using SacreBLEU and show rankings given by both metrics in Table 1. Results show that BLEU and chrF produce two different rankings. For instance, for the Ja→En task, NiuTrans system is the best according to BLEU by being 1.1 points better than the Tohoku-AIP-NTT system ranked second. In most MT papers, such a difference in BLEU points would be considered as a *significant* evidence of the superiority of an MT system and as an improvement in translation quality. Relying only on these BLEU scores without any statistical significance testing nor human evaluation would thus lead to the conclusion that NiuTrans system is the best. However, according to another metric that better correlates with human

[6] We did not count variants of the same metric and excluded metrics only proposed for an evaluation at segment level.

[7] SacreBLEU (short) signatures: chrF2+l.{ja-en,zh-en}+n.6+s.false+t.wmt20+v.1.5.0 and BLEU+c.mixed+l.{ja-en,zh-en}+#.1+s.exp+t.wmt20+tok.13a+v.1.5.0

[8] data.statmt.org/wmt20/translation-task/

Rank	Japanse-to-English (Ja→En)				Chinese-to-English (Zh→En)			
	BLEU	System	chrF	System	BLEU	System	chrF	System
1	26.6♠	NiuTrans	0.536	Tohoku-AIP-NTT	36.9	WeChat_AI	0.653	Volctrans
2	25.5	Tohoku-AIP-NTT	0.535	NiuTrans	36.8	Tencent_Translation	0.648♦	Tencent_Translation
3	24.8♦	OPPO	0.523♦	OPPO	36.6	DiDi_NLP	0.645♦	DiDi_NLP
4	22.8♦	NICT_Kyoto	0.507♦	Online-A	36.6	Volctrans	0.644♦	DeepMind
5	22.2♦	eTranslation	0.504♦	Online-B	35.9♦	THUNLP	0.643♦	THUNLP

Table 1: Rankings of WMT20 top 5 submissions for the News Translation Shared Tasks according to BLEU and chrF scores. Superscripts indicate systems that are significantly worse (♦) and better (♠) according to each metric (p-value < 0.05) than Tohoku-AIP-NTT and Volctrans systems for Ja→En and Zh→En, respectively.

judgment, i.e., chrF, this does not hold: Tohoku-AIP-NTT system is better. Similar observations are made for the Zh→En task.[9] These observations have often been made by the MT community, for instance at WMT shared tasks, but nonetheless rarely seen in research papers.

We assume that MT researchers largely ignore new metrics in their research papers for the sake of some comparability with previous work or simply because differences between BLEU scores may seem more meaningful or easier to interpret than differences between scores of a rarely used metric. Most papers even qualify differences between BLEU scores as "small," "large," or "significant" (not necessarily statistically), implying that there is a scientific consensus on the meaning of differences between BLEU scores. As we show in the following sections, all these considerations are illusory. Moreover, BLEU may also be directly requested by reviewers, or even worse, other metrics may be requested to be dropped.[10] We believe that the exclusive reliance on BLEU can be ended and the use of better metrics should be encouraged, in addition to or in lieu of BLEU, by the adoption of a guideline for automatic MT evaluation (see Section 4).

3.2 The Disappearing Statistical Significance Testing

Statistical significance testing is a standard methodology designed to ensure that experimental results are not coincidental. In MT, statistical significance testing has been used on automatic metric scores and more particularly to assess whether a particular difference of metric scores between two MT systems is not coincidental. Two methods are prevalent in MT: the paired bootstrap test (Koehn, 2004) and the approximate randomization test (Riezler and Maxwell, 2005), for instance respectively implemented in Moses[11] and MultEval.[12]

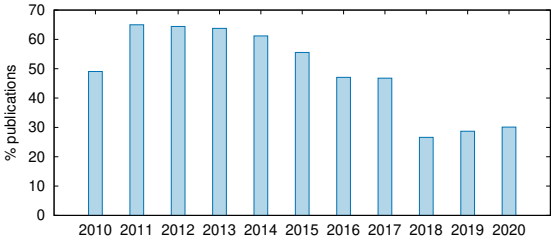

Figure 2: Percentage of papers testing statistical significance of differences between metric scores.

Dror et al. (2018) report that while the natural language processing (NLP) community assigns a great value to experimental results, statistical significance testing is rarely used. We verified if this applies to MT evaluations based on our annotations for A3. Figure 2 shows the percentage of papers that performed statistical significance testing. We found out that the observations by Dror et al. (2018) apply to MT since never more than 65.0% of the publications in a year (2011) performed statistical significance testing. Furthermore, our meta-evaluation shows a sharp decrease of its use since 2016. Most papers did not check whether their results are not coincidental but drew conclusions from them.

MT papers mainly relied on the amplitude of the differences between metric scores to state whether they are significant or not. This was also observed by Dror et al. (2018) for NLP in general.

For illustration, we also performed statistical significance testing[13] with BLEU and chrF scores

[9] For both Ja→En and Zh→En tasks, systems ranked first by chrF were also ranked first by the human evaluation.

[10] Examples of such requests or related comments by reviewers can be found in the ACL 2017 review corpus (github.com/allenai/PeerRead), e.g., in the review ID 369 we read: "I am also rather suspicious of the fact that the authors present only METEOR results and no BLEU."

[11] github.com/moses-smt/mosesdecoder

[12] github.com/jhclark/multeval

[13] For all the statistical significance testing performed in this paper, we used the paired bootstrap test with 1,000 samples and 1,000 iterations.

Ja→En			Zh→En		
System	BLEU	chrF	System	BLEU	chrF
Tohoku-AIP-NTT	25.5	0.536	Volctrans	36.6	0.653
Custom 1	25.5	0.536	Custom 1	36.6	0.653
Custom 2	18.7	0.503	Custom 2	32.2	0.638

Table 2: BLEU and chrF scores of the customized Tohoku-AIP-NTT and Volctrans outputs from which only one sentence has been modified. The first row shows the results of the original WMT20 submissions. Custom 1 replaced the last sentence with an empty line, while Custom 2 replaced the last sentence with a sequence repeating 10k times the same token. None of these systems are significantly different according to statistical significance testing on these scores.

on the WMT20 submissions in Table 1. For Ja→En, NiuTrans system is significantly better in BLEU than Tohoku-AIP-NTT system. In contrast, they are not significantly different in chrF. Using only BLEU, we would conclude that NiuTrans system is significantly the best. This is not confirmed by chrF hence we need to report on more than one metric score to conduct a credible evaluation, even when performing statistical significance testing.

Furthermore, to show that the significance of a difference between metric scores is independent from its amplitude, we performed additional experiments by modifying only one sentence, replacing it with an empty line or by the repetition of the same token many times,[14] from Tohoku-AIP-NTT and Volctrans systems' outputs. Results in BLEU and chrF are reported in Table 2. We observe that a difference in only one sentence can lead to a difference in BLEU of 6.8 points (Ja→En, Custom 2).[15] Nonetheless, our statistical significance tests did not find any system significantly better than the others.

While the importance of statistical significance testing is regularly debated by the scientific community (Wasserstein et al., 2019), it remains one of the most cost-effective tools to check how trustworthy a particular difference between two metric scores is.[16]

[14]This could be considered as a simulation of potential defects from an MT framework or model, e.g., when translating extremely long sequences.

[15]For "Custom 2," BLEU greatly penalized the increase of the number of tokens in the output. This is indicated by the length ratio reported by SacreBLEU but rarely shown in MT papers.

[16]Wasserstein et al. (2019) give several recommendations for a better use of statistical significance testing.

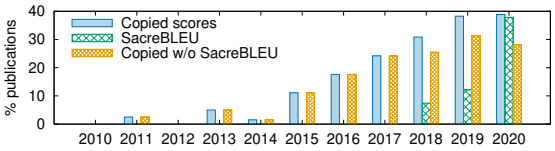

Figure 3: Percentage of papers copying scores from previous work ("Copied scores"), using SacreBLEU ("SacreBLEU"), and copying scores without using SacreBLEU ("Copied w/o SacreBLEU").

3.3 The Copied Results

An MT paper may compare the automatic metric scores of proposed MT systems with the scores reported in previous work. This practice has the advantage to save the time and cost of reproducing competing methods. Based on our annotations for A4, we counted how often papers copied the scores from previous work to compare them with their own scores. As pointed out by Figure 3, copying scores (mostly BLEU) from previous work was rarely done before 2015. In 2019 and 2020, nearly 40% of the papers reported on comparisons with scores from other papers. While many papers copied and compared metric scores across papers, it is often unclear whether they are actually comparable. As demonstrated by Post (2018), BLEU, as for most metrics, is not a single metric. It requires several parameters and is dependent on the pre-processing of the MT output and reference translation used for scoring. In fact, Post (2018) pointed out that most papers do not provide enough information to enable the comparability of their scores with other work. Post (2018) proposed a tool, SacreBLEU, to standardize metrics[17] in order to guarantee this comparability, provided that all the scores compared are computed with SacreBLEU.[18] This is the only tool of this kind used by the papers we annotated. However, based on our annotations for A5, Figure 3 shows that SacreBLEU is still far from widely adopted by the MT community, even though it is gradually getting more popular since its emergence in 2018. Moreover, papers that copy BLEU scores do not always use SacreBLEU, even in 2020.

To illustrate how deceiving a comparison of copied scores can be, we report on BLEU and chrF scores using different processing,[19] commonly

[17]Currently BLEU, chrF, and TER.

[18]SacreBLEU also generates a "signature" to further ensure this comparability: two scores computed through SacreBLEU with an identical signature are comparable.

[19]For all our processing, we used Moses (code version mmt-

Processing	Tohoku-AIP-NTT (Ja→En) BLEU	chrF	Volctrans (Zh→En) BLEU	chrF
original	25.5	0.536	36.6	0.653
fully lowercased	26.9	0.549	38.2	0.664
norm. punct.	25.5	0.537	37.8	0.657
tokenized	26.7	0.541	37.1	0.653
+ norm. punct.	26.8	0.541	38.5	0.659
+ aggressive	27.8	0.541	39.5	0.659

Table 3: BLEU and chrF scores computed by SacreBLEU after applying different processing on some WMT20 MT system outputs (from Tohoku-AIP-NTT and Volctrans) and on the reference translations. None of these rows are comparable.

adopted by MT researchers, applied to some MT system outputs and reference translations of the WMT20 news translation shared tasks. Our results are presented in Table 3. The first row presents original SacreBLEU scores, i.e., detokenized. Second and third rows respectively show the impact of lowercasing and punctuation normalization on metric scores. Scores are increased. Last three rows show the results on tokenized MT outputs. Applying both punctuation normalization and aggressive tokenization with Moses scripts leads to BLEU scores several points higher than the original SacreBLEU scores. Obviously, none of the scores in different rows are comparable. Nonetheless, MT papers still often report on *tokenized* BLEU scores compared with tokenized, or even detokenized, BLEU scores from other papers without exactly knowing how tokenization has been performed. Tokenized BLEU scores reported in MT papers are often computed using the multi-bleu script of Moses even though it displays the following warning:[20] "*The scores depend on your tokenizer, which is unlikely to be reproducible from your paper or consistent across research groups.*"

Even though the work of Post (2018) is a well-acclaimed initiative towards better MT evaluation, we believe that it can only be a patch for questionable evaluation practices. A comparison with a copied score is *de facto* associated with the absence of statistical significance testing since the MT output used to compute the copied score is not available. We also observed several misuses of SacreBLEU, such as the comparison of scores obtained by SacreBLEU against scores obtained by other tools. SacreBLEU signatures are also often not reported despite being required to ensure the comparability between SacreBLEU scores.

Ultimately, comparisons with copied scores must be avoided. As we will show in the next section, copying scores also calls for more pitfalls.

3.4 The Data Approximation

In MT, datasets are mostly monolingual or parallel texts used in three different steps of an experiment: training a translation model, tuning/validating the model, and evaluating it. Henceforth, we denote these datasets as training, validating, and testing data, respective to these three steps. How these datasets are pre-processed strongly influences translation quality. MT papers regularly propose new methods or algorithms that aim at better exploiting training and/or validating data. Following the scientific method, we can then define these new methods/algorithms and datasets as independent variables of an MT experiment while the translation quality, approximated by metric scores, would be our dependent variable that we want to measure. Testing the impact of a new algorithm on our dependent variable requires to keep all other independent variables, such as datasets, unchanged. In other words, changing datasets (even slightly) and methods/algorithms in the same experiment cannot answer whether the change in metric scores is due to the datasets, methods/algorithms, or the combination of both.

Relying on our annotation for A6, we examined how often MT papers compared MT systems for which the datasets and/or their pre-processing[21] described in the papers are not exactly identical. Note that we only performed this comparison for papers that copied and compared metric scores from previous work. Here, we also excluded comparisons between systems performed to specifically evaluate the impact of new datasets, pre-processing methods, and human intervention or feedback (e.g., post-editing and interactive MT). If we had any doubt whether a paper belongs or not to this category, we excluded it. Consequently, our estimation can be considered as the lower bound.

To illustrate the impact of modifications of these datasets on metric scores, we conducted experiments using the training, validating, and testing data of the WMT20 news translation tasks. We

mvp-v0.12.1-2851-gc054501) scripts.

[20] This warning has been added on 20 Oct. 2017. Insightful discussions on this commit can be found there: github.com/moses-smt/mosesdecoder/commit/545eee7e75487aeaf45a8b077c57e189e50b2c2e.

[21] For pre-processing, we checked, for instance, tokenization (framework and parameters), casing, subword segmentations (method and vocabulary size), data filtering, etc.

```
--type transformer --mini-batch-fit --valid-freq
 5000 --save-freq 5000 --workspace 10000
--disp-freq 500 --beam-size 12 --normalize
 1 --valid-mini-batch 16 --overwrite
--early-stopping 5 --cost-type ce-mean-words
--valid-metrics bleu --keep-best --enc-depth
 6 --dec-depth 6 --transformer-dropout
 0.1 --learn-rate 0.0003 --lr-warmup 16000
--lr-decay-inv-sqrt 16000 --lr-report
--label-smoothing 0.1 --devices 0 1 2 3 4 5 6
 7 --optimizer-params 0.9 0.98 1e-09 --clip-norm
 5 --sync-sgd --exponential-smoothing --seed 1234
```

Table 4: Hyper-parameters of Marian used for training our NMT systems.

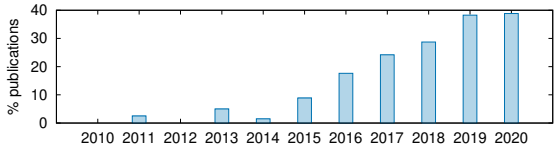

Figure 4: Percentage of papers that compared MT systems using data that are not identical.

trained neural MT (NMT) systems with Marian[22] (Junczys-Dowmunt et al., 2018), using the hyper-parameters in Table 4, on all the provided parallel data ("all" configurations) and removed sentence pairs based on their length ("Max Len."). This simple filtering step is usually applied for a more efficient training or due to some limits of the framework, method, or algorithm used. Yet, it is so common as a pre-processing step that it is rarely described in papers. As shown in Table 5, we observed that BLEU scores vary by several points depending on the maximum length used for filtering. Another common pre-processing step is the truecasing of the datasets. While it is rather commonly performed by participants in the WMT translation shared tasks, how casing is handled is rarely mentioned in research papers. In our experiments, applying this step changed BLEU scores by more than 0.5 points. Further experiments applying language identification filtering or removing one corpus from the training data also lead to variations in metric scores. The best configurations according to metric scores do not use truecasing and has a maximum sentence length set at 120 (second row). A comparison of this configuration with the third row, which uses truecasing and a different maximum sentence length, cannot lead to the conclusion that truecasing decreases translation quality, since we changed two variables at the same time.

While these observations may be expected or even obvious, Figure 4 shows that we found in

[22]Version: v1.7.6 1d4ba73 2019-05-11 17:16:31 +0100

Data	Max Len.	tc	En→De BLEU	En→De chrF	Ja→En BLEU	Ja→En chrF
all	120	✓	30.9	0.599	20.4	0.478
all	120		31.5♠	0.604♠	21.1♠	0.481
all	100	✓	30.8	0.597	20.5	0.476
all	80	✓	29.8♦	0.584♦	20.0	0.471♦
all	60	✓	26.6♦	0.549♦	18.5♦	0.453♦
lid filtered	120	✓	30.3♦	0.596	20.7	0.480
- 1 corpus	120	✓	30.7	0.600	19.6♦	0.468♦

Table 5: BLEU and chrF scores of systems using differently pre-processed parallel data for training. "Max Len." denotes that sentence pairs with sentence longer than the specified number, in terms of subword tokens, are removed. "tc" denotes whether truecasing is done or not. If not, original case of the data is kept for all datasets. "lid filtered" denotes that the training data are filtered with language identification tools. Last row denotes that we remove one corpus from the training data: "Rapid" for En→De and "OpenSubtitles" for Ja→En. ♦ and ♠ respectively denote systems that are significantly worse and better (p-value < 0.05), according to the metric, than the system in the first row.

our meta-evaluation an increasing amount of MT papers (38.5% for the 2019–2020 period) drawing conclusions of the superiority of a particular method or algorithm while also using different data. While their conclusions may be valid, the evaluation conducted in these papers is scientifically flawed and cannot support the conclusions. We assume that this is mainly due to a rather common lack of detailed experimental settings. Consequently, it makes a specific experiment often impossible to be reproduced identically. In most cases, ensuring the comparability with the published scores of an MT system is only possible by replicating the MT system by ourselves. There have been initiatives towards the release of pre-processed datasets for MT, for instance by the WMT conference that released pre-processed data for WMT19.[23] Nonetheless, we only noticed a very small number of papers exploiting pre-processed training/validating/testing data publicly released by previous work.[24] We believe that the current trend should be reversed. Reviewers should also request more rigor to the authors by checking the configurations of the compared MT systems to make sure that their comparison can, indeed, answer whether the proposed method/algorithm improves MT inde-

[23]This effort has not been conducted for WMT20.

[24]For instance, Ma et al. (2020) and Kang et al. (2020) used exactly the same pre-processed data for research on document-level NMT released by Maruf et al. (2019).

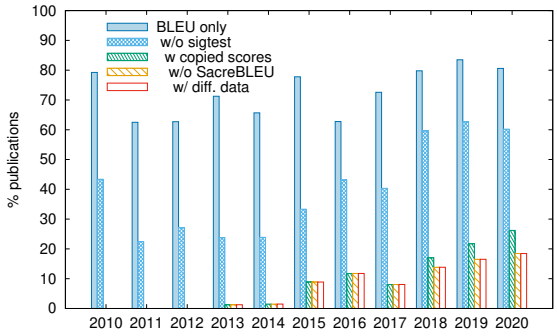

Figure 5: Percentage of papers affected by the accumulation of pitfalls. Each bar considers only the papers counted by the previous bar, e.g., the last bar considers only papers that compared MT systems exploiting different datasets while exclusively using BLEU ("BLEU only"), without performing statistical significance testing ("w/o sigtest"), to measure differences with BLEU scores copied from other papers ("w/ copied scores") while not using SacreBLEU ("w/o SacreBLEU").

pendently of the data and their pre-processing.

4 A Guideline for MT Meta-Evaluation

4.1 Motivation

The MT community is well-aware of all the pitfalls described in Section 3. They have all been described by previous work. Nonetheless, our meta-evaluation shows that most MT publications are affected by at least one of these pitfalls. More puzzling are the trends we observed. Figure 5 shows that an increasing number of publications accumulate questionable evaluation practices. In the period 2019–2020, 17.4% (38 papers) of the annotated papers exclusively relied for their evaluation on differences between BLEU scores of MT systems, of which at least some have been copied from different papers, without using SacreBLEU nor statistical significance testing, while exploiting different datasets.

While these pitfalls are known and relatively easy to avoid, they are increasingly ignored and accumulated. We believe that a clear, simple, and well-promoted guideline must be defined for automatic MT evaluation. Such a guideline would be useful only if it is adopted by authors and its application is checked by reviewers. For the latter, we also propose a simple scoring method for the meta-evaluation of MT.

Note that the proposed guideline and scoring method only cover the aspects discussed in this paper. Thus, their strict adherence can only guarantee a better evaluation but not a flawless evaluation.

4.2 The Guideline

This guideline and the scoring method that follows are proposed for MT papers that rely on automatic metric scores for evaluating translation quality.

1. An MT evaluation may not exclusively rely on BLEU. Other automatic metrics that better correlate with human judgments, or a human evaluation, may be used in addition or in lieu of BLEU.

2. Statistical significance testing may be performed on automatic metric scores to ensure that the difference between two scores, whatever its amplitude, is not coincidental.

3. Automatic metric scores copied from previous work may not be compared. If inevitable, copied scores may only be compared with scores computed in exactly the same way, through tools guaranteeing this comparability, while providing all the necessary information to reproduce them.

4. Comparisons between MT systems through their metric scores may be performed to demonstrate the superiority of a method or an algorithm only if the systems have been trained, validated, and tested with exactly the same pre-processed data, unless the proposed method or algorithm is indeed dependent on a particular dataset or pre-processing.

The purpose of the following scoring method is to assess the trustworthiness of an automatic evaluation performed in an MT paper. Ultimately, it can be used for authors' self-assessment or by MT program committees to identify trustworthy papers.

Each "yes" answer to the following questions brings 1 point to the paper for a maximum of 4 points.

1. Is a metric that better correlates with human judgment than BLEU used or is a human evaluation performed?

2. Is statistical significance testing performed?

3. Are the automatic metric scores computed for the paper and not copied from other work? If copied, are all the copied and compared scores computed through tools that guarantee their comparability (e.g., SacreBLEU)?

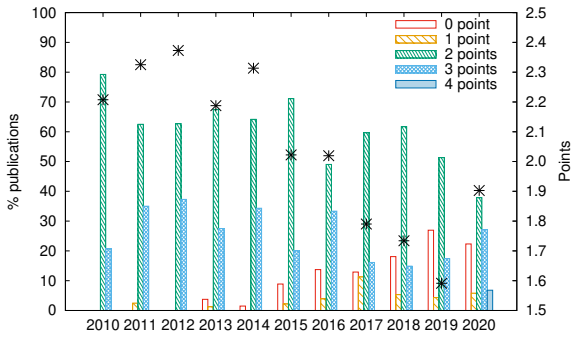

Figure 6: Average meta-evaluation score per year.

4. If comparisons between MT systems are performed to demonstrate the superiority of a method or an algorithm that is independent from the datasets exploited and their preprocessing, are all the compared MT systems exploiting exactly the same pre-processed data for training, validating, and testing? (if not applicable, give 1 point by default)

We scored all the annotated papers, and report on the average score and score distribution for each year in Figure 6. Based on this meta-evaluation, MT evaluation worsens.

5 Conclusion

Our meta-evaluation identified pitfalls in the MT evaluation in most of the annotated papers. The accumulation of these pitfalls and the concerning trends we observed lead us to propose a guideline for automatic MT evaluation. We hope this guideline, or a similar one, will be adopted by the MT community to enhance the scientific credibility of MT research.

This work also has its limitations since it does not cover all the pitfalls of MT evaluation. For instance, we noticed that MT papers regularly rely on the same language pairs to claim general improvements of MT. They also almost exclusively focus on translation from or into English. Another, more positive observation, is that MT papers tend to use stronger baseline systems, following some of the recommendations by Denkowski and Neubig (2017), than at the beginning of the last decade when baseline systems were mostly vanilla MT systems. For future work, we plan to extend our meta-evaluation of MT to publications at conferences in other research domains, such as Machine Learning and Artificial Intelligence.

As a final note, we would like to encourage NLP researchers to perform a similar meta-evaluation in their respective area of expertise. As we showed, it can unveil pitfalls and concerning trends that can be reversed before becoming prevalent.

Acknowledgments

We would like to thank the reviewers for their insightful comments and suggestions. This work was partly supported by JSPS KAKENHI grant numbers 20K19879 and 19H05660.

References

Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, Ann Arbor, USA. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 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*, Trento, Italy. Association for Computational Linguistics.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver, Canada. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on*

Empirical Methods in Natural Language Processing, pages 944–952, Cambridge, USA. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. 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)*, pages 3092–3102, Minneapolis, USA. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 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*, pages 4984–4997, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA. Association for Computational Linguistics.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Stefan Riezler and John T. Maxwell. 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*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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*, pages 223–231, Cambridge, USA. Association for Machine Translation of the Americas.

Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. 2019. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1):1–19.