Difficulty-Aware Machine Translation Evaluation

Runzhe Zhan[*] **Xuebo Liu**[*] **Derek F. Wong**[†] **Lidia S. Chao**

NLP²CT Lab, Department of Computer and Information Science, University of Macau
`nlp2ct.{runzhe,xuebo}@gmail.com, {derekfw,lidiasc}@um.edu.mo`

Abstract

The high-quality translation results produced by machine translation (MT) systems still pose a huge challenge for automatic evaluation. Current MT evaluation pays the same attention to each sentence component, while the questions of real-world examinations (e.g., university examinations) have different difficulties and weightings. In this paper, we propose a novel *difficulty-aware MT evaluation* metric, expanding the evaluation dimension by taking translation difficulty into consideration. A translation that fails to be predicted by most MT systems will be treated as a difficult one and assigned a large weight in the final score function, and conversely. Experimental results on the WMT19 English↔German Metrics shared tasks show that our proposed method outperforms commonly-used MT metrics in terms of human correlation. In particular, our proposed method performs well even when all the MT systems are very competitive, which is when most existing metrics fail to distinguish between them. The source code is freely available at https://github.com/NLP2CT /Difficulty-Aware-MT-Evaluation.

1 Introduction

The human labor needed to evaluate machine translation (MT) evaluation is expensive. To alleviate this, various automatic evaluation metrics are continuously being introduced to correlate with human judgements. Unfortunately, cutting-edge MT systems are too close in performance and generation style for such metrics to rank systems. Even for a metric whose correlation is reliable in most cases, empirical research has shown that it poorly correlates with human ratings when evaluating competitive systems (Ma et al., 2019; Mathur et al., 2020),

---
[*]Equal contribution
[†]Corresponding author

limiting the development of MT systems.

Current MT evaluation still faces the challenge of how to better evaluate the overlap between the reference and the model hypothesis taking into consideration *adequacy* and *fluency*, where all the evaluation units are treated the same, i.e., all the matching scores have an equal weighting. However, in real-world examinations, the questions vary in their difficulty. Those questions which are easily answered by most subjects tend to have low weightings, while those which are hard to answer have high weightings. A subject who is able to solve the more difficult questions can receive a high final score and gain a better ranking. MT evaluation is also a kind of examination. For bridging the gap between human examination and MT evaluation, it is advisable to incorporate a *difficulty* dimension into the MT evaluation metric.

In this paper, we take translation difficulty into account in MT evaluation and test the effectiveness on a representative MT metric BERTScore (Zhang et al., 2020) to verify the feasibility. More specifically, the difficulty is first determined across the systems with the help of pairwise similarity, and then exploited as the weight in the final score function for distinguishing the contribution of different sub-units. Experimental results on the WMT19 English↔German evaluation task show that difficulty-aware BERTScore has a better correlation than do the existing metrics. Moreover, it agrees very well with the human rankings when evaluating competitive systems.

2 Related Work

The existing MT evaluation metrics can be categorized into the following types according to their underlying matching sub-units: n-gram based (Papineni et al., 2002; Doddington, 2002; Lin and Och, 2004; Han et al., 2012; Popović, 2015),
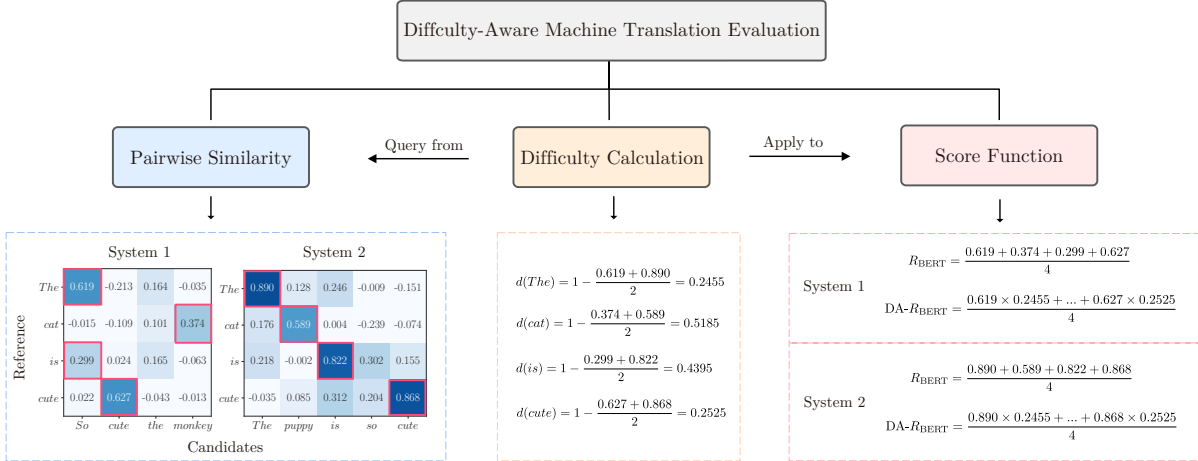
Figure 1: Illustration of combining difficulty weight with BERTScore. R_{BERT} denotes the vanilla recall-based BERTScore while DA-R_{BERT} denotes the score augmented with translation difficulty.

edit-distance based (Snover et al., 2006; Leusch et al., 2006), alignment-based (Banerjee and Lavie, 2005), embedding-based (Zhang et al., 2020; Chow et al., 2019; Lo, 2019) and end-to-end based (Sellam et al., 2020). BLEU (Papineni et al., 2002) is widely used as a vital criterion in the comparison of MT system performance but its reliability has been doubted on entering neural machine translation age (Shterionov et al., 2018; Mathur et al., 2020). Due to the fact that BLEU and its variants only assess surface linguistic features, some metrics leveraging contextual embedding and end-to-end training bring semantic information into the evaluation, which further improves the correlation with human judgement. Among them, BERTScore (Zhang et al., 2020) has achieved a remarkable performance across MT evaluation benchmarks balancing speed and correlation. In this paper, we choose BERTScore as our testbed.

3 Our Proposed Method

3.1 Motivation

In real-world examinations, the questions are empirically divided into various levels of difficulty. Since the difficulty varies from question to question, the corresponding role a question plays in the evaluation does also. Simple question, which can be answered by most of the subjects, usually receive of a low weighting. But a difficult question, which has more discriminative power, can only be answered by a small number of good subjects, and thus receives a higher weighting.

Motivated by this evaluation mechanism, we measure difficulty of a translation by viewing the

MT systems and sub-units of the sentence as the subjects and questions, respectively. From this perspective, the impact of the sentence-level sub-units on the evaluation results supported a differentiation. Those sub-units that may be incorrectly translated by most systems (e.g., polysemy) should have a higher weight in the assessment, while easier-to-translate sub-units (e.g., the definite article) should receive less weight.

3.2 Difficulty-Aware BERTScore

In this part, we aim to answer two questions: 1) how to automatically collect the translation difficulty from BERTScore; and 2) how to integrate the difficulty into the score function. Figure 1 presents an overall illustration.

Pairwise Similarity Traditional n-gram overlap cannot extract semantic similarity, word embedding provides a means of quantifying the degree of overlap, which allows obtaining more accurate difficulty information. Since BERT is a strong language model, it can be utilized as a contextual embedding $\mathbf{O}_{\text{BERT}}$ (i.e., the output of BERT) for obtaining the representations of the reference $\mathbf{t}$ and the hypothesis $\mathbf{h}$. Given a specific hypothesis token h and reference token t, the similarity score $\text{sim}(t, h)$ is computed as follows:

$$\text{sim}(t, h) = \frac{\mathbf{O}_{\text{BERT}}(t)^{\mathsf{T}}\mathbf{O}_{\text{BERT}}(h)}{\|\mathbf{O}_{\text{BERT}}(t)\| \cdot \|\mathbf{O}_{\text{BERT}}(h)\|} \quad (1)$$

Subsequently, a similarity matrix is constructed by pairwise calculating the token similarity. Then the token-level matching score is obtained by greedily

Metric	En→De (All)			En→De (Top 30%)			De→En (All)			De→En (Top 30%)		
	$|r|$	$|\tau|$	$|\rho|$	$|r|$	$|\tau|$	$|\rho|$	$|r|$	$|\tau|$	$|\rho|$	$|r|$	$|\tau|$	$|\rho|$
BLEU	0.952	0.703	0.873	0.460	0.200	0.143	0.888	0.622	0.781	0.808	0.548	0.632
TER	0.982	0.711	0.873	0.598	0.333	0.486	0.797	0.504	0.675	**0.883**	0.548	0.632
METEOR	0.985	0.746	0.904	0.065	0.067	0.143	0.886	0.605	0.792	0.632	0.548	0.632
BERTScore	0.990	0.772	0.920	0.204	0.067	0.143	0.949	0.756	0.890	0.271	0.183	0.316
DA-BERTScore	**0.991**	**0.798**	**0.930**	**0.974**	**0.733**	**0.886**	**0.951**	**0.807**	**0.932**	0.693	**0.548**	**0.632**

Table 1: Absolute correlations with system-level human judgments on WMT19 metrics shared task. For each metric, higher values are better. Difficulty-aware BERTScore consistently outperforms vanilla BERTScore across different evaluation metrics and translation directions, especially when the evaluated systems are very competitive (i.e., evaluating on the top 30% systems).

searching for the maximal similarity in the matrix, which will be further taken into account in sentence-level score aggregation.

Difficulty Calculation The calculation of difficulty can be tailored for different metrics based on the overlap matching score. In this case, BERTScore evaluates the token-level overlap status by the pairwise semantic similarity, thus the token-level similarity is viewed as the bedrock of difficulty calculation. For instance, if one token (like "cat") in the reference may only find identical or synonymous substitutions in a few MT system outputs, then the corresponding translation difficulty weight ought to be larger than for other reference tokens, which further indicates that it is more valuable for evaluating the translation capability. Combined with BERTScore mechanism, it is implemented by averaging the token similarities across systems. Given K systems and their corresponding generated hypotheses $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K$, the difficulty of a specific token t in the reference $\mathbf{t}$ is formulated as

$$d(t) = 1 - \frac{\sum_{k=1}^{K} \max_{h \in \mathbf{h}_k} \text{sim}(t, h)}{K} \quad (2)$$

An example is shown in Figure 1: the entity "cat" is improperly translated to "monkey" and "puppy", resulting in a lower pairwise similarity of the token "cat", which indicates higher translation difficulty. Therefore, by incorporating the translation difficulty into the evaluation process, the token "cat" is more contributive while the other words like "cute" are less important in the overall score.

Score Function Due to the fact that the translation generated by a current NMT model is fluent enough but not adequate yet, F-score which takes into account the *Precision* and *Recall*, is more appropriate to aggregate the matching scores, instead

of only considering precision. We thus follow vanilla BERTScore in using F-score as the final score. The proposed method directly assigns difficulty weights to the counterpart of the similarity score **without any hyperparameter**:

$$\text{DA-}R_{\text{BERT}} = \frac{1}{|\mathbf{t}|} \sum_{t \in \mathbf{t}} d(t) \max_{h \in \mathbf{h}} \text{sim}(t, h) \quad (3)$$

$$\text{DA-}P_{\text{BERT}} = \frac{1}{|\mathbf{h}|} \sum_{h \in \mathbf{h}} d(h) \max_{t \in \mathbf{t}} \text{sim}(t, h) \quad (4)$$

$$\text{DA-}F_{\text{BERT}} = 2 \cdot \frac{\text{DA-}R_{\text{BERT}} \cdot \text{DA-}P_{\text{BERT}}}{\text{DA-}R_{\text{BERT}} + \text{DA-}P_{\text{BERT}}} \quad (5)$$

For any $h \notin \mathbf{t}$, we simply let $d(h) = 1$, i.e., retaining the original calculation. The motivation is that the human assessor keeps their initial matching judgement if the test taker produces a unique but reasonable alternative answer. We regard DA-F_{BERT} as the DA-BERTScore in the following part.

There are many variants of our proposed method: 1) designing more elaborate difficulty function (Liu et al., 2020; Zhan et al., 2021); 2) applying a smoothing function to the difficulty distribution; and 3) using other kinds of F-score, e.g., $F_{0.5}$-score. The aim of this paper is not to explore this whole space but simply to show that a straightforward implementation works well for MT evaluation.

4 Experiments

Data The WMT19 English↔German (En↔De) evaluation tasks are challenging due to the large discrepancy between human and automated assessments in terms of reporting the best system (Bojar et al., 2018; Barrault et al., 2019; Freitag et al., 2020). To sufficiently validate the effectiveness of

SYSTEM	BLEU ↑	TER ↓	METEOR ↑	BERTScore ↑	DA-BERTScore ↑	HUMAN ↑
Facebook.6862	0.4364 (⇓5)	0.4692 (⇓5)	0.6077 (⇓3)	0.7219 (⇓4)	**0.1555 (✓0)**	**0.347**
Microsoft.sd.6974	0.4477 (⇓1)	0.4583 (⇓1)	0.6056 (⇓3)	0.7263 (✓0)	0.1539 (⇓1)	0.311
Microsoft.dl.6808	0.4483 (⇑1)	0.4591 (⇓1)	0.6132 (⇑1)	0.7260 (✓0)	0.1544 (⇑1)	0.296
MSRA.6926	**0.4603 (⇑3)**	**0.4504 (⇑3)**	**0.6187 (⇑3)**	**0.7267 (⇑3)**	0.1525 (✓0)	0.214
UCAM.6731	0.4413 (✓0)	0.4636 (✓0)	0.6047 (⇓1)	0.7190 (⇓1)	0.1519 (⇓1)	0.213
NEU.6763	0.4460 (⇑2)	0.4563 (⇑4)	0.6083 (⇑3)	0.7229 (⇑2)	0.1521 (⇑1)	0.208
sum($|\triangle_{\text{Rank}}|$)	12	14	14	10	**4**	0

Table 2: Agreement of system ranking with human judgement on the top 30% systems (k=6) of WMT19 En→De Metrics task. ⇑/⇓ denotes that the rank given by the evaluation metric is higher/lower than human judgement, and ✓ denotes that the given rank is equal to human ranking. DA-BERTScore successfully ranks the best system that the other metrics failed. Besides, it also shows the lowest rank difference.
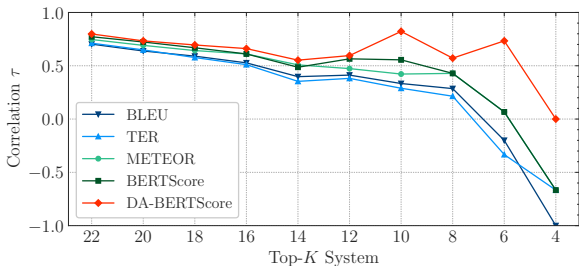

Figure 2: Effect of top-K systems in the En→De evaluation. DA-BERTScore is highly correlated with human judgment for different values of K, especially when all the systems are competitive (i.e., $K \leq 10$).

our approach, we choose these tasks as our evaluation subjects. There are 22 systems for En→De and 16 for De→En. Each system has its corresponding human assessment results. The experiments were centered on the correlation with system-level human ratings.

Comparing Metrics In order to compare with the metrics that have different underlying evaluation mechanism, four representative metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2014), BERTScore (Zhang et al., 2020), which are correspondingly driven by n-gram, edit distance, word alignment and embedding similarity, are involved in the comparison experiments without losing popularity. For ensuring reproducibility, the original[1][2] and widely used implementation[3] was used in the experiments.

Main Results Following the correlation criterion adopted by the WMT official organization, Pearson's correlation r is used for validating the system-

level correlation with human ratings. In addition, two rank-correlations Spearman's ρ and original Kendall's τ are also used to examine the agreement with human ranking, as has been done in recent research (Freitag et al., 2020). Table 1 lists the results. DA-BERTScore achieves competitive correlation results and further improves the correlation of BERTScore. In addition to the results on all systems, we also present the results on the top 30% systems where the calculated difficulty is more reliable and our approach should be more effective. The result confirms our intuition that DA-BERTScore can significantly improve the correlations under the competitive scenario, e.g., improving the $|r|$ score from 0.204 to 0.974 on En→De and 0.271 to 0.693 on De→En.

Effect of Top-K Systems Figure 2 compares the Kendall's correlation variation of the top-K systems. Echoing previous research, the vast majority of metrics fail to correlate with human ranking and even perform negative correlation when K is lower than 6, meaning that the current metrics are ineffective when facing competitive systems. With the help of difficulty weights, the degradation in the correlation is alleviated, e.g., improving τ score from 0.07 to 0.73 for BERTScore ($K = 6$). These results indicate the effectiveness of our approach, establishing the necessity for adding difficulty.

Case Study of Ranking Table 2 presents a case study on the En→De task. Existing metrics consistently select MSRA's system as the best system, which shows a large divergence from human judgement. DA-BERTScore ranks it the same as human (4th) because most of its translations have low difficulty, thus lower weights are applied in the scores. Encouragingly, DA-BERTScore ranks Facebook's system as the best one, which implies that it overco-

---

	BERTS.	+DA	Sentence
Src	-	-	"I'm standing **right here** in front of you," one woman said.
Ref	-	-	„Ich stehe genau hier vor Ihnen ", sagte eine Frau.
MSRA	**0.9656**	0.0924	„Ich stehe **hier vor** Ihnen ", sagte eine Frau.
Facebook	0.9591	**0.1092**	„Ich stehe **hier direkt vor** Ihnen ", sagte eine Frau.
Src	-	-	France has more than 1,000 **troops on the ground** in the war-wracked country.
Ref	-	-	Frankreich hat über 1.000 Bodensoldaten in dem kriegszerstörten Land im Einsatz.
MSRA	**0.6885**	0.2123	Frankreich hat mehr als 1.000 **Soldaten vor Ort** in dem kriegsgeplagten Land.
Facebook	0.6772	**0.2414**	Frankreich hat mehr als 1000 **Soldaten am Boden** in dem kriegsgeplagten Land stationiert.

Table 3: Examples from the En→De evaluation. BERTS. denotes BERTScore. Words indicate the difficult translations given by our approach on the top 30% systems. DA-BERTScores are more in line with human judgements.
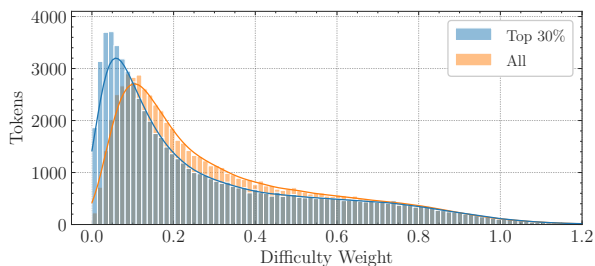

Figure 3: Distribution of token-level difficulty weights extracted from the En→De evaluation.

mes more challenging translation difficulties. This testifies to the importance and effectiveness of considering translation difficulty in MT evaluation.

Case Study of Token-Level Difficulty Table 3 presents two cases, illustrating that our proposed difficulty-aware method successfully identifies the omission errors ignored by BERTScore. In the first case, the Facebook's system correctly translates the token "right", and in the second case, uses the substitute "Soldaten am Boden" which is lexically similar to the ground-truth token "Bodensoldaten". Although the MSRA's system suffers word omissions in the two cases, its hypotheses receive the higher ranking given by BERTScore, which is inconsistent with human judgements. The reason might be that the semantic of the hypothesis is highly close to the reference, thus the slight lexical difference is hard to be found when calculating the similarity score. By distinguishing the difficulty of the reference tokens, DA-BERTScore successfully makes the evaluation focus on the difficult parts, and eventually correct the score of the Facebook's system, thus giving the right rankings.

Distribution of Difficulty Weights The difficulty weights can reflect the translation ability of a group of MT systems. If the systems in a group are of higher translation ability, the calculated dif-

ficulty weights will be smaller. Starting from this intuition, we visualize the distribution of difficulty weights as shown in Figure 3. Clearly, we can see that the difficulty weights are centrally distributed at lower values, indicating that most of the tokens can be correctly translated by all the MT systems. For the difficulty weights calculated on the top 30% systems, the whole distribution skews to zero since these competitive systems have better translation ability and thus most of the translations are easy for them. This confirms that the difficulty weight produced by our approach is reasonable.

5 Conclusion and Future Work

This paper introduces the conception of difficulty into machine translation evaluation, and verifies our assumption with a representative metric BERTScore. Experimental results on the WMT19 English↔German metric tasks show that our approach achieves a remarkable correlation with human assessment, especially for evaluating competitive systems, revealing the importance of incorporating difficulty into machine translation evaluation. Further analyses show that our proposed difficulty-aware BERTScore can strengthen the evaluation of word omission problems and generate reasonable distributions of difficulty weights.

Future works include: 1) optimizing the difficulty calculation; 2) applying to other MT metrics; and 3) testing on other generation tasks, e.g., speech recognition and text summarization.

Acknowledgement

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, Michigan. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*, pages 441–450, Mumbai, India. The COLING 2012 Organizing Committee.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. 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, Pennsylvania, 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.

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

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3):217–235.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021. Meta-curriculum learning for domain adaptation in neural machine translation. In *the 35th AAAI Conference on Artificial Intelligence, AAAI2021*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.