Investigating Data Variance in Evaluations of Automatic Machine Translation Metrics

Jiannan Xiang[♡*], Huayang Li[♠], Yahui Liu[★], Lemao Liu[♣], Guoping Huang[♣], Defu Lian[♢], Shuming Shi[♣]

[♡]Carnegie Mellon University, [♠]Nara Institute of Science and Technology

[♢]University of Science and Techology of China

[★]University of Trento, Italy, [♣]Tencent AI Lab

jiannanx@cs.cmu.edu, li.huayang.lh6@is.naist.jp

yahui.liu@unitn.it, lemaoliu@gmail.com, {donkeyhuang, shumingshi}@tencent.com

Abstract

Current practices in metric evaluation focus on one single dataset, *e.g.*, Newstest dataset in each year's WMT Metrics Shared Task. However, in this paper, we qualitatively and quantitatively show that the performances of metrics are sensitive to data. The ranking of metrics varies when the evaluation is conducted on different datasets. Then this paper further investigates two potential hypotheses, *i.e.*, insignificant data points and the deviation of Independent and Identically Distributed (i.i.d) assumption, which may take responsibility for the issue of data variance. In conclusion, our findings suggest that when evaluating automatic translation metrics, researchers should take data variance into account and be cautious to claim the result on a single dataset, because it may leads to inconsistent results with most of other datasets.

1 Introduction

Assessing the quality of machine translation (MT) systems is always crucial to promote MT research (Marie et al., 2021). Since it is costly and time-consuming for human graders to evaluate machine translation (MT) systems, designing automatic metrics for MT has drawn booming attention during the past decades, and many metrics such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) have been proposed consequently.

Generally, it is non-trivial to measure automatic metrics. Conference Machine Translation (WMT) (Ma et al., 2019, 2018; Macháček and Bojar, 2013a,b; Bojar et al., 2016) thereby holds the Metric Shared Task to evaluate the performance of automatic metrics. In each year, WMT organizers collect a dataset consisting of many MT outputs annotated with human judgments, and automatic metrics are evaluated on the dataset in terms of their correlations to human judgments. Over the past ten years, the official evaluation reports only independently analyzed the results of that year. To the best of our knowledge, there are no studies to put the evaluation results of ten years together and make a more systematic analysis. Therefore, some key questions remain unknown: are the evaluation results consistent across different years? Are the results on each dataset reliable?

One may simply summarize the existing results from the official evaluation reports of the past years and answer the above questions accordingly. However, the existing results use Pearson's correlation for evaluation which suffers from sensitivity to outlier data points as argued by Mathur et al. (2020). Besides, involved metrics in the evaluation are different year by year, thus it is difficult to directly compare the results among different years. To this end, in this work, we firstly re-evaluate ten popular metrics on all available datasets in the past ten years, with the Error Number evaluation method (Mathur et al., 2020). We then empirically investigate the fluctuation of metric evaluation results. Surprisingly, our experiments show that the evaluation result is sensitive to the choice of datasets, which suggests that the results on some datasets may not be reliable (§3).

Then we investigate two potential hypotheses about the emergence of data variance, i.e., the insignificant data points (§4.1) and deviation of Independent and Identically Distributed (i.i.d) assumption (§4.2). First, we show that the data variance issue is substantially alleviated when the insignificant data points are removed. To further understand the variance that cannot be alleviated by the first hypothesis, we design a simple method to measure the distributional differences between datasets, and hypothesize that the deviation of the i.i.d assumption may contribute to the variance. For future metric evaluation, we recommend WMT community pay attention to the potential issue of data variance

[*] Work done while J. Xiang was an intern at Tencent AI Lab.

Findings of the Association for Computational Linguistics: ACL 2022, pages 150 - 157

Dataset	Size	System Number	Link
Newssyscombtest2010	2,034	31	http://www.statmt.org/wmt10/results.html
Newssyscombtest2011	2,000	26	http://www.statmt.org/wmt11/results.html
Newstest2012	3,003	16	http://www.statmt.org/wmt12/results.html
Newstest2013	3,000	23	http://www.statmt.org/wmt13/results.html
Newstest2014	3,003	13	http://www.statmt.org/wmt14/results.html
Newstest2015	2,169	13	http://www.statmt.org/wmt15/results.html
Newstest2016	2,999	10	http://www.statmt.org/wmt16/results.html
Newstest2017	3,004	11	http://www.statmt.org/wmt17/results.html
Newstest2018	2,998	16	http://www.statmt.org/wmt18/results.html
Newstest2019	2,000	16	http://www.statmt.org/wmt19/results.html

Table 1: The data statistics for German-English language pair.

when conducting evaluations.

Metrics	Features	Average Type
BLEU	n-grams	macro
WER	Levenshtein distance	macro
TER	edit distance	macro
PER	edit distance	macro
chrF	character n-grams	micro
chrF+	character n-grams	micro
BEER	char. n-grams, trees	micro
CharacTER	char. edit distance	micro
BERTScore	neural representations	micro
MoverScore	neural representations	micro

Table 2: Features for the metrics we use in the paper. Note that we implement PER by ourselves.

2 Experiment Settings

2.1 Datasets and evaluation metrics

We collect the testing set data and the human assessments of the WMT Metrics Task from 2010 to 2019. In this work, we mainly conduct experiments on the De⇒En task and more details about datasets are shown in Table 1. However, as shown in §3.1, our conclusions are consistent on other translation tasks, such as Ru⇒En.

Since participating metrics in the WMT Metrics Task varied over years, we collect ten most popular metrics and re-evaluate them on all ten datasets to study their performance.These metrics are summarized as follows: BLEU (Papineni et al., 2002), WER (Morris et al., 2004), PER (Tillmann et al., 1997), TER (Snover et al., 2006), chrF (Popović, 2015), chrF+ (Popović, 2017), BEER (Stanojević and Sima'an, 2014), CharacTER (Wang et al., 2016), BERTScore (Zhang et al., 2020), and MoverScore (Zhao et al., 2019). The first 4 metrics are in system-level (i.e., macro) while others are in sentence-level (i.e., micro), as shown in Table 2. Since extending sentence-level metrics to system-level is more natural (Zhang et al., 2020), we only compare them on the system-level.

For each system pair, metrics or humans give a comparison result about whether one system is better than another. Following Graham et al. (2014), we use statistical significance tests to detect if the difference in scores (metrics or humans) between two systems is significant. Specifically, for RR scores, we use the bootstrap method (Koehn, 2004). For DA scores, we apply the Wilcoxon rank-sum test. For macro-average metrics, *i.e.*, BLEU, WER, PER, and TER, we use the bootstrap method (Koehn, 2004). For other micro-average metrics, we use the paired t-test method.

2.2 Measuring Automatic Metrics

The previous WMT Metrics Tasks used Pearson's r to measure the ability of a metric to evaluate MT systems. However, as pointed out by Mathur et al. (2020), Pearson's r is unstable for a small sample size and sensitive to outlier systems. Besides, in practice, metric scores are always used to compare pairs of MT systems[1]. Thus following Mathur et al. (2020), we measure an automatic metric with the number of errors made by the metric when comparing system pairs. Error Number can be considered as an absolute view of measuring a metric.

Error Number Following Mathur et al. (2020), we measure the performance of a metric by its consistency with humans. Specifically, each metric or human can judge whether a system performs better compared to another system (details of system comparison process are presented in the appendix), and the error number is the number of contrary cases between the results of metric and human. As mentioned by Graham and Liu (2016), when the number of compared MT systems are too small on a dataset, differences among different metrics

[1]Unless otherwise specified, a system always denotes MT system in our work, rather than an evaluation metric.

Metric	Dataset									
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
BERTScore	**1** / 24.4	**1** / 37.1	**2** / 28.9	**1** / 10.6	**2** / 20.4	**1** / 14.7	**1** / 14.5	**6** / 24.6	**2** / 15.3	**3** / 37.0
CharacTER	**6** / 27.6	**1** / 37.1	**1** / 24.2	**6** / 18.0	**1** / 17.3	**1** / 14.7	**3** / 17.6	**1** / 20.8	**1** / 14.4	**4** / 38.2
MoverScore	**2** / 25.2	**4** / 39.3	**2** / 28.8	**2** / 11.7	**2** / 20.3	**1** / 14.7	**2** / 16.0	**5** / 23.9	**2** / 15.4	**1** / 36.6
chrF	**3** / 26.7	**1** / 37.8	**4** / 29.7	**2** / 12.1	**2** / 20.8	**4** / 17.7	**4** / 18.9	**2** / 22.9	**2** / 15.3	**1** / 37.0
BEER	**3** / 26.3	**5** / 45.3	**5** / 33.5	**4** / 13.4	**6** / 25.0	**5** / 19.0	**5** / 19.5	**2** / 23.2	**2** / 15.2	**6** / 38.4
chrF+	**3** / 26.9	**5** / 45.8	**6** / 35.1	**4** / 13.8	**7** / 26.4	**6** / 19.2	**5** / 20.2	**2** / 23.3	**2** / 15.2	**4** / 37.7
BLEU	**8** / 32.3	**8** / 58.3	**8** / 42.3	**7** / 20.9	**8** / 29.3	**7** / 23.1	**8** / 21.2	**7** / 26.3	**9** / 18.1	**7** / 41.3
WER	**7** / 31.7	**7** / 57.7	**7** / 40.8	**8** / 23.4	**9** / 32.3	**7** / 22.9	**5** / 19.7	**8** / 27.2	**7** / 17.0	**7** / 40.9
TER	**9** / 35.0	**9** / 61.2	**9** / 43.9	**9** / 24.7	**10** / 36.2	**7** / 22.7	**8** / 20.9	**10** / 28.6	**7** / 17.2	**9** / 43.0
PER	**10** / 38.6	**9** / 61.7	**10** / 48.0	**10** / 26.9	**5** / 23.8	**7** / 22.8	**10** / 28.2	**8** / 27.6	**9** / 18.4	**9** / 43.5

Table 3: Metric evaluation results on De⇒En datasets from 2010 to 2019. The tuple "**R** / E" shows the performance of a metric, where **R** denotes Significant Ranking (§2.3) among all metrics and E denotes the Error Rate (Error Number divided by the total number of system pairs).

may be insignificant. Thus, the results of the metric evaluation can be highly inconclusive. We indeed observe similar results in our experimental setting. Therefore, we use the hybrid super-sampling method (Graham and Liu, 2016) to create a large number of hybrid system pairs: on each dataset, we synthesize 142 systems in total, which form 10K system pairs. Finally, we calculate the error number of each metric on all 10K system pairs.

2.3 Measuring Data Variance

Significant Ranking Based on the measurement of error number, a qualitative approach to know whether those metrics perform consistently on different datasets is to evaluate the variance of their rankings. To make the ranking more reliable, we propose a *significant ranking* method, which conducts significant test when sorting the error numbers of metrics. For example, in Table 3, the significant ranking of all metrics on 2010 dataset is "1, 6, 2, 3, 3, 3, 8, 7, 9, 10" where chrF, BEER and chrF+ are with the same relative ranking of 3. This is because they are not significantly different, although their absolute error numbers are slightly different. We employ the bootstrap re-sampling method (Koehn, 2004) to test if the number of errors of one metric is significantly less than the others. For the bootstrap method, we repeat re-sampling 1000 times and set the p-value to 0.05 for all the significance tests.

Disagreement Number In addition, we also propose a method to quantitatively measure the variance between two datasets $\mathcal{D}$ and D', namely, *disagreement number*. Specifically, we construct a set $\mathcal{S}_D$ by collecting all pairwise metrics that one is significantly better than the other on dataset $\mathcal{D}$. Then to measure the mismatch between $\mathcal{D}$ and $\mathcal{D}'$,

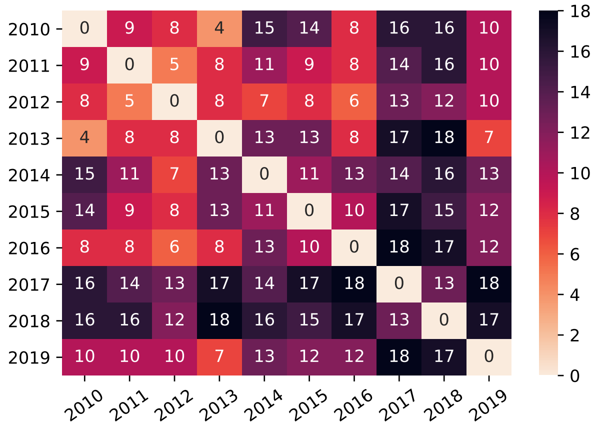

Figure 1: The heatmap for the disagreement numbers between every two datasets on De⇒En task.

we count the **disagreement number** between the pairwise metrics in $\mathcal{S}_{\mathcal{D}}$ and that in $\mathcal{S}_{\mathcal{D}'}$. For example, disagreement number plus one, if BLEU is significantly better than TER on $\mathcal{D}$ and worse than TER on $\mathcal{D}'$. Although this number is linear to Kendall's Tau (Kendall, 1938), it is able to show more informative difference between two overall rankings. For example, two metrics with totally different ranks may just have a slight difference on disagreement number. As a result, we employ disagreement number rather than Kendall's Tau to show the quantitative difference between two overall rankings more intuitively for more detailed analysis. It is worth mentioning that the disagreement number is at most 45 in our setting where there are 10 metrics in total.

3 Data Variance in Metric Evaluations

3.1 Variance of Different Datasets

We conduct experiments on all 10 datasets. We have 10 metrics, which can form 45 metric pairs. On each dataset, for each metric, we calculate its

Metric	Dataset				
	2015	2016	2017	2018	2019
BERTScore	**2** / 18.0	**3** / 16.7	**2** / 32.1	**3** / 24.5	**1** / 35.8
CharacTER	**6** / 20.5	**6** / 19.4	**1** / 30.4	**1** / 22.3	**4** / 37.2
MoverScore	**1** / 14.9	**1** / 15.1	**2** / 31.4	**3** / 24.5	**1** / 36.1
chrF	**3** / 18.7	**2** / 15.7	**2** / 31.9	**2** / 24.0	**4** / 37.1
BEER	**3** / 19.0	**3** / 17.0	**5** / 33.2	**3** / 24.3	**9** / 39.6
chrF+	**5** / 19.7	**5** / 17.6	**5** / 33.3	**3** / 24.5	**4** / 36.9
BLEU	**10** / 27.9	**7** / 21.0	**9** / 34.8	**8** / 25.0	**9** / 39.8
WER	**8** / 23.4	**7** / 21.5	**5** / 33.2	**8** / 25.0	**8** / 37.9
TER	**8** / 23.4	**10** / 23.3	**9** / 34.5	**10** / 25.8	**4** / 36.9
PER	**6** / 21.1	**7** / 21.5	**7** / 34.0	**2** / 23.4	**1** / 35.7

Table 4: Metric evaluation results on Ru⇒En datasets from 2015 - 2019.

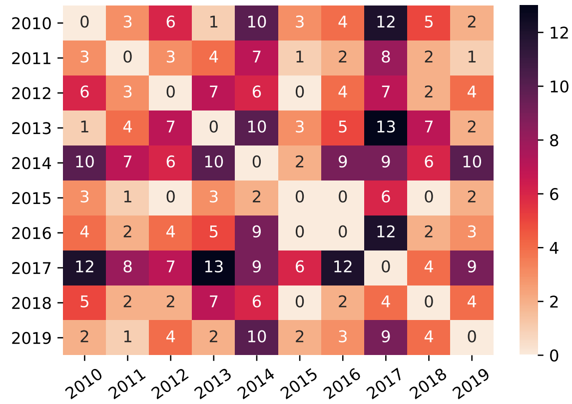

Figure 2: The heatmap for the disagreement numbers between every two datasets on De⇒En task. Insignificant system pairs according to human assessments are removed.

error number (described in Section 2.2). In addition, we perform a statistical significance test for each metric pairs in terms of both error numbers, from which we can obtain a ranking result among 10 metrics accordingly.

Table 3 illustrates the error numbers and ranks on 10 datasets. It shows that the ranks are always variant along with different datasets. For example, on the dataset of 2011, the error rate of MoveScore is larger than chrF (39.3 v.s. 37.8), and the former ranks 4 while the latter ranks 1. However, it is opposite on the dataset of 2015, where MoveScore ranks 1 with an error rate of 14.7 while chrF ranks 4 with an error rate of 17.7. As shown in Table 4, we observe a similar trend on the Ru⇒En task.

As shown in Figure 1, there is a high inconsistency between the results of different datasets and none of the dataset pairs achieve zero disagreements. The difference between the datasets in 2010 and 2013 is the smallest (i.e., only 4 disagreed metric pairs). However, most of the disagreement numbers are larger than 10 (the maximum achieves 18). Moreover, datasets from 2017 to 2019 not only have a high disagreement number with datasets of early years, but also have high variances among themselves. This finding is a little surprising, because in our sense the quality of WMT's datasets must be improved year by year.

4 Potential Reasons for Data Variance

Many factors may contribute to the data variance issue, but lots of them are difficult to be evaluated, such as the personal preferences of humans. In this section, we propose to analyze two potential factors that can be quantitatively evaluated.

4.1 Insignificant Data Points

Intuitively, if the translations H_A from system A are much better than those H_B from system B in translation quality according to human evaluation, then it is easy to judge the better system even for a weak automatic metric. In contrast, if H_A is similar to H_B in translation quality, it is typically difficult to judge the better system even for a good metric. This motivates us that such an insignificant data point $\langle H_A, H_B \rangle$ may take responsibility for the data variance issue.

To validate the above intuition, we remove the system pairs that are not significantly different according to human evaluation, and compute the disagreement number between any two datasets again. The experimental results are shown in Figure 2. We can see the disagreement number decreases greatly comparing to the results in Figure 1. In the previous experiment, most of the disagreement numbers are greater than 10, while in the new experiment most of them are less than 5, and some of them even achieve 0, such as the number between 2012 and 2015, which means the ranks of metrics are exactly the same on those datasets. The results indicate that part of the data variance issue can be explained by system pairs that humans think are not significantly different.

However, After the removal of insignificant data points, some disagreement number are still high, e.g., the number between 2013 and 2017 is 13. It demonstrates that there are still some other underlying problems that give rise to the data variance phenomenon. In addition, the datasets for both 2014 and 2017 do not agree well with others. This indicates that we should be cautious to report overall results on some datasets, e.g., 2014 and 2017.

4.2 Deviation of I.I.D Assumption

How to interpret the high variance on datasets, e.g., 2014 and 2017, remains to be an open question. In this section we try to give a hypothesis based on the i.i.d assumption. According to the principle of statistical sampling, if two samples are drawn from the same distribution, then a statement made on one sample is likely to hold on the other sample. Therefore, one hypothesis about the high variance may be that datasets from different years deviate i.i.d assumption. In fact, this may be true in our scenario because each dataset is generated by a set of translation systems but the set of systems is variant each year.

To this end, we design an experiment to measure the extent to which each dataset is drawn from the same distribution during the past ten years. Since the standard method such as Kolmogorov-Smirnov test (Massey Jr, 1951) is difficult to scale with respect to feature dimension, we employ adversarial validation to distinguish the difference between two datasets (Pan et al., 2020). Its basic idea is to formulate the i.i.d test as a classification problem and train a classifier between two datasets. If the classifier can accurately distinguish the data from different datasets, then the distributions of the two datasets are regarded as highly different. Since it is too slow to train classifiers for all pairs of datasets, we conduct experiments on three years from 2017 to 2019. More details are shown in appendix.

The results on two kinds of datasets are shown in Table 5, where higher accuracy indicates clearer distributional differences between two datasets. Note that accuracy scores in main diagonal are got from two subsets of each year via randomly splitting. As shown in Table 5, the distributional differences between MT datasets have been introduced by source sentences. After accompanied with the system outputs, the distributional differences are more severe between different years. This fact shows that some datasets in past ten years deviate the i.i.d assumption, which may be related to the inconsistency of metrics.

4.3 Suggestions

According to those potential factors, we propose some suggestions to alleviate some potential issues for metric evaluation due to data variance in future. First, it would be better if pay more attention to those insignificant data points such that their manual annotations are as good as possible. Since it is costly to hire more annotators for data points, it would be possible to ask more annotators only for those insignificant data points. Second, it would be helpful to construct a dataset with diverse domains and explicitly show the evaluation results for each subset with the same domain rather than a single evaluation result for the entire dataset. Generally, although inconsistent results from different domains are possible, however, the inconsistency in the same domain may be undesirable. Thus, showing the domain information could help researchers to better promote the datasets and metrics.

	17	18	19
17	50.4	52.8	65.8
18	52.8	51.4	67.5
19	65.8	67.5	50.9

(a) Src

	17	18	19
17	51.4	75.3	80.2
18	75.3	55.6	79.2
19	80.2	79.2	52.2

(b) Src+Output

Table 5: The accuracy of classifiers. The higher value means two datasets deviate i.i.d assumption. We run the model with 5 different random seeds to calculate the average accuracy.

5 Conclusion

Over the past ten years, the official evaluation reports of WMT Metrics Shared Task only independently analyzed the results of that year. In this paper, we re-evaluate ten popular metrics on all available datasets in the past ten years and comparatively analyze the evaluation results among different years together. We show the problem of conducting evaluations with only one dataset. In addition, we analyze its potential reasons that the insignificant data points and deviation of i.i.d assumption may induce the issue of data variance. This fact suggests that future researches on evaluating automatic translation metrics should take data variance into account and be cautious to conclude the result on a single dataset.

Acknowledgement

We would like to thank the anonymous reviewers for their constructive comments.

References

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of wmt evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yvette Graham and Qun Liu. 2016. Achieving accurate conclusions in evaluation of automatic machine translation metrics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–10, San Diego, California. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

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.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. 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*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013a. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013b. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. 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 and the 11th International Joint Conference on Natural Language Processing*, pages 7297–7306, Online. Association for Computational Linguistics.

Frank J Massey Jr. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

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.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Jing Pan, Vincent Pham, Mohan Dorairaj, Huigang Chen, and Jeong-Yoon Lee. 2020. Adversarial validation approach to concept drift problem in user targeting automation systems at uber. *arXiv preprint arXiv:2004.03045*.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Settings for Adversarial Validation

To train the classifier, we need to construct a binary classification dataset first. Since the difference between distributions may come from both the source sentences and system outputs, we consider two types of classification datasets correspondingly. The first kind of dataset only considers the source information. Supposing that we want to compare the distributions of source sentences of MT datasets from year Y1 and Y2, we follow the three steps below to construct the classification dataset:

1. We label the source sentences from Y1 and Y2 with 0 and 1, respectively;

2. We split the data from Y1 and Y2 to train, dev, and test sets without overlap;

3. We merge the data from Y1 and Y2 according to their split.

For each pairs of MT datasets from year 2010 to 2019, we construct a classification dataset following the steps above. Besides the source information, we also construct another kind of classification datasets that also consider the information of system outputs. The procedure to construct this kind of dataset is almost similar to the previous one, except that we concatenate each system outputs with its source sentences after the Step-2 is finished. In our experiments, we use mBERT (Devlin et al., 2019; Wolf et al., 2020) as the classifier, thus an unified structure can be used for the two kinds of datasets.