Towards a Better Evaluation of Metrics for Machine Translation

Peter Stanchev **Weiyue Wang** **Hermann Ney**
Human Language Technology and Pattern Recognition, Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany
<surname>@i6.informatik.rwth-aachen.de

Abstract

An important aspect of machine translation is its evaluation, which can be achieved through the use of a variety of metrics. To compare these metrics, the workshop on statistical machine translation annually evaluates metrics based on their correlation with human judgement. Over the years, methods for measuring correlation with humans have changed, but little research has been performed on what the optimal methods for acquiring human scores are and how human correlation can be measured. In this work, the methods for evaluating metrics at both system- and segment-level are analyzed in detail and their shortcomings are pointed out.

1 Introduction

In the past, machine translation (MT) metrics have been extensively studied and evaluated, at both system- and segment-level (Bojar et al., 2016, 2017; Ma et al., 2018, 2019). When performing system-level evaluation, the average score of a MT system is taken into account. Segment-level evaluation uses each sentence (segment) separately to compute correlation. The results of these metric evaluations are critical to the way MT metrics are perceived. In particular the correlation with human judgment is of great importance.

For this reason, an understanding for the workings of the evaluation method is required. Proposals to identify relevant system-level human scores have been discussed (Koehn, 2012; Sakaguchi et al., 2014), but no comprehensive analysis on this topic has been conducted. In particular, detailed studies on the segment-level evaluation are neglected, although it is an integral part of the metric evaluation.

Since the goal of a metric is to evaluate a translation as close as possible to a human's rating, it is important to clearly define the methods of determining human score and the methods of correlation measurement. This work aims to present an overview of the methods used in the evaluation, analyze their strengths and weaknesses, and propose solutions to some of the pitfalls of the methods.

2 Human Scores

To measure the correlation between the score of a metric and the score of a human, a method of determining human scores is required. Thus, a person has to judge the quality of a translated sentence. This is not a simple task, as different people may have different opinions about the exact quality of the translation. Another aspect to consider is that in order to calculate correlation, the score must be quantifiable in some way. Thus, the methods used to detect human judgment must use a sufficient number of human judges for them to be reproducible.

In the Workshop on Statistical Machine Translation (WMT), three different methods are used to determine the human score: direct assessment (DA) (Graham et al., 2017), relative ranking (RR) (Stanojevic et al., 2015) and, in recent years, relative ranking out of direct assessment (DARR) (Bojar et al., 2017).

2.1 Direct Assessment

The DA measures the quality of a translation on a scale from 0 to 100, based on the adequacy and fluency of the sentence. To obtain the score, the human judges are provided with a reference translation and the output of a single MT system, which makes the evaluation process monolingual. To ensure reproducibility, a large number of judges are needed – at least 15 (Ma et al., 2019). Additionally, scores are standardized (Graham et al., 2017) to eliminate individual distortions, such as judges who only provide high or low scores. Furthermore, a form of quality control is applied to filter out

judges who exhibit a high variance in comparison to their peers.

Overall, DA is one of the best ways to obtain human judgement. It provides a numerical score that can be easily used in common statistical methods, such as Spearman's ρ (Spearman, 1987) or Pearson's r (Pearson and Galton, 1895), at both the segment- and the system-level. However, to obtain statistically significant correlation measurements and ensure reproducibility, a high number of human scores are required. For the segment-level it is therefore infeasible to obtain DA scores. This leads to the need to use a completely different method for determining human judgements at the segment-level. Another possibility is to establish a relative ranking of the few obtained DA scores (DARR).

2.2 Relative Ranking

The RR produces, as the name implies, a ranking between multiple translations. In WMT, the judges are presented with five system outputs with the corresponding source and reference sentence, making the evaluation process bilingual. Each judge ranks the five sentences from the best to worst, taking equality (tie) into account. To simplify the evaluation, identical sentences from different systems are collapsed into one.

The resulting relative ranking of five tuples is not as straightforward to use for correlation calculation, since most correlation coefficients rely on absolute ranking information. One approach to obtaining a correlation is to use a variant of Kendall's τ (Kendall, 1938; Macháček and Bojar, 2014). This entails converting the scores produced by metrics into relative rankings. Naturally, this has the disadvantage that the fine granularity of the scores is lost. However, this method can be used for both segment- and system-level correlation calculations.

Another option used in WMT16 (Bojar et al., 2016) is to convert relative rankings to absolute rankings through TrueSkill (Herbrich et al., 2006; Sakaguchi et al., 2014). This method uses the relative rankings to estimate an absolute score for each system, which is then used to calculate the correlation (by Pearson's r or Spearman's ρ). The score of each system is represented by a Gaussian distribution, with the mean of the predicted score of the system and the variance of the confidence in that prediction. Due to the nature of the method, it can only be used for the system-level correlation calculation. This, in turn, makes it difficult to interpret the results since normally two different correlation calculation methods must be used for the different evaluation levels.

2.3 DARR

Due to the difficulty of obtaining enough DA scores for a statistically significant segment-level correlation calculation, Bojar et al. (2017) introduced the concept of obtaining a relative ranking from the DA used at the system-level and termed DARR. For this purpose, all possible sentence pairs, for which a DA score is available, are generated between all participating systems. These sentence pairs are then filtered to remove ties. The criterion used by Ma et al. (2019) is to remove sentence pairs, whose difference on the DA scale is less than 25. This should lead to the removal of all ties and produce an RR that scores the systems only as better or worse. However, this is not the case. Table 1 shows the RR of sentences with a sentence identifier (`SID`) on different language pairs (`LP`). The system that has achieved a better translation according to the DA score for these sentences is under the column `better`. In this case, the sentences generated by both systems are completely identical, as can be seen in Table 2, although they have been classified as different according to the DARR method. Such identical sentences occur across multiple language pairs in the WMT19 data set.

Another important aspect is that tie filtering is not applied to the metrics scores and therefore ties are possible for metrics. This makes the correlation calculation, especially for identical sentences, a difficult task. It is therefore of interest to determine how many identical sentences are present after filtering. For this reason, a brief analysis is carried out on the basis of the WMT19 data using six language pairs, which is shown in Table 3. There are no identical sentences for the language pairs Gujarati→English (`gu-en`) and Kazakh→English (`kk-en`). However, for all other language pairs, especially Chinese→English (`zh-en`), there are identical sentences. Note that these identical sentences are present after the tie filtering. Table 3 also shows the amount of ties produced by two metrics: YiSi-1 (Lo, 2019) and EED (Stanchev et al., 2019). It is clear that a significant amount of the ties for the two metrics come from identical sentences.

In addition, a considerable amount of data is eliminated. Figure 1 depicts the effect of varying the equivalence threshold, i.e. cases, in which the

LP	data	SID	better	worse
de-en	newstest2019	1200	uedin.6749	UCAM.6461
zh-en	newstest2019	604	Baidu-system.6940.zh-en	MSRA.MASS.6996.zh-en
en-zh	newstest2019	1351	NEU.6830	UEDIN.6158

Table 1: RR human scores for segment-level with their corresponding sentences from WMT19[1].

LP	SID	system-generated hypotheses (identical for both systems)
de-en	1200	Music cabaret: Gender understanding with heart - Wolbeck - Westphalian News
zh-en	604	China Resources Beer closed at HK $28.85 on Friday, down nearly 4.5% in the past month.
en-zh	1351	他还希望赋予议会更大的权力来建造新的住房。

Table 2: Corresponding sentences from Table 1 for the two systems.

LP	#sentences	#identical sentences	YiSi-1 #ties	EED #ties
gu-en	31k	0	2	27
kk-en	27k	0	74	115
zh-en	31k	152	336	361
en-gu	11k	5	8	13
en-kk	18k	23	53	64
en-zh	19k	84	205	455

Table 3: Number of identical sentences vs. ties in the WMT19 corpus used for human correlation.

difference in DA scores is below the specified value, are considered as ties. Note that the threshold influences the amount of data used immensely. By having virtually no threshold (a threshold of 1) the average number of sentences is five times higher than when using a threshold of 50. The threshold of 25 used by WMT19 almost halves the amount of data used to acquire the correlation.

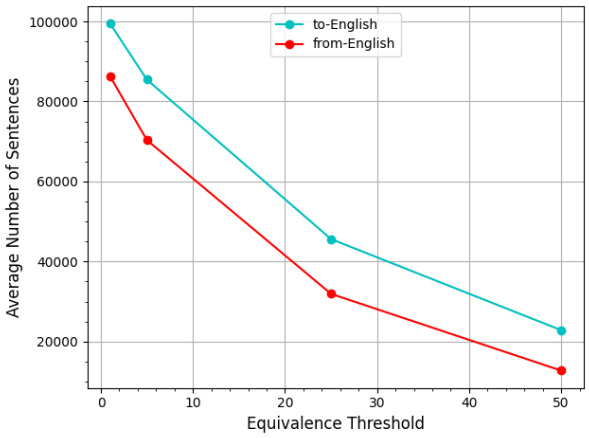

Figure 1: The average number of sentences over different language pairs (`to-English` and `from-English` directions) when excluding ties based on various equivalence thresholds.

Overall DARR provides a method for calculating the correlation at the segment-level in a scenario where there is not enough DA data. However, removing ties as part of the human component makes the evaluation unfair. This is aggravated by the fact that after DA-based tie filtering, not all ties are successfully removed. One possible solution, which remains to be tested, is to consider the ties of the human component carefully. This would at least equal the domain of metrics and human scores.

3 Measuring the Correlation

Obtaining human scores is only part of the correlation calculation. The other one is to use both human and metric scores to compute their similarity or correlation. The case, where both human and metric scores are represented by absolute values, is straightforward to compute using methods such as Pearson's r or Spearman's ρ. However, DA relies on a large amount of annotators that cannot always be guaranteed, especially at the segment-level. In the case where RR or DARR is used for human scores, this task is not that easy. For this reason, the focus here is on the case where a form of RR is used – typically for the segment-level correlation calculation.

As previously mentioned, WMT uses a form of Kendall's τ to obtain a correlation given the relative ranking. The coefficient definition in its most general form is shown in Equation (1)

$$\tau = \frac{|\text{concordant} - \text{discordant}|}{|\text{concordant} + \text{discordant}|}, \qquad (1)$$

where the concordant pairs denote cases in which there is agreement between the metric and the hu-

[1]Scripts and data from:
http://ufallab.ms.mff.cuni.cz/~bojar/wmt19-metrics-task-package.tgz

man score, and the discordant pairs cases in which there is disagreement.

To formally define agreement and disagreement, a matrix can be used as described by Macháček and Bojar (2014). The various matrix formulations that have been used in WMT over the years are shown in Table 4. For metric scores to be interpretable in these matrices, a relative ranking must be constructed from the absolute scores for each participating metric. This is achieved by performing a pairwise comparison of the participating systems at the segment-level, taking into account ties. All three matrices treat matches and mismatches identically:

- discordant pairs are always cases where there are disagreements between the human and metric scores: $\{<, >\}$ or $\{>, <\}$,
- concordant pairs are always cases where the scores match: $\{<, <\}$ or $\{>, >\}$.

The only difference between the three methods is the treatment of ties.

Table 4a ignores the existence of ties. However, this is not desirable since ties are possible for metrics. Therefore, metrics are not evaluated on the same amount of sentences. This can be particularly detrimental to metrics that produce a large number of ties. For example, a metric with 99 ties and 1 concordant pair would achieve perfect correlation, while a metric without ties and 70 concordant and 30 discordant pairs would give a correlation of 0.4. The discrepancy in the results due to the data difference is evident.

On the other hand, incorporating ties while not considering human score ties can also lead to undesirable results. In Table 4c, which is used in WMT19, metric ties are considered as a discordant pair $\{<, =\}$ and $\{>, =\}$. Since ties are not defined (or included) in human scores, every tie produced by a metric results in a discordant pair. This in turn reduces its correlation. Thus, a "perfect" metric would never produce a tie between two sentences. This assumption does not reflect reality. In addition, the matrix is not symmetric since there are more possible discordant pairs than concordant ones. This means that a reasonable interpretation of the negative correlation is not possible. Therefore, metrics that have a negative correlation, such as TER (Snover et al., 2006), CHARACTER (Wang et al., 2016) and EED (Stanchev et al., 2019), must be mapped from an error (or edit) rate (E) to an accuracy score to ensure a relatively fair evaluation. This is not trivial, as there is no standard way to convert these metrics into the accuracy rate: neither $1 - E$ nor $-E$ is optimal.

A middle ground between the penalization and the ignoring of ties is the matrix in Table 4b. The ties are not penalized directly, but affect the overall correlation since they are part of the denominator:

$$\tau = \frac{|\text{concordant} - \text{discordant}|}{|\text{concordant} + \text{discordant} + \text{ties}|} \quad (2)$$

Since there is no hard penalization for metrics that produce more ties, such metrics are at a disadvantage. For example, a metric with 20 ties and 80 concordant pairs would achieve a correlation of $80/(80 + 20) = 0.8$, although all non-tie pairs achieve perfect correlation. On the other hand, a metric that overproduces ties, for example, with 80 ties and 20 concordant pairs, would have a correlation of $20/(20 + 80) = 0.2$. It can also be argued that measuring the correlation on metrics with a too high percentage of ties is not significant, since there are too few sentence pairs that are concordant or discordant.

One possible solution to the problem is shown in Table 5. The cases where there is clear agreement or disagreement between humans and metrics remain unchanged. In cases of tie disagreements, a soft penalization is added. This soft penalization is realized the same manner as in Table 4b using Equation (2). In the case where both the metric and human scores tie the two systems, a concordant pair (1) for accuracy-based metrics and a discordant pair (-1) for error rate-based metrics are given. This allows the process to be symmetrical and avoids the problem of having to map error rate to accuracy or vice versa. In addition, ties can now positively affect the correlation and all metrics are evaluated on the same amount of data. Naturally, this alteration of the evaluation method requires that ties be included in the RR. When using DARR, this can be achieved by considering all pairs, where the DA score difference is less than 25, and where the system translations are identical, as ties. A disadvantage of this method is that a distinction has to be made between metrics that aim for a strong negative correlation and metrics that aim for a strong positive correlation. Moreover, the exact range, where a tie is considered, is not necessarily clear.

		Metric		
		<	=	>
Human	<	1	X	-1
	=	X	X	X
	>	-1	X	1

(a) No tie penalization

		Metric		
		<	=	>
Human	<	1	0	-1
	=	X	X	X
	>	-1	0	1

(b) Soft tie penalization

		Metric		
		<	=	>
Human	<	1	-1	-1
	=	X	X	X
	>	-1	-1	1

(c) Hard tie penalization

Table 4: Kendall's τ evaluation matrices (Machácek and Bojar, 2014; Ma et al., 2019).

		Metric		
		<	=	>
Human	<	1	0	-1
	=	0	{1,-1}	0
	>	-1	0	1

Table 5: Integration of human ties in Kendall's τ.

4 Discussion

The MT metric evaluation is an area that needs further investigation. This work gives an overview of the methods used so far and highlights some of their shortcomings. The system-level assessment currently seems to be good, but the evaluation methods at the segment-level still need to be explored (in particular, if there is not enough DA data to directly calculate the correlation at the segment-level):

- It might not be a good idea to rule out tie cases: in theory, there are identical translations and translations of the same quality, and the metrics should be able to give them the same score; in practice, we have shown that excluding all tie cases eliminated a large proportion of the scores collected, which will have a significant impact on the final results. However, it is difficult to clearly define the tie cases for human evaluations, as in DA, on a scale from 0 to 100, different human annotators can give different scores for identical translations.

- The threshold for tie cases is not well defined. Further studies on the threshold value can be carried out. And also whether a threshold should be applied to the automatic metric scores. This study itself may not be a theoretically well-defined task, but some insight could be gained by examining the performance of various metrics under different thresholds.

- The used correlation coefficient is not symmetrical. Then the metrics with negative correlations have to be preprocessed before the evaluation, which can lead to inconsistencies. The proposed solution may also have potential problems as described, but it is worth doing further studies to define a better correlation coefficient.

In general, the task of creating a metric evaluation that is fair and reproducible for all metric types remains to be solved and deserves more attention and study.

Acknowledgements

This work has received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS") and the Deutsche Forschungsgemeinschaft (DFG; grant agreement NE 572/8-1, project "CoreTec"). The work reflects only the authors' views and none of the funding parties is responsible for any use that may be made of the information it contains.

References

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 489–513, Copenhagen, Denmark.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Milos Stanojevic. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 199–231. The Association for Computer Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation sys-

tems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 569–576. MIT Press.

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

Philipp Koehn. 2012. Simulating human judgment in machine translation evaluation campaigns. In *2012 International Workshop on Spoken Language Translation, IWSLT 2012, Hong Kong, December 6-7, 2012*, pages 179–184. ISCA.

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*, pages 706–712, Florence, Italy. 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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 671–688. 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, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 62–90. Association for Computational Linguistics.

Matous Machácek and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 293–301. The Association for Computer Linguistics.

Karl Pearson and Francis Galton. 1895. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 1–11. The Association for Computer Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

C. Spearman. 1987. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 713–719, Florence, Italy. Association for Computational Linguistics.

Milos Stanojevic, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 256–273. The Association for Computer Linguistics.

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, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 505–510. The Association for Computer Linguistics.