Revisiting Round-Trip Translation for Quality Estimation

Jihyung Moon
Naver Papago

Hyunchang Cho
Naver Papago

Eunjeong L. Park
Naver Papago

{jihyung.moon, hyunchang.cho, lucypark}@navercorp.com

Abstract

Quality estimation (QE) is the task of automatically evaluating the quality of translations without human-translated references. Calculating BLEU between the input sentence and round-trip translation (RTT) was once considered as a metric for QE, however, it was found to be a poor predictor of translation quality. Recently, various pretrained language models have made breakthroughs in NLP tasks by providing semantically meaningful word and sentence embeddings. In this paper, we employ semantic embeddings to RTT-based QE. Our method achieves the highest correlations with human judgments, compared to previous WMT 2019 quality estimation metric task submissions. While backward translation models can be a drawback when using RTT, we observe that with semantic-level metrics, RTT-based QE is robust to the choice of the backward translation system. Additionally, the proposed method shows consistent performance for both SMT and NMT forward translation systems, implying the method does not penalize a certain type of model.

1 Introduction

A good machine translation (MT) system converts one language to another while preserving the meaning of a sentence. Given a pair of well-performing translation systems between two languages, the meaning of a sentence should remain intact even after a round-trip translation (RTT) – the process of translating text from the source to target language (forward translation, FT) and translating the result back into the source language (backward translation, BT). If the MT systems work reasonably well and no human-produced reference translations are provided, using RTT for translation evaluation seems like a natural choice.

Input (en)	'We know it won't change students' behaviour instantly.
Reference (de)	Wir wissen, dass es das Verhalten der Studenten nicht sofort ändern wird.
Output (de)	„Wir wissen, dass es das Verhalten der Schüler nicht sofort ändern wird.
Round-trip (en)	"We know that it will not change student behavior immediately.
RTT-SENTBLEU: 14.99 (rank: 1947/1997)	
RTT-SBERT(*): 98.07 (rank: 1001/1997)	
RTT-BERTSCORE(*): 97.04 (rank: 1033/1997)	

Table 1: A sample of RTT-based evaluation methods with an example from the WMT19 English–German evaluation set. * denotes our proposed semantic-level methods (Detailed definitions are described in Section 3). Note that SENTBLEU could not capture the similarity of the input and RTT.

However, in the early 2000s, this practice was not recommended to be used as a translation evaluation method (Huang, 1990; Somers, 2005; van Zaanen and Zwarts, 2006). This argument was largely supported by the poor correlation between BLEU (Papineni et al., 2002) for reference and translated output and BLEU for input and RTT (We address this method again in Section 3 as RTT-BLEU). However, BLEU only measures surface-level lexical similarity, thus penalizing paraphrased sentences resulting from the round-trip translation as shown in Table 1.

On the other hand, human evaluations con-

ducted on input sentences and translated outputs show a significant positive correlation with human evaluations on input sentences and round-trip sentences (Aiken and Park, 2010). The result implies if a suitable semantic-level metric is provided, RTT-based method can be used for MT evaluation. Meanwhile, recently introduced pre-trained language models e.g., BERT (Devlin et al., 2019) and SBERT (Reimers and Gurevych, 2019), are effective for many natural language processing tasks including semantic similarity detection (Cer et al., 2017). BERTSCORE (Zhang et al., 2019a) and YISI (Lo, 2019) leveraged such models for MT evaluation and confirmed the efficacy.

In this paper, we revisit RTT with recently proposed semantic-level metrics for MT quality estimation. Quality estimation (QE) aims to measure how good a translation is without any human-translated references (Fonseca et al., 2019) as opposed to reference-based metrics such as BLEU or CHRF. Therefore, with these metrics, it is easy to evaluate translations beyond reference-ready domains, e.g., user logs in commercial services.

We start by investigating RTT-based QE metrics on different BT systems to choose a proper BT system to examine RTT-based methods across different language pairs. Then we compare the methods on NMT with statistical machine translation (SMT) systems and demonstrate the compatibility of our methods. Across the experiments, RTT-based QE metrics with semantic-level similarities outperform lexical-based similarity metrics. We find the results are related to the metric's ability of detecting paraphrases.

The main contributions of this work are as follows:

- We reconsider RTT with suitable semantic-level metrics, specifically SBERT and BERTSCORE in our settings, and show it can be used to measure translation quality.
- We observe RTT methods using SBERT and BERTSCORE are robust to the choice of BT systems.
- We present RTT with semantic similarity measurements consistently exhibit high-performance across different FT systems: SMT and NMT.
- We find the paraphrase detection ability of metrics is related to the performance of RTT-based QE.

2 Related Work

2.1 Quality Estimation

One goal of QE is to estimate the quality for machine translated sentences without reference translations, but the definition of quality has gradually changed. Traditional QE aimed to estimate the required amount of post-editing efforts for a given translation in the word, sentence or document level. In the sentence level, this can be understood as estimating the Human Translation Error Rate (HTER) (Snover et al., 2006), or the rate of edit operations which include the insertions of words, deletions or replacements. The recently proposed view of "QE-as-a-metric" (Fonseca et al., 2019)[1] differs from traditional quality estimation in that it directly aims to estimate the absolute score of a translation, and can be directly compared with previous reference-based metrics. While reference-based metrics easily achieve above 0.9 Pearson correlation with direct human assessments in the system-level and up to 0.4 correlation in the sentence-level, QE-based metrics typically score less (Ma et al., 2019).

YISI (Lo, 2019) is the best performing QE metric from the recent QE-as-a-metrics subtask submitted to the quality estimation shared task of WMT19 (Ma et al., 2019)[2]. It takes contextual embeddings extracted from BERT and computes F-scores of semantic phrases using the cosine similarity of words weighting by their inverse document frequency (idf). YISI has variants for both situations where the references exist (YISI-1) or does not exist (YISI-2).

2.2 Round-trip Translation

RTT had frequently been used for a means of evaluating MT systems until Somers (2005) and van Zaanen and Zwarts (2006) claimed that RTT is inappropriate as a QE metric for translations. The idea was supported by the low correlations between a BLEU score for the input and RTT (RTT-BLEU) and a BLEU score for the reference and output. However, BLEU is not an adequate metric to validate RTT for QE. When Aiken and Park (2010) re-assessed RTT with human judgments, there was a significant positive correlation

[1] Since WMT20, this was modified to the "sentence-level direct assessment task".

[2] We excluded UNI and its variants from consideration, since they do not have any open publications to refer to. See Table 2 in (Ma et al., 2019).

between the human scores of round-trip translations and one-way translations.

Recently, RTT has been employed for other purposes: generating paraphrased sentences and modeling purposes. Yu et al. (2010) exploit RTT-based features to estimate the quality of spoken language translation and improve the accuracy of QE model. Mallinson et al. (2017) reassess using RTT for generating paraphrases in the context of NMT. Junczys-Dowmunt and Grundkiewicz (2017) and Lichtarge et al. (2019) generate large amounts of artificial data to train an automatic post editing model and grammatical error correction, respectively. Vaibhav et al. (2019) also uses RTT to augment bilingual data for NMT. Lample et al. (2018) measures RTT-BLEU for model selection purposes and Hassan et al. (2018) uses RTT as a feature to re-rank translation hypotheses.

2.3 Sentence Similarity Methods

Lexical metrics, such as BLEU (Papineni et al., 2002) and CHRF (Popović, 2015), have long and widely been used for translation evaluation. Both metrics compute strict matching between translation output and reference at the surface level. BLEU counts the n-gram matches of the output and reference over the number of tokens of output as well as the length similarity of the output and reference. CHRF computes F-score based on character-level n-grams. However, they cannot capture the semantic similarity of output and reference sentences beyond lexical relatedness or overlap. In this sense, lexical-based metrics may not be the best way to measure the similarity of paraphrases.

BERT (Devlin et al., 2019), a pre-trained language representation model, made breakthroughs on many natural language processing tasks, including the sentence similarity prediction task (Cer et al., 2017). The methods using BERT's embedding vectors were also introduced to MT evaluation, the task that needs semantic-level similarity measurement, and show the best performance (Ma et al., 2019; Lo, 2019; Zhang et al., 2019a). BERTSCORE (Zhang et al., 2019a) leverages BERT wordpiece embeddings to compute sentence similarity of two monolingual sentences. When BERTSCORE is applied to the output and reference, it outperforms BLEU and CHRF. Meanwhile, SENTENCE-BERT (SBERT) (Reimers and Gurevych, 2019), a fine-tuned BERT, is introduced to derive more semantically meaningful sentence-level representation than BERT. From the encouraging results of the embedding-based methods, we would expect the embeddings to catch the semantic similarity of input and round-trip sentences.

3 RTT-based QE Metrics

Given an input sentence $x = (x_1, x_2, x_3, ..., x_n)$ and a round-trip sentence $\hat{x} = (\hat{x}_1, \hat{x}_2, \hat{x}_3, ..., \hat{x}_m)$, an RTT-based QE metric f is a scalar function computing the similarity of x and $\hat{x}$. We consider the scalar output as a quality for the translation of x. The validity of f is assessed primarily by Pearson correlation against the human judgments.

Previously, only surface-level similarity metrics were used for f. In this paper, we propose to use semantic-level metrics which can capture higher-level concepts of the similarity. Detailed implementations are described in the Appendix A.

3.1 Surface-level Metrics

RTT-BLEU / RTT-SENTBLEU BLEU (Papineni et al., 2002) has originally designed to measure system-level translation performance. To evaluate a sentence-level translation, SENTBLEU, the smoothed version of BLEU, has been used (Ma et al., 2019; Ma et al., 2018). Since system-level BLEU and sentence-level BLEU exploit different computation method, we also separate BLEU-based RTT QE metric for the system-level and sentence-level. Specifically, RTT-BLEU is either BLEU or SACREBLEU-BLEU (Post, 2018) on system-level input sentences and round-trip sentences while RTT-SENTBLEU is SENTBLEU on a single input sentence and round-trip sentence.

RTT-CHRF Sentence-level score is produced by CHRF and system-level score is the average of the segment score obtained by SACREBLEU-CHRF[3] (Post, 2018).

3.2 Semantic-level Metrics

In our settings, the semantic-level metrics are represented by the cosine similarity of SBERT embeddings and BERTSCORE. For all metrics, system-level score is an averaged sentence-level scores.

RTT-SBERT RTT-SBERT calculates the cosine similarity of x and $\hat{x}$ embedding vectors extracted from SBERT (Reimers and Gurevych, 2019). We

[3]It is widely known that their scores are slightly different from the average of CHRF even with the same parameters. Since SACREBLEU is standard, we take SACREBLEU-CHRF for the system-level score.

use a publicly available pre-trained SBERT[4]. Note that released models support Arabic, Chinese, Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, and Turkish.

RTT-BERTSCORE RTT-BERTSCORE computes F-score based on wordpiece-level embedding similarities of x and $\hat{x}$ weighted by inverse document frequency (idf), where each embedding is taken from BERT. The idf weights penalize common wordpiece similarities, such as end of sentence symbols. Given L input sentences $\{x^k\}_{k=1}^{L}$, the idf score of x_i is defined as:

$$\text{idf}(x_i) = -\log \frac{1}{L} \sum_{i=1}^{L} \mathbb{1}[x_i \in x^k]$$

4 Experimental Settings

We compare RTT-based semantic-level QE metrics to lexical-level QE metrics in various conditions. Initially, we prepare different BT systems to see the impact of the BT system to the performance change in RTT-based metrics. Then, with a suitable BT system, we observe the proposed metrics on WMT 2019 metrics task evaluation dataset. We also examine whether our methods are biased to the certain type of FT system. Furthermore, we investigate relations of the performance of RTT-based QE metrics and their paraphrase detection ability.

4.1 Data

WMT metrics task evaluation set The WMT19 dataset includes translations from English to Czech, German, Finnish, Gujarati, Kazakh, Lithuanian, Russian, and Chinese, and from the same set except Czech to English. Translation outputs were provided by the WMT19 submitted systems where all were NMT. Each system was not necessarily present in all language pairs, therefore, English–German received 22 submissions whereas German–English received 16 (see n in Table 3). The human scores were gathered by using Direct Assessment (DA) for the translations of all systems on a scale of 0-100 points then standardized for each annotator. System's performance is an average over all assessed sentences produced by the given system and sentence-level golden truth is a relative ranking of DA judgments (DARR). In WMT19, QE-as-a-metrics were also assessed by the same standard as the reference-based metrics, namely Pearson correlation coefficient and Kendall's τ-like formulation against DARR, therefore, performance could be compared directly with BLEU (Papineni et al., 2002) and CHRF (Popović, 2015).

We also use WMT12 metrics task evaluation set to assess RTT-based metrics on SMT. It includes translations from English to Czech, French, German, and Spanish and vice versa. We select English–German, German–English, and English–Czech which also appeared in WMT 2019. The annotators were asked to evaluate sentences by ranking translated outputs from randomly selected 5 systems. A ratio of wins is used for the system's performance (Callison-Burch et al., 2012).

PAWS PAWS (Paraphrase Adversaries from Word Scrambling) (Zhang et al., 2019b) is a paraphrase identification dataset constructed from sentences in Wikipedia (Wiki) and Quora Question Pairs (QQP) corpus. We denote dataset as $\text{PAWS}_{\text{Wiki}}$ and PAWS_{QQP} respectively.

Paraphrased sentences are generated by controlled word swapping and back translation, followed by fluency and paraphrase judgments by human raters. Paraphrase and non-paraphrase pairs are mixed and to make dataset more challenging, both pairs have high lexical overlap.

4.2 Backward Translation (BT)

To estimate the quality of MT systems with RTT, a BT system is required. The choice of the BT system seemingly has the potential to largely affect the performance of RTT-based QE metrics, so we run experiments to verify the effect of BT system qualities. We compare two types of models–the system trained solely on WMT19 news translation task training corpus and online system–with different performance in terms of BLEU. The BT systems trained on the WMT19 news dataset could be considered adequate to evaluate the WMT19 submitted FT systems since both systems are trained on the same domain. On the other hand, online systems could also be desirable, because the online systems are trained on a huge amount of corpus mixed with various domains and would outperform the trained models on WMT19 dataset. If the online systems show more favorable results, then RTT-based QE metrics can be more practical in terms of the easy access to a BT system on any

[4] https://github.com/UKPLab/sentence-transformers

Backward translations	Pearson correlations					Variance ($\times 10^{-4}$)	
Systems	BLEU	RTT-BLEU	RTT-CHRF	RTT-SBERT	RTT-BERTSCORE	RTT-SBERT	RTT-BERTSCORE
Google	**46.96**	0.797	0.853	0.941	0.951	5.08	1.96
Microsoft	42.68	**0.845**	**0.877**	**0.948**	0.955	5.12	2.07
Amazon	40.89	0.776	0.804	0.941	**0.956**	4.86	1.88
Facebook-FAIR	42.17	0.788	0.865	0.940	0.934	4.84	1.27
Transformer Big (100k)	38.96	0.739	0.818	0.939	0.937	4.58	1.57
Transformer Big (40k)	36.38	0.707	0.795	0.938	0.935	4.22	1.36
Transformer Big (20k)	34.75	0.617	0.759	0.931	0.860	3.97	1.15
Transformer Big (10k)	31.30	0.509	0.749	0.908	0.789	3.17	0.91

Table 2: Performance of RTT-based QE metrics on 22 English–German FT systems with various German–English BT systems. The variance of the best metrics, RTT-SBERT and RTT-BERTSCORE, are described, additionally.

language pair.

To examine the impact of the BT systems, we choose English–German, which is the most submitted language pair. For trained BT systems, we use Facebook-FAIR[5], the best system in WMT19 on German–English, and the Transformer Big model (Vaswani et al., 2017) saved at 10k, 20k, 40k, and 100k iterations during training on the WMT19 corpus. Details of the Transformers are described in Appendix B. We also try three online systems, namely Google, Microsoft, and Amazon, showing different BLEU on WMT19 German–English evaluation set. Each system was requested on Oct 2019, Nov 2019, and Dec 2019.

4.3 Forward Translation (FT)

The metric might penalize or favor a certain type of models. For instance, BLEU has been argued to penalize rule-based systems against statistical systems (Hovy, 2007).

To investigate whether RTT-based QE metrics penalize FT systems based on their architecture, we assess RTT-based QE metrics on both NMT and SMT. As the all models submitted to WMT19 are NMT (Ma et al., 2019), and the models submitted to WMT12 are SMT or rule-based model (Callison-Burch et al., 2012), we denote the former as NMT and the latter as SMT. We compare RTT-based QE metrics' performance with Pearson correlation coefficient for the language pairs both appeared on WMT19 and WMT12, English–Czech, English–German, and German–English.

5 Results

5.1 Sensitivity to Backward Translation

Due to the nature of RTT-based QE metrics, a BT system is needed. We use a variety of BT systems in terms of the training recipe and BLEU on WMT19 German–English testset and observe the performance of RTT-based QE metrics evaluated by Pearson correlation (r) with human scores (Ma et al., 2019; Ma et al., 2018). Note that well-performing metrics achieve high correlation coefficient.

According to Table 2, RTT-BERTSCORE and RTT-SBERT not only outperform the other metrics but are also robust to the type and performance of the BT systems. On the other hand, RTT-BLEU and RTT-CHRF are sensitive to the performance of the BT systems, and the correlations fall behind RTT-BERTSCORE and RTT-SBERT. Since BT systems scoring low BLEU have less chance of having same word orders in RTT as with input sentences, the performance of surface-form metrics, RTT-BLEU and RTT-CHRF, decrease more sharply than RTT-SBERT and RTT-BERTSCORE.

The best correlation of each metric is accomplished when the online system is used for the BT system. Even though Microsoft and Facebook-FAIR exhibit a similar BLEU score, metrics are more successful when using the Microsoft system. This can be explained by a variance of RTT-based QE metrics score. In average, the variance of RTT-SBERT and RTT-BERTSCORE using the online BT systems is higher than that of trained ones. The trained BT systems might over-translate a fault translation output similar to the original input, e.g., Kim Jong Un – Kim – Kim Jong Un, that make QE metrics hard to distinguish good systems to the bad ones.

Surprisingly, the best BT system in terms of BLEU does not always guarantee the best RTT-based QE metrics. Despite Google's highest BLEU score, the performance of RTT-based QE metrics is lower than or similar to that of Microsoft. This assures that BLEU is not the only feature that affect

[5]Submitted model is publicly available via PyTorch (`https://pytorch.org/hub/pytorch_fairseq_translation`).

src lang tgt lang	de en	fi en	gu en	kk en	lt en	ru en	zh en	avg. (std.)	en cs	en de	en fi	en gu	en kk	en lt	en ru	en zh	avg. (std.)
n	16	12	11	11	11	14	15		11	22	12	11	11	12	12	12	
BLEU*	.849	.982	.834	.946	.961	.879	.899	.907 (.057)	.897	.921	.969	.737	.852	.989	.986	.901	.907 (.084)
CHRF*	.917	.992	.955	.978	.940	.945	.956	.955 (.025)	.990	.979	.986	.841	.972	.981	.943	.880	.947 (.056)
SACREBLEU-BLEU*	.813	.985	.834	.946	.955	.873	.903	.901 (.065)	.994	.969	.966	.736	.852	.986	.977	.801	.910 (.100)
SACREBLEU-CHRF*	.910	.990	.952	.969	.935	.919	.955	.947 (.028)	.983	.976	.980	.841	.967	.966	.985	.796	.937 (.074)
QE as a Metric																	
Individual Best*	.850	.930	.566	.324	.487	.808	.947	- (-)	.871	.936	.907	.314	.339	.810	.919	.118	- (-)
YiSi-2*	**.796**	.642	.566	.324	.442	.339	**.940**	.578 (.232)	.324	.924	.696	.314	.339	.055	.766	.097	.439 (.319)
RTT-BLEU	.130	**.827**	.641	.859	.596	.295	.825	.596 (.284)	-.625	.797	.417	.608	.930	-.334	.572	-.599	.221 (.637)
RTT-CHRF	.495	.810	**.778**	.776	.692	.524	.875	.707 (.146)	-.408	.842	.487	.586	.423	-.153	.750	-.310	.277 (.493)
RTT-SBERT	.761	-	-	-	-	**.867**	.889	.839 (.005)	.470	.941	.804	.710	.950	**.410**	.833	**.256**	**.672** (.261)
RTT-BERTSCORE	.654	.819	.729	**.889**	**.712**	.816	.912	**.790** (.095)	**.473**	**.951**	**.819**	**.737**	**.966**	.342	**.869**	.071	.654 (.324)

Table 3: Pearson correlations of system-level metrics with human judgments on WMT19. The best correlations of QE-as-a-metric within the same language pair are highlighted in bold. * denotes that reported correlations are from WMT19 metrics task (Ma et al., 2019).

src lang tgt lang	de en	fi en	gu en	kk en	lt en	ru en	zh en	avg. (std.)	en cs	en de	en fi	en gu	en kk	en lt	en ru	en zh	avg. (std.)
n	85k	38k	31k	27k	22k	46k	31k		27k	100k	32k	11k	18k	17k	24k	19k	
SENTBLEU*	.056	.233	.188	.377	.262	.125	.323	.223 (.111)	.367	.248	.396	.465	.392	.334	.469	.270	.368 (.081)
CHRF*	.122	.286	.256	.389	.301	.180	.371	.272 (.096)	.455	.326	.514	.534	.479	.446	.539	.301	.449 (.091)
QE as a Metric																	
Individual Best*	.022	.211	-.001	.096	.075	.089	.253	- (-)	.069	.236	.351	.147	.187	.003	.226	.044	- (-)
YiSi-2*	**.068**	.126	-.001	.096	.075	**.053**	.253	.096 (.080)	**.069**	**.212**	.239	.147	.187	.003	-.155	.044	.093 (.131)
RTT-SENTBLEU	-.169	.095	.111	.140	.086	-.104	.168	.047 (.130)	-.122	-.001	.088	.374	.399	-.110	.157	-.106	.085 (.211)
RTT-CHRF	-.114	.141	**.184**	.130	.099	-.050	.195	.083 (.119)	-.093	.055	.119	.395	.310	-.069	.195	-.075	.105 (.185)
RTT-SBERT	-.066	-	-	-	-	-.013	.225	.049 (.024)	.025	.169	.268	.444	.503	**.070**	.371	**.064**	.239 (.185)
RTT-BERTSCORE	-.085	**.185**	.167	**.204**	**.118**	-.020	**.255**	**.118** (.125)	.065	.194	**.292**	**.494**	**.579**	.069	**.391**	.056	**.268** (.205)

Table 4: Kendall's τ formulation of segment-level metric scores with human judgments on WMT19. The best correlations of QE-as-a-metric within the same language pair are highlighted in bold. For some language pairs, QE metrics obtain negative correlations. * denotes that reported correlations are from WMT19 metrics task (Ma et al., 2019).

the performance of the RTT-based QE metrics.

5.2 Performance across Language Pairs

Provided from the results in Section 5.1, we use one of the online systems to get RTT for all language pairs in WMT19. Specifically, we use Google Translate, because of its coverage of supported language pairs and its overall performance across all language pairs.

We conduct the same experiments as in the WMT19 metrics shared task to directly compare with the previous QE-as-a-metrics. Individual best results of previous methods and YISI-2 are provided to compare RTT-based QE metrics within the same reference-free metrics. Note that YISI was the only QE-as-a-metric scoring on all language pairs, at the same time, achieving the best performance in total (Ma et al., 2019). We also include commonly used reference requiring metrics, namely BLEU, CHRF, SACREBLEU-BLEU, and SACREBLEU-CHRF, to see how far QE metrics can get without reference translation. The metrics are evaluated in system-level and sentence-level for all language pairs. Specifically, Pearson correlation is applied to assess system-level metrics and Kendall's τ-like formulation against the DARR to measure sentence-level metrics.

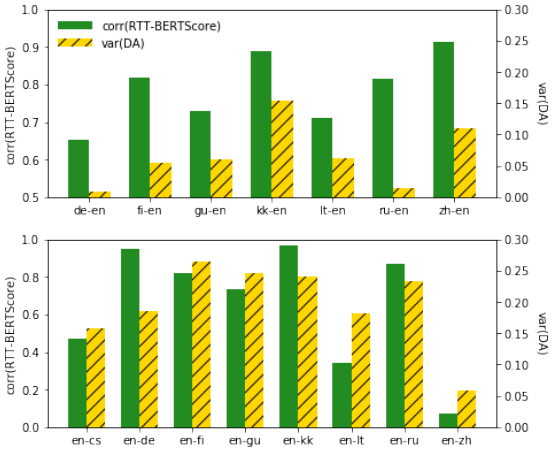


Figure 1: System-level correlations of RTT-BERTSCORE and variance of DA scores.

Table 3 illustrates the system-level correlations with human judgments on both to-English and out-of-English language pairs. Across all language

Language Pairs	Systems (n)	Pearson correlations				
		BLEU	RTT-BLEU	RTT-CHRF	RTT-SBERT	RTT-BERTSCORE
English–Czech	SMT (12)	0.615	0.261	0.342	0.482	**0.620**
	NMT (11)	0.897	-0.625	-0.408	0.470	**0.473**
English–German	SMT (12)	0.582	0.523	0.553	0.742	**0.765**
	NMT (22)	0.921	0.797	0.842	0.941	**0.951**
German–English	SMT (13)	0.841	0.530	0.374	**0.712**	0.682
	NMT (16)	0.849	0.130	0.495	**0.761**	0.654

Table 5: Pearson correlations of BLEU and RTT-based QE metrics where FT systems are SMT and NMT. We reveal the number of systems in parenthesis.

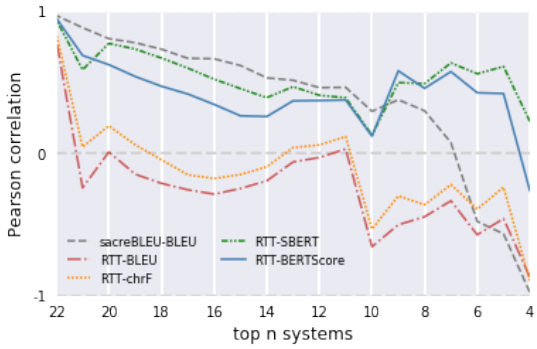


Figure 2: Pearson correlations of RTT-based QE metrics and SACREBLEU-BLEU for English–German system-level evaluation for all systems (left) down to top 4 systems (right).

Metrics	$\text{PAWS}_{\text{WiKi}}$	PAWS_{QQP}
SENTBLEU	0.639	0.354
CHRF	0.584	0.405
SBERT	0.656	**0.545**
BERTSCORE	**0.718**	0.509

Table 6: AUC scores of precision-recall curves of BERT-based metrics on $\text{PAWS}_{\text{WiKi}}$ and PAWS_{QQP} testing set.

pairs except German–English, BERT-based RTT QE metrics outperform RTT-BLEU and RTT-CHRF. For some language pairs, Gujarati–English, Kazakh–English, Russian–English, English–German, English–Gujarati, and English–Kazakh, RTT-BERT-based metrics show comparable result to BLEU, however, QE metrics still fall behind the reference-based metrics on average. Surprisingly enough, the high Pearson correlation coefficients are mostly achieved on low-resource language pairs. Results in Figure 1 suggest that this might due to the high variance of the system's DA scores which implies distinguishing good systems to the bad ones is relatively easy.

To present a more reliable view, we draw plots of Pearson correlation while reducing MT systems to top n ones as in Ma et al. (2019). Figure 2 depicts English–German, and all language pairs are in Appendix C. In general, correlations of SACREBLEU-BLEU and RTT-based QE metrics tend towards 0 or negative, whereas the reference-based metric shows a rather continuous degradation than RTT-based metrics. RTT-SBERT and RTT-BERTSCORE are better at retaining positive correlations compared to RTT-BLEU and RTT-CHRF, however, their consistency is weaker than SACREBLEU-BLEU except for some language pairs (English–German, English–Gujarati, English–Kazakh, and Finnish–English).

Metrics performance on sentence-level is described in Table 4[6]. Sentence-level quality estimation is considered as a more difficult task than that of system-level. This is supported by the poor correlation coefficients of even SENTBLEU and CHRF. Similar to the system-level results, QE metrics fall short of the reference-based metrics. For language pairs with high DA score variance, again, RTT-BERT-based metrics provide comparable performance with reference-based metrics.

5.3 Sensitivity to Forward Translation

A certain type of FT system could be penalized by one metric according to its computation method. For this reason, we observe the performance of RTT-based QE metrics on different FT systems: SMT and NMT. SMT denotes the systems submitted to WMT12 and NMT represents the systems submitted to WMT19. Same as the Section 5.2, we use Google Translate for the BT system and evaluate the metrics with Pearson correlation coefficient. Results are shown in Table 5.

RTT-SBERT and RTT-BERTSCORE demonstrate the most promising performance regardless of the FT systems. In contrast, RTT-BLEU and RTT-CHRF seem to favor SMT. The correlation

[6]Instead of BLEU, we report SENTBLEU.

Case	Sentences	Label	Ranks (out of 677)	
(a)	**sentence 1**: What are some example of *deep web and dark web* ? **sentence 2**: What are some example of *dark web and deep web* ?	1	SENTBLEU: 534 CHRF: 416	SBERT: 242 BERTSCORE: 101
(b)	**sentence 1**: What was the CD that Deanna and family *were* listening to at the beginning of Try (S5E15) of Season 5 of the Walking Dead and why was they listening to it ? **sentence 2**: What was the CD that Deanna and family *was* listening to at the beginning of Try (S5E15) of Season 5 of the Walking Dead and why were they listening to it ?	1	SENTBLEU: 100 CHRF: 142	SBERT: 14 BERTSCORE: 2
(c)	**sentence 1**: How is dark/vacuum energy created with the universe conserved if it is not *created* ? Can infinite of these be *conserved* ? **sentence 2**: How is dark/vacuum energy created with the universe conserved if it is not *conserved* ? Can infinite of these be *created* ?	0	SENTBLEU: 119 CHRF: 90	SBERT: 353 BERTSCORE: 501

Table 7: Example sentences on PAWS_{QQP} dataset. Label 1 indicates paraphrased, and 0 represents dissimilarity. The higher the metric rank, the more similar the two sentences are.

Case	Sentences	Label	Ranks (out of 8000)	
(d)	**sentence 1**: Other famous spa towns include Sandanski , Hisarya , *Kyustendil , Devin , Bankya* , Varshets , and Velingard . **sentence 2**: Other famous spa towns include Sandanski , Hisarya , *Bankya , Devin , Kyustendil* , Varshets and Velingard .	1	SENTBLEU: 2510 CHRF: 1521	SBERT: 884 BERTSCORE: 533
(e)	**sentence 1**: *Southport Tower* is the first new tower to be built at the *southern* end of *the Macleod Trail* in almost 20 years . **sentence 2**: *Macleod Trail* is the first new tower to be built at the *south* end of *Southport Tower* in almost 20 years .	0	SENTBLEU: 2942 CHRF: 1446	SBERT: 4374 BERTSCORE: 7505

Table 8: Example sentences on $\text{PAWS}_{\text{WiKi}}$ dataset. Label 1 indicates paraphrased, and 0 represents dissimilarity. The higher the metric ranks, the more similar the two sentences are.

coefficient gap between BLEU and both RTT-BLEU and RTT-CHRF is smaller when FT system is SMT.

5.4 Paraphrase Detection

The results from all the previous sections consistently show the outstanding performance of RTT-SBERT and RTT-BERTSCORE. We see this in a view of paraphrase detection ability of SBERT and BERTSCORE. To confirm our assumption, we compare the area-under-curve (AUC) scores of precision-recall curves of the four metrics used to measure input and RTT on PAWS dataset. The higher the score is, the better the metric at paraphrase detection. Table 6 depicts the results. Note that SBERT indicates the cosine similarity of the embedding vectors of two sentence pairs extracted from the model.

As expected, BERTSCORE and SBERT outperform SENTBLEU and CHRF. In case (a) of Table 7 and case (d) of Table 8, we can find SENTBLEU and CHRF are sensitive to the change of word order. Additionally, they are hard to distinguish paraphrases on long sentences. From case (b), (c), (d), and (e), lexical-based metrics constantly view the sentences are not paraphrased.

The results imply that metrics capability to measure the semantic similarity is highly correlated to the performance of RTT-based QE metrics.

6 Conclusions

We have presented round-trip translation for translation quality estimation. It can be used for QE with suitable semantic-level similarity metrics like SBERT (Reimers and Gurevych, 2019) and BERTSCORE (Zhang et al., 2019a). RTT-SBERT and RTT-BERTSCORE are robust to the choice of a BT system, which alleviates the disadvantages of RTT being dependent on the BT system. Moreover, both QE metrics significantly outperform the state-of-the-art QE metric, YISI-2. When the performance gap between the FT systems is large, RTT-SBERT and RTT-BERTSCORE provide comparable performance to BLEU. They also perform well irrespective of the modeling architecture of FT systems. In future work, it would be interesting to investigate when RTT-based metrics become more reliable or unreliable.

We find the high performance of RTT-SBERT and RTT-BERTSCORE is owing to SBERT and BERTSCORE's ability to detect paraphrased sentences. If better sentence similarity measurements appear, the performance of RTT-based metrics would increase as well. With the growing amount

of the data and the advance of computing power, there certainly be a better measurement, thus RTT-based QE metric is also promising.

References

Aiken, Milam and Park, Mina. 2010. The efficacy of round-trip translation for MT evaluation. *Translation Journal* Volume 14 1–10.

Cer, Daniel and Diab, Mona and Agirre, Eneko and Lopez-Gazpio, Inigo and Specia, Lucia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation* (SemEval-2017) 1–14.

Callison-Burch, Chris and Koehn, Philipp and Monz, Christof and Post, Matt and Soricut, Radu and Specia, Lucia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the Seventh Workshop on Statistical Machine Translation* 10–51.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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) 4171–4186.

Federmann, Christian and Elachqar, Oussama and Quirk, Chris. 2019. Multilingual Whispers: Generating Paraphrases with Translation. *Proceedings of the 5th Workshop on Noisy User-generated Text* (W-NUT 2019) 17–26.

Fonseca, Erick and Yankovskaya, Lisa and Martins, André FT and Fishel, Mark and Federmann, Christian. 2019. Findings of the WMT 2019 Shared Tasks on Quality Estimation. *Proceedings of the Fourth Conference on Machine Translation* (Volume 3: Shared Task Papers, Day 2) 1–10.

Hassan, Hany and Aue, Anthony and Chen, Chang and Chowdhary, Vishal and Clark, Jonathan and Federmann, Christian and Huang, Xuedong and Junczys-Dowmunt, Marcin and Lewis, William and Li, Mu and others 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*

Hovy, Edward. 2007. Investigating why BLEU penalizes non-statistical systems. *Proceedings of the eleventh MT Summit*

Huang, Xiuming. 1990. A machine translation system for the target language inexpert. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*

Junczys-Dowmunt, Marcin and Grundkiewicz, Roman. 2017. An Exploration of Neural Sequence-to-Sequence Architectures for Automatic Post-Editing. *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) 120–129.

Lample, Guillaume and Ott, Myle and Conneau, Alexis and Denoyer, Ludovic and Ranzato, Marc'Aurelio 2018. Phrase-based & neural unsupervised machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 5039–5049.

Lichtarge, Jared and Alberti, Chris and Kumar, Shankar and Shazeer, Noam and Parmar, Niki and Tong, Simon. 2019. Corpora Generation for Grammatical Error Correction. *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) 3291–3301.

Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin. 2019. Roberta: A robustly optimized bert pre-training approach. *arXiv preprint arXiv:1907.11692*

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

Ma, Qingsong and Wei, Johnny and Bojar, Ondřej and Graham, Yvette. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. *Proceedings of the Fourth Conference on Machine Translation* (Volume 2: Shared Task Papers, Day 1) 62–90

Ma, Qingsong and Bojar, Ondřej and Graham, Yvette. 2018. Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance. *Proceedings of the third conference on machine translation: shared task papers* 671–688.

Mallinson, Jonathan and Sennrich, Rico and Lapata, Mirella. 2017. Paraphrasing revisited with neural machine translation. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Volume 1, Long Papers) 881–893.

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics* 311–318

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. *Proceedings of the*

Tenth Workshop on Statistical Machine Translation 392–395

Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. *WMT 2018* 186

Reimers, Nils and Gurevych, Iryna. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP) 3973–3983.

Somers, Harold. 2005. Round-trip translation: What is it good for? *Proceedings of the Australasian Language Technology Workshop* 127–133.

Snover, Matthew and Dorr, Bonnie and Schwartz, Richard and Micciulla, Linnea and Makhoul, John 2006. A study of translation edit rate with targeted human annotation. *Proceedings of association for machine translation in the Americas* Volume 200 6.

Vaibhav, Vaibhav and Singh, Sumeet and Stewart, Craig and Neubig, Graham 2019. Improving robustness of machine translation with synthetic noise. *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) 1916–1920.

van Zaanen, Menno and Zwarts, Simon. 2006. Unsupervised measurement of translation quality using multi-engine, bi-directional translation. In *Proceedings of the 19th Australian joint conference on Artificial Intelligence: advances in Artificial Intelligence* 1208–1214.

Vaswani, Ashish and Shazeer, Noam and Parmar Niki and Uszkoreit, Jakob and Jones, Llion and N. Gomez, Aidan and Kaiser, Łukasz and Polosukhin, Illia. 2017. Attention is all you need. *Advances in neural information processing systems* 5998–6008.

Yu, Dong and Wei, Wei and Jia, Lei and Xu, Bo. 2010. Confidence estimation for spoken language translation based on Round Trip Translation. In *7th International Symposium on Chinese Spoken Language Processing* 426–429.

Zhang, Tianyi and Kishore, Varsha and Wu, Felix and Weinberger, Kilian Q and Artzi, Yoav. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*

Zhang, Yuan and Baldridge, Jason and He, Luheng. 2019b. PAWS: Paraphrase Adversaries from Word Scrambling. *Proc. of NAACL*

Appendix A Metrics Implementation

RTT-BLEU / RTT-SENTBLEU SACREBLEU-BLEU (Post, 2018) is used for system-level score and SENTBLEU for sentence-level which is a smoothed version of BLEU. Following WMT19 metrics task (Ma et al., 2019), we ran SACREBLEU-BLEU[7] with `BLEU+case.lc+lang.de-en+numrefs.1+smooth.exp+tok.intl+version.1.3.6` and SENTBLEU with `sentence-Bleu` in the Moses toolkit[8]. Since Chinese tokenization is not supported by the `tok.intl` included in the package, we preprocess Chinese sentences with `tokenizeChinese.py`[9].

RTT-CHRF We also take the same computation procedure with WMT19 SACREBLEU-CHRF[7] and CHRF[10]. We ran `chrF3+case.mixed+lang.de-en+numchars.6+numrefs.1+space.False+tok.13a+version.1.3.6.` and python script `chrF++.py` with the parameters `-nw 0 -b 3` respectively.

RTT-SBERT We use `bert-large-nli-mean-tokens` for English and `distiluse-base-multilingual-cased` for the others.

RTT-BERTSCORE BERTSCORE is publicly available[11] and it uses the same model and layer applied in RTT-BERT.

Appendix B German–English Transformer big model configurations

Hyperparameters of German–English transformer model used in Section 4.2 generally followed transformer big configuration of Vaswani et al. (2017), except for three shared embedding matrices of encoder input, decoder input, and decoder output. In other words, we set the matrices' variables independently.

For training data, we used all downloadable parallel corpus on WMT19 news translation task for German–English: Europarl, ParaCrawl, CommonCrawl corpus, News Commentary, Wiki titles, and Rapid corpus of EU press releases. Then, we cleaned corpora by filtering sentence pairs whose token length ratio is bigger than 1.5 or less than 0.66 and left 37,066,883 parallel lines.

We normalized corpora with `normalize-punctuation.perl` in the Moses toolkit[12] and tokenized them using bype-pair encoding implemented in Google's SentencePiece[13]. Encoding models for German and English are separately trained with vocabulary size 32K.

Finally, we trained model with mini batch containing approximately 35K tokens of English and 35K of German for each iteration.

[7] https://github.com/mjpost/sacreBLEU

[8] https://github.com/moses-smt/mosesdecoder/tree/master/mert/sentence-bleu.cpp

[9] http://hdl.handle.net/11346/WMT17-TVXH

[10] https://github.com/m-popovic/chrF/chrF++.py

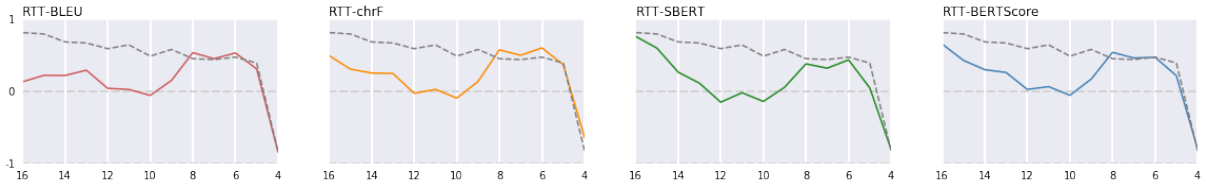
[11] https://github.com/Tiiiger/bert_score

[12] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/normalize-punctuation.perl

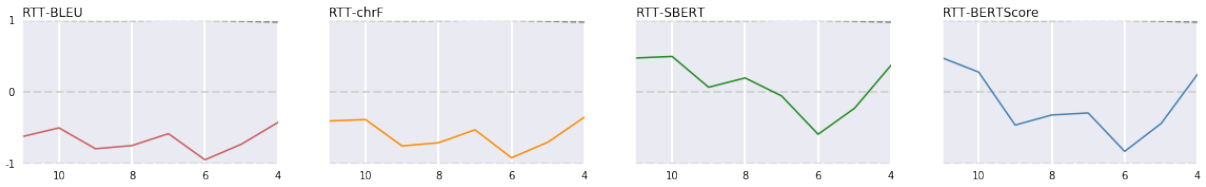
[13] https://github.com/google/sentencepiece

Appendix C Correlations for Top-N Systems

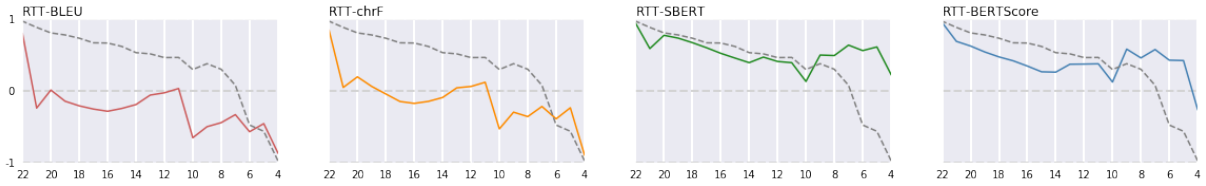
C.1 de-en



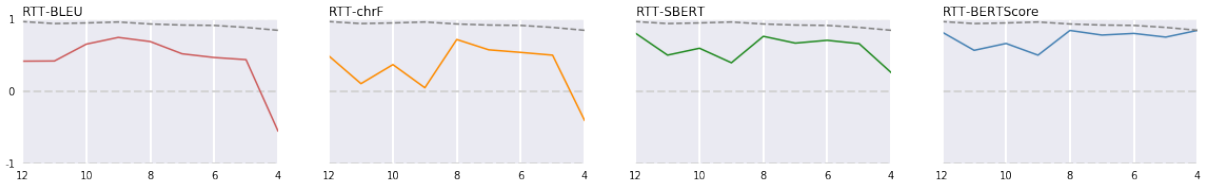
C.2 en-cs



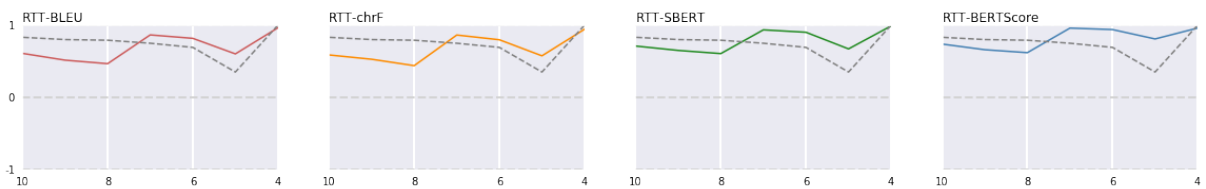
C.3 en-de



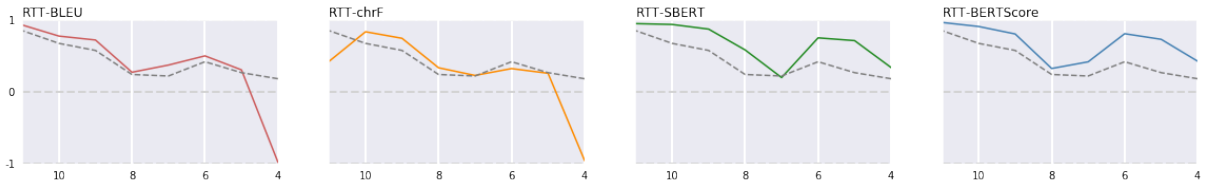
C.4 en-fi



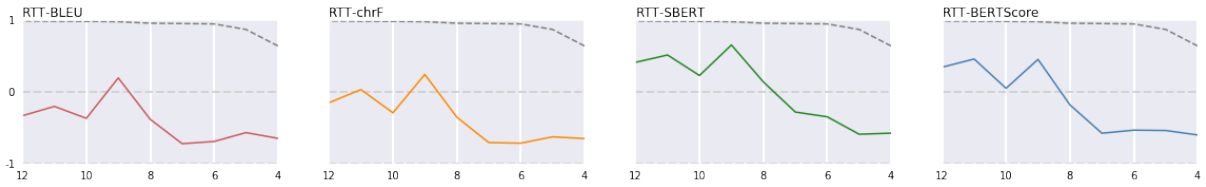
C.5 en-gu



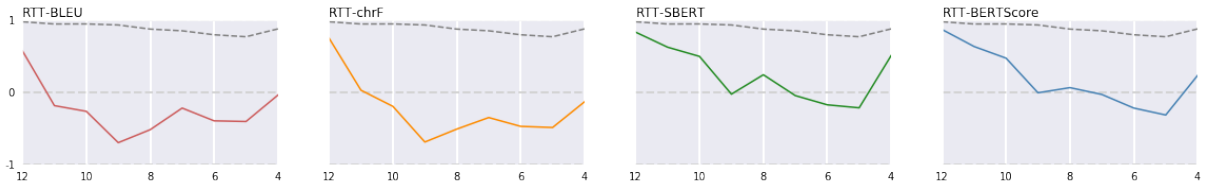
C.6 en-kk



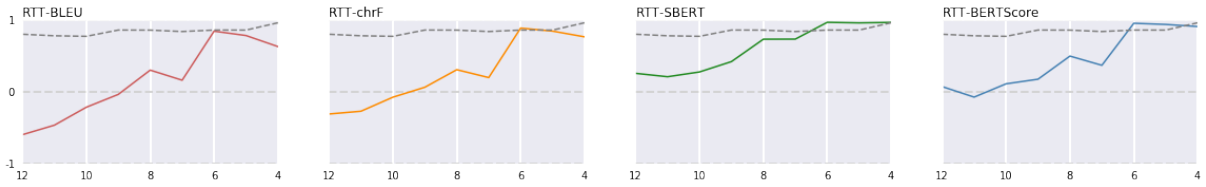
C.7 en-lt



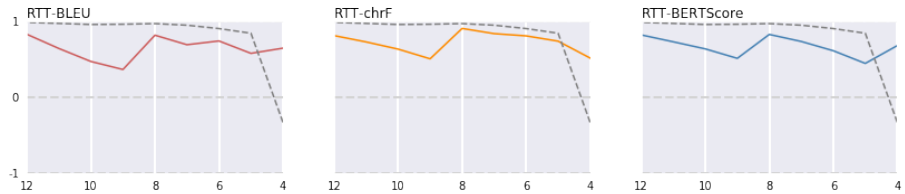
C.8 en-ru



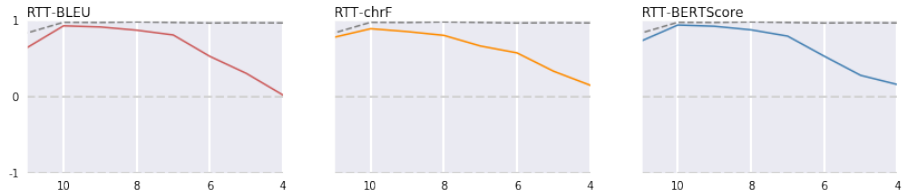
C.9 en-zh



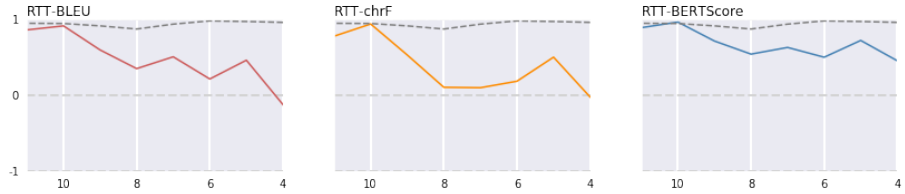
C.10 fi-en



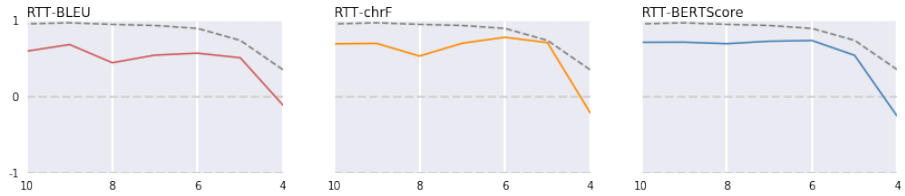
C.11 gu-en



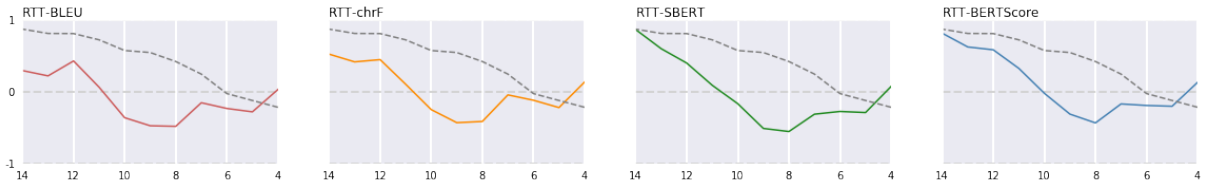
C.12 kk-en



C.13 lt-en



C.14 ru-en



C.15 zh-en