Better Evaluation of Very Good Translation Results

Zehao Guan, Liang Wu, Sean Zhang
{zehaog, liangwu, xiaoronz}@andrew.cmu.edu

Our code is available at https://github.com/zach96guan/11-731_Project.

1 Introduction

As the MT systems nowadays get better and better, it is more important to have evaluation metrics that can properly compare the best models. On the WMT19 Metrics Shared Task this year, although many evaluation metrics achieve a high correlation with human evaluation overall, most metrics have poor performance on very good machine translation systems [1]. As we begin to consider only the top-N MT systems, the correlation between the metrics and human evaluation falls quickly. It is important to understand why this phenomenon happens, so we want to explore the results and try to figure out in what ways these evaluation metrics are diverging from human assessment. We investigate the effect of semantic features in an evaluation metric [2], and add such features to improve existing metrics. We also try to fine-tune the context embedding-based metric ESIM using data of human evaluation of the best models [3].

2 Current Metrics

For the automatic evaluation of machine translation tasks, BLEU is by far the most commonly used metric. Nevertheless, a variety of other metrics are proposed, and they broadly fall into the following categories [2]:

- N-gram Matching: These metrics compare the candidate sentence and the reference sentence by looking at the amount of n-grams that they match in.
  - BLEU matches the 1-4 grams of the two sentences, and it includes a brevity penalty for short candidates. Some downsides of BLUE are that it does not take into account semantically similar words, and it does not consider word dependencies of long length.
  - METEOR uses unigram matching, except that it allows for synonyms and paraphrases in the candidate sentence.
- Edit Distance: Several metrics use word edit distance or word error rate, which is based on the number of edit operations required to get from the candidate to the reference. Some of these metrics include TER and the more recent characTER.
- Word Embeddings
  - Yisi-1 computes the word embeddings and matches the embeddings of both sentences to compute similarities.
  - BERTScore is similar, except that the contextual embedding based on BERT is used.

– From the results of WMT19 metrics shared task, metrics based on embeddings achieved the highest performance [1].

- Learned Metrics: Several metrics are trained using human judgements as supervision, including BEER and RUSE. But it's possible that these human judgements do not generalize well to texts in new domains.

3 WMT19 Results

	Metric	Features	Learned?	Scoring Level Seg	Scoring Level Sys
Baselines	sentBLEU	n-grams		•	–
	BLEU	n-grams		–	•
	NIST	n-grams		–	•
	WER	Levenshtein distance		–	•
	TER	edit distance, edit types		–	•
	PER	edit distance, edit types		–	•
	CDER	edit distance, edit types		–	•
	chrF	character n-grams		•	⊘
	chrF+	character n-grams		•	⊘
	sacreBLEU-BLEU	n-grams		–	•
	sacreBLEU-chrF	n-grams		–	•
Metrics	BEER	char. n-grams, permutation trees	yes	•	⊘
	BERTr	contextual word embeddings		•	⊘
	characTER	char. edit distance, edit types		•	⊘
	EED	char. edit distance, edit types		•	⊘
	ESIM	learned neural representations	yes	•	⊘
	LEPORa	surface linguistic features		•	⊘
	LEPORb	surface linguistic features		•	⊘
	Meteor++_2.0 (syntax)	word alignments		•	⊘
	Meteor++_2.0 (syntax+copy)	word alignments		•	⊘
	PReP	psuedo-references, paraphrases		•	⊘
	WMDO	word mover distance		•	⊘
	YiSi-0	semantic similarity		•	⊘
	YiSi-1	semantic similarity		•	⊘
	YiSi-1_srl	semantic similarity		•	⊘
QE Systems	IBM1-morpheme	LM log probs., IBM1 lexicon		•	⊘
	IBM1-pos4gram	LM log probs., IBM1 lexicon		•	⊘
	LP	contextual word emb., MT log prob.	yes	•	⊘
	LASIM	contextual word embeddings	yes	•	⊘
	UNI	?	?	•	⊘
	UNI+	?	?	•	⊘
	USFD	?	?	•	⊘
	USFD-TL	?	?	•	⊘
	YiSi-2	semantic similarity		•	⊘
	YiSi-2_srl	semantic similarity		•	⊘

Figure 1: Participants of 2019 WMT Metrics Shared Task

Figure 1 lists all the metrics that participated in the WMT19 Metrics Shared Task. According to the results, the metrics that are based on contextual word embeddings such as ESIM, BERTr and YiSi perform better in judging the best systems.

Figure 2 shows the correlation between evaluation metrics and human evaluations when restricted to top-N systems on the de-en language pair. We notice a significant decrease in correlation for

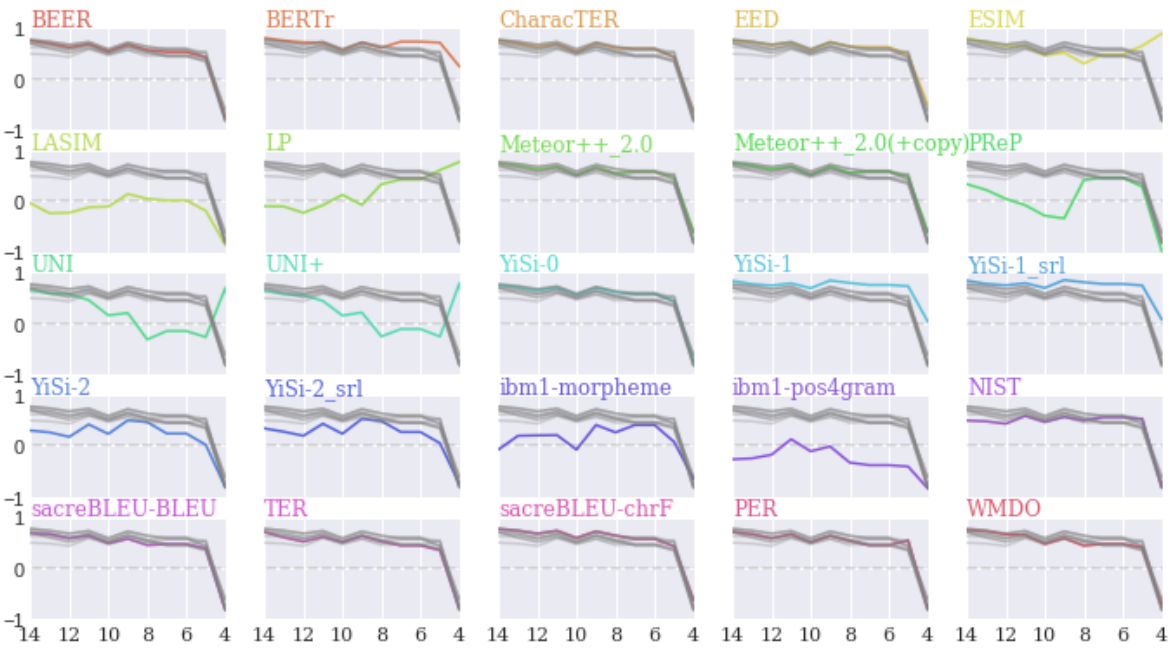

Figure 2: Correlation with human evaluation for top-N systems for aach metric on the de-en data

most metrics when n goes down to 4. Similar phenomenon happens in en-de, zh-en, and, to a slight extent, in ru-en and en-kk.

4 Dataset

Same as the WMT19 Metrics Shared Task, we are going to use the source and reference texts from the newstest2019 dataset. It contains 18 language pairs, each having approximately 2,000 sentences. In addition, the system outputs come from 233 systems across those language pairs from the WMT19 News Translation Task. The manual quality assessment that the metrics are compared with is done by Direct Assessment. When we consider correlation in this study, we only consider the system-level evaluation like Figure 2.

In our study, we will focus on the German-English language pair for now. The German-English language pair is a typical pair that has a downward trend in correlation as we consider top-N systems. As you can see in Figure 2, most metrics, including the baseline metrics drawn in grey, are dropping to zero or negative in terms of correlation when we start to consider the top-6 systems. We will evaluate our metrics using the same methodology, namely, finding the correlation of our scores with human evaluation for the top-N systems.

5 Qualitative Analysis of WMT19 Results

We focus on the language pair de-en, since this is where we observe a strong decrease in performance for top systems. We examined the entire WMT19 test set for de-en, which contains 2000

pairs of sentences. We looked at the original German text, the reference output, and the system outputs for the top systems. Our findings are as follows.

5.1 Quality of Reference Translation

Since the metric evaluation systems compares the reference translation to the candidate output, if the reference translation contains errors, the scoring of the best candidate outputs will be negatively affected.

The de-en dataset consists of news articles from various news websites. We notice that the quality of the reference translation varies from article to article. Some types of errors include omitting information, adding words, and incorrect translation that changes the meaning. Some examples of these errors can be found in appendix A.1.

We only listed a few examples out of the ones we found. We feel that for a dataset of 2000 sentence pairs, the amount of errors means that the reference data might not be a good basis to judge the top translations produced by MT systems.

5.2 Translation for Domain-Specific Text

We note that a significant amount of the test data is sports news, especially soccer news. However, the reference might not get the translation right, due to unfamiliarity with the sport or with the correct terms for the sport. Some examples of the errors can be found in appendix A.2.

We notice that for these domain-specific words, sometimes both the reference and the system outputs got the incorrect translation. However, sometimes the reference is wrong while the system outputs produce the correct sentence.

5.3 Document-Level vs. Sentence-Level Translation

In addition, we also notice an interesting phenomenon in the reference translation: some of the translations are completed at a sentence level, while others are done article by article. An example of this is shown in appendix A.3.

Note that this test set is used for the evaluation of all MT systems, including sentence-level ones and document-level ones. We can imagine that, for example, a good document-level system will be punished when evaluated using a reference translated sentence by sentence, and vice versa.

5.4 Discussion

The analysis above suggests that an imperfect reference might played a large role in why most evaluation metrics fail to distinguish the top models. It also explains why some of them do well: in Figure 2, some of the metrics that performed well for top systems are ESIM, LP, UNI, and UNI+. Other than ESIM, all other three are quality estimation systems used as a metric. These

metrics do not look at the reference translation, and this might be the reason why they outperform almost all other evaluation metrics.

6 Approaches

Based on the results and observations in the WMT19 Metrics Shared Task, we decided on the following directions to improve the existing metrics.

6.1 Evaluate BERTScore on top-N MT systems in WMT19

BERTScore is a recent paper that appeared after WMT19, in which the authors proposed a metric based on contextual word embedding calculated using BERT [2]. The paper shows that the BERTScore achieves good results when compared to the human evaluation in all WMT18 systems, but it does not show an analysis when we restrict to the top systems. We will evaluate its performance on top MT systems in WMT19, and compare the correlation results to other metrics in the WMT19 metrics shared task. This will give us more evidence on whether contextual embeddings are useful for judging the top systems.

6.2 Combining BERTScore with Quality Estimation via Cross-Lingual Embedding

From the discussion in Section 5, we believe that quality estimation can provide an advantage in distinguishing the top systems, especially when the reference translation is not perfect. We propose to combine BERTScore, a metric based on the embedding of the reference and candidate sentences, with a quality estimation approach, where we compute a metric based on the word embedding of the source and the candidate sentences.

To do that, we find the MUSE tool developed by Facebook which provides cross-lingual word embedding. Cross-lingual word embeddings essentially map tokens from different languages into the same vector space, and thus we can get the source-translation interaction. In our de-en language pair, we use MUSE's tokenizers, and convert the source and translation into vectors. Then we use the same methodology as BERTScore by calculating the precision, recall and F1 score for the cosine similarities between each pair of words from the source and translation. In the end, we want to combine the results from source-translation interaction and translation-reference interaction. A weighted average is calculated to get the final score, where the weights are determined by experimenting with different numbers from 0 to 1 with step size 0.1.

6.3 Fine-tuning ESIM using Human Evaluation of Top Systems

In the paper that comes up with the metric ESIM, the authors use BERT to generate word embeddings, and train a high-performing RNN model on the Natural Language Inference task on top of the embeddings to generate a score for each sentence. They train the model so that it generates scores as close to human evaluation as possible. Indeed, the results in WMT19 shows

that ESIM is one of the best metrics as it tops several language pairs it participated in. In addition, ESIM sometimes performs better than the baselines in ranking the best models.

We want to optimize the training methodology of ESIM by starting with their metric, and fine-tuning only on the human evaluation of the outputs of best models. In this way, we hope that the new model can better learn the minute differences among the best models. In this case, we have to find other data as our training data. We decide to use data from the Metric Shared Task in previous years.

7 Results and Analysis

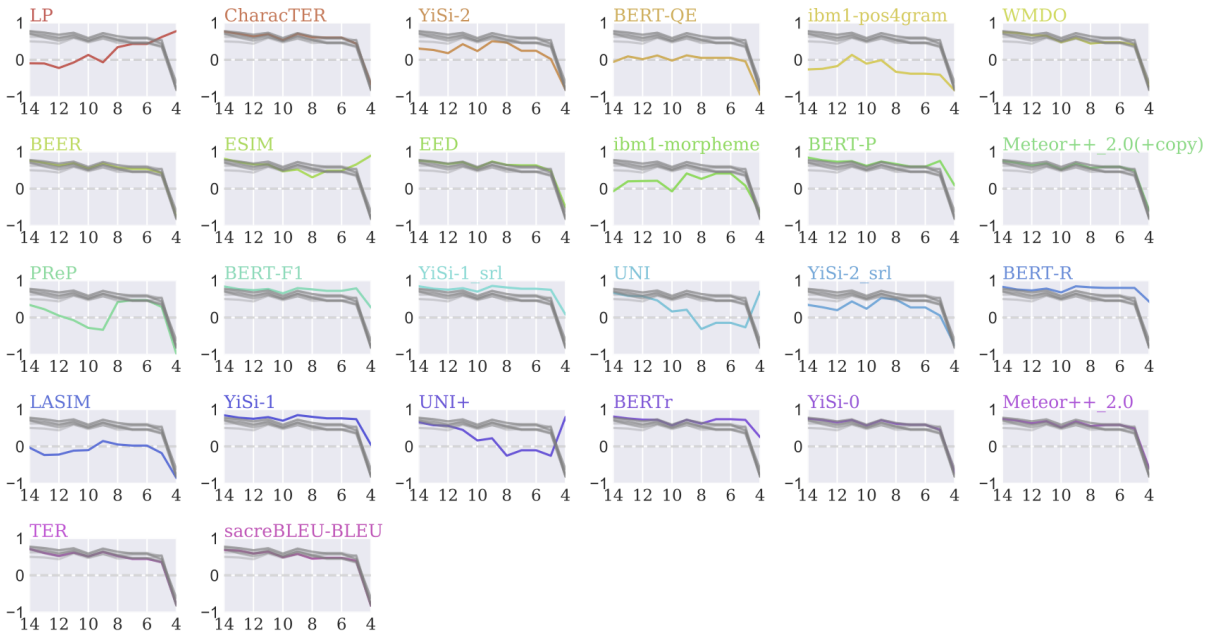

Figure 3: Add results of BERT-R, BERT-P, BERT-F1, BERT-QE over de-en.

Figure 3 and Figure 4 show the results of running BERTScore (BERT-R for recall, BERT-P for precision, and BERT-F1 for the F1 score) on the WMT19 de-en data. While most metrics show the downward trend as we have seen previously, BERTScore shows a promising trend that is better than others. Namely, the correlation doesn't go to zero or negative, but stays relatively positive, just like BERTr and ESIM. This is yet another proof that contextual embedding is capturing some trend in the best systems.

For BERTScore with QE, we see that the system is not very well correlated with human evaluations. Upon further examination of the system outputs, we see that every system has a score between 0.59 to 0.61, and in fact, most sentences have scores close to 0.6 range. This suggests that (1) the cross-lingual embedding does not map two languages exactly the same, since otherwise we expect more sentences to have score close to 1, (2) the fact that a lot of the sentences scored very similarly overall suggests that the precision-recall approach used by BERTScore may not be a good way to perform quality estimation, since it cannot distinguish a lot of the sentences.

	DA	BERT-P	BERT-R	BERT-F1	BERT-QE	BERTr	ESIM	YiSi-1	YiSi-2
DA	1.0	0.947	0.946	0.949	-0.274	0.926	0.941	0.949	0.796
DA	1.0	0.913	0.907	0.912	0.259	0.897	0.896	0.914	0.612
DA	1.0	0.842	0.832	0.841	-0.058	0.809	0.810	0.843	0.303
DA	1.0	0.768	0.759	0.769	0.092	0.758	0.720	0.778	0.266
DA	1.0	0.730	0.734	0.741	0.016	0.725	0.656	0.749	0.179
DA	1.0	0.744	0.784	0.773	0.116	0.720	0.655	0.794	0.423
DA	1.0	0.620	0.681	0.657	-0.022	0.569	0.470	0.697	0.236
DA	1.0	0.734	0.847	0.797	0.119	0.720	0.520	0.848	0.503
DA	1.0	0.669	0.818	0.763	0.051	0.617	0.307	0.801	0.468
DA	1.0	0.597	0.804	0.723	0.052	0.737	0.486	0.761	0.244
DA	1.0	0.597	0.804	0.723	0.052	0.737	0.486	0.761	0.244
DA	1.0	0.751	0.802	0.791	-0.043	0.718	0.656	0.738	0.021
DA	1.0	0.086	0.431	0.273	-0.941	0.251	0.895	0.045	-0.809
DA	1.0	0.321	0.039	0.190	-0.848	0.129	0.778	0.196	-0.882

Figure 4: Correlation of BERT-R, BERT-P, BERT-F1, BERT-QE and best metrics against DA.

8 Future Work

- Given the results of our improved metrics, we should continue to test their performance on other language pairs available in the newstest2019 dataset. We already see that different language pairs exhibit different characteristics. Some pairs do not have the downward trend, and some even have upward trend. It will be interesting to see how the added features perform in those pairs as well.
- In the BERTr paper, it describes some of the features that the metrics cannot catch, e.g. paraphrase of the same meaning, change in a few words that changes the meaning entirely. It is important to think of a way to take these features into consideration when designing new metrics on top of contextual embedding.
- Just as we use simple cross-lingual word embedding for source-translation interaction, it might be better to have some methods to find the cross-lingual contextual embedding.
- It would be interesting to see how contextual embedding perform if we consider not only single tokens, but two, three, or more tokens together. We can concatenate the contextual embeddings of consecutive tokens, just like how BLEU score looks at n-grams, and find the BERTScore thereof.

9 Contributions

Zehao Guan worked on experiments for BERTScore with WMT19 dataset and fine-tuned ESIM. Liang Wu worked on improving metrics with context evaluation. Sean Zhang worked on improving BERTScore with quality estimation via cross-lingual embeddings as well as the qualitative analysis of WMT19 results.

References

[1] Ma, Qingsong, et al. *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). 2019.

[2] Zhang, Tianyi, et al. *BERTScore: Evaluating Text Generation with BERT.* (2019).

[3] Mathur, Nitika et al. *Putting Evaluation in Context: Contextual Embeddings Improve Machine Translation Evaluation.* ACL (2019).

[4] Post, Matt. *A call for clarity in reporting BLEU scores.* (2018).

[5] Devlin, Jacob, et al. *Bert: Pre-training of deep bidirectional transformers for language understanding.* (2018).

[6] Kané, Hassan, et al. *Towards Neural Similarity Evaluator.* (2019).

[7] Barrault, Loïc, et al. Findings of the 2019 conference on machine translation (wmt19). *Proceedings of the Fourth Conference on Machine Translation* (Volume 2: Shared Task Papers, Day 1). 2019.

A Some examples of errors in the reference translation

In these examples, blue highlights the keywords and correct translations, while red indicates an incorrect translation.

A.1 Simple translation errors

German	Nach sechs Siegen zum Start fiel Liverpool auf Platz zwei zurück.
Reference	Following six wins at the start Liverpool went back two places.
Our translation	Following six wins at the start, Liverpool dropped to second place.
Facebook_FAIR	After six straight wins, Liverpool dropped to second place.
RWTH_Aachen	After six wins to start, Liverpool dropped to second.

German	Was Lohan zu dieser Aktion bewogen hat, ist derzeit völlig unklar.
Reference	What provoked Lindsay Lohan to such very strange actions is currently completely unclear.
Our translation	What prompted Lohan to take this action is completely unclear at this time.
Facebook_FAIR	It is unclear at this time what prompted Lohan to take this action.
RWTH_Aachen	What prompted Lohan to take this action is currently completely unclear.

A.2 Translation errors in domain-specific texts

German	[...]Ilkay Gündogan kam beim neuen Spitzenreiter nicht zum Einsatz.
Reference	[...]Ilkay Gündogan as a new attacking player was not used.
Our translation	[...]Ilkay Gündogan did not feature for the new front-runner.
Facebook_FAIR	[...]Ilkay Gündogan did not feature for the new leaders.
RWTH_Aachen	[...]Ilkay Gündogan did not feature in the new front-runner.

German	Der Brasilianer Neymar brachte PSG nach 22 Minuten in Führung und stellte in der Nachspielzeit den Endstand her.
Reference	Brazilian Neymar gave PSG the lead after 22 minutes and till the extra-time.
Our translation	Brazilian Neymar gave PSG the lead after 22 minutes and produced the final scoreline during the injury time.
Facebook_FAIR	Brazilian Neymar gave PSG the lead after 22 minutes and set up the final score in injury time.
RWTH_Aachen	Brazilian Neymar put PSG ahead after 22 minutes and produced the final score in the aftermath.

German	Zum Auftakt der Königsklasse hatte Tuchels Team beim 2:3 in Liverpool die bislang einzige Pflichtspiel-Niederlage in dieser Saison hinnehmen müssen.
Reference	At the start of the premier class the Tuchel's team with 2:3 in Liverpool is suffering only one defeat this season.
Our translation	At the start of the top-level tournament, Tuchel's team's 2:3 loss in Liverpool was the only defeat in official matches that they suffered this season.
Facebook_FAIR	Tuchel's team suffered their only defeat in the Premier League so far this season in the 2-3 draw at Liverpool.
RWTH_Aachen	At the start of the premier class, Tuchel's team had suffered the only mandatory defeat so far this season in the 2-3 draw at Liverpool.

A.3 Discrepancy between document-level and sentence-level translations

An example of an article translated sentence by sentence is as follows

German	Von az, aktualisiert am 04.05.2018 um 11:11 ⟨eos⟩ [...] ⟨eos⟩ Hvar - Flirten, kokettieren, verführen - keine einfachen Aufgaben für unsere Mädchen. ⟨eos⟩
Reference	From A-Z, updated on 04/05/2018 at 11:11 ⟨eos⟩ [...] ⟨eos⟩ Hvar with its flirting, coquetting, and seduction is not an easy task for our girls. ⟨eos⟩

Here Hvar is the location where the article was written, but reference translation assumed that it is part of the sentence. It's clear that the translator was not given the context of the article.

German	29. September 2018 um 19:45 Uhr ⟨eos⟩ Nizza Trainer Thomas Tuchel eilt mit Paris Saint-Germain in Frankreichs Fußball-Meisterschaft weiter von Sieg zu Sieg. ⟨eos⟩
Reference	29 September 2018 at 19:45 ⟨eos⟩ Nizza - At football championship in France coach Thomas Tuchel rushes with Paris Saint-Germain from victory to victory. ⟨eos⟩
Facebook_FAIR	September 29, 2018 at 7: 45 PM ⟨eos⟩ Nice coach Thomas Tuchel is rushing from victory to victory with Paris Saint-Germain in France's football championship. ⟨eos⟩
RWTH_Aachen	September 29, 2018 at 7:45 p.m. ⟨eos⟩ Nice coach Thomas Tuchel continues to rush from victory to victory with Paris Saint-Germain in France's football championship. ⟨eos⟩

Here same as the previous example, "Nizza" (Nice) refers to the city where the article was written. Yet due to incorrect segmentation, it appears as if the source sentence is talking about the coach of Nice (although the correct German would be "Nizza-Trainer"). Hence, both system outputs got the translation wrong, but the reference translation is correct, suggesting that this article was first translated and then segmented.