Better Evaluation of Machine Translation Systems

Liang Wu, Zach Guan, Sean Zhang

School of Computer Science, Carnegie Mellon University

Introduction

As the MT systems nowadays are getting better and better, it is important to have evaluation metrics that can properly compare them. On the **WMT19 Metrics Shared Task** this year, most evaluation metrics have poor performance on very good machine translation systems, while many of them achieve a high correlation with human evaluation in general.

As we begin to consider only the top-N MT systems, the correlation between the metrics and human evaluation falls quickly even to negative values. It would be important to know why this is, so we want to explore the results and try to figure out in what ways these evaluation metrics are diverging from human assessment. We will investigate the effect of **semantic features** in evaluation metrics and add such features to improve existing metrics.

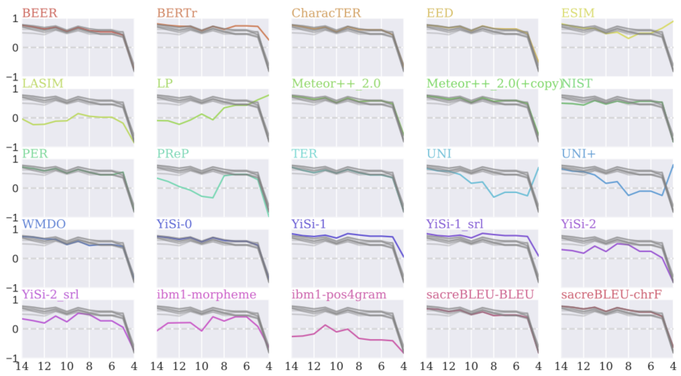

Figure1: Correlations for Top-N de-en MT System

Metrics

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

- **N-gram Matching**
  - ***BLUE*** 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 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 Embedding**
  - ***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 WMT19 metrics shared task, metrics based on embeddings achieved the highest performance.
- **Learned Metrics**
  several metrics were 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.

Baseline

	Metric	Features	Learned?	Scoring Seg	Scoring 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		•	⊘

Figure2: Participants of Metrics Shared Task with Corresponding Features

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.

Methods

- **Evaluating BERTScore on top MT systems in WMT19 and comparing with other metrics 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. However, the paper does not show an analysis **when we restrict to the top systems**. We will reproduce the experiments in the BERTScore paper and evaluate its performance on top MT systems in WMT19. We will then optimize the hyperparameters of BERTScore for top correlation.
- **Combining BERTScore with BLEU**. We note that the best results in WMT19 metrics shared task, in terms of correlation with human evaluation, are achieved by metrics that use word or sentence embedding. These embedding provide **semantic information** that can be used to decide the similarity of two sentences. On the other hand, BLEU is a commonly used evaluation metrics that does not take into account the semantic information of words at all. We will try to improve BLEU by considering the contextual word embeddings generated by BERT while matching n-grams of the sentences. We will evaluate our new metric on the WMT19 dataset.

References

- Mathur, Nitika et al. "Putting Evaluation in Context: Contextual Embeddings Improve Machine Translation Evaluation." *ACL* (2019).
- 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.
- Zhang, Tianyi, et al. "BERTScore: Evaluating Text Generation with BERT." *arXiv preprint arXiv:1904.09675* (2019).