Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing

Brian Thompson
Johns Hopkins University
brian.thompson@jhu.edu

Matt Post
Johns Hopkins University
post@cs.jhu.edu

Abstract

We frame the task of machine translation evaluation as one of scoring machine translation outputs with a sequence-to-sequence paraphraser, which works by force-decoding between an output and a human reference. We propose training the paraphraser as a multilingual NMT system, thereby treating paraphrasing as a *zero-shot translation* task (e.g., Czech to Czech). This setting produces a model that is less lexically and syntactically biased than monolingually-trained paraphrasers, because the mode of our model's output probability is centered around a copy of the sequence, which corresponds to the best-case scenario where an MT system perfectly matches a human reference. Our method is simple and intuitive, requiring no data other than that used to train standard MT systems. Our single model (trained in 39 languages) outperforms or statistically ties with all prior metrics on the WMT19 segment-level shared metrics task in all languages, and our approach outperforms settings that use monolingually-trained paraphrasers. We also explore using our model conditioned on the source instead of the reference, and find that, although it is not as good as reference-based evaluation, it still outperforms every "quality estimation as a metric" system from the WMT19 shared task on quality estimation by a statistically significant margin in every language pair.

1 Introduction

Machine Translation (MT) systems have improved dramatically in the past several years. This is largely due to advances in neural MT (NMT) methods, but the pace of improvement would not have been possible without automatic MT metrics, which provide immediate feedback on MT quality without the time and expense associated with obtaining human judgments of MT output.

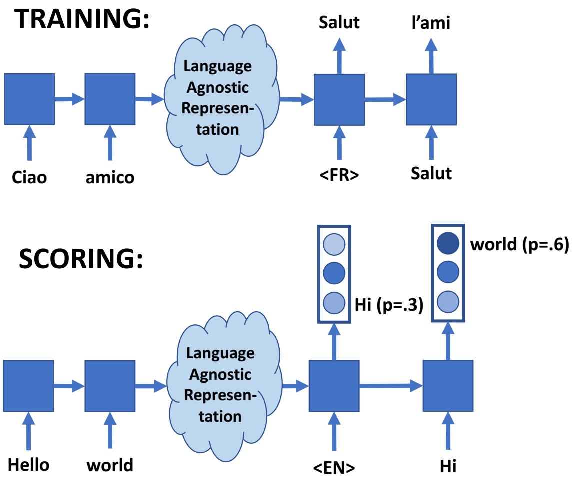

Figure 1: PRISM is trained on multilingual parallel examples such as *Ciao amico*, here translated to French as *Salut l'ami.* At evaluation time, the model is used in zero-shot mode to score MT system outputs conditioned on their corresponding human references. For example, the MT system output *Hi world* conditioned on the human reference *Hello world* is found to have token probabilities [0.3, 0.6].

Since its introduction nearly two decades ago, BLEU (Papineni et al., 2002) has been the dominant metric for machine translation. However, BLEU has a number of fairly serious weaknesses. To begin with, it is a surface-based metric that can only compute observable overlap between a system output and a reference—and usually, only one such reference is available. BLEU is continually outperformed by other metrics in the WMT metrics tasks (cf. Ma et al. (2018, 2019), and recent work has shown that its relatively poor correlation with human judgments have at best been *underestimated* (Mathur et al., 2020). Over-reliance on BLEU could cause good ideas to be mistakenly rejected or bad ones to stick around too long.

We propose a modern metric that addresses many of these issues. Our approach is to use a sentential, sequence-to-sequence paraphraser to score MT outputs by force-decoding to them, conditioned on their corresponding human references. Our model implicitly represents the entire (exponentially large) set of potential paraphrases of a sentence, both valid and invalid; by "querying" the model with a particular system output, we can use the model score to measure how well the system output paraphrases the human reference translation.

The best possible MT output is one which perfectly matches a human reference; therefore, for evaluation, an ideal paraphraser would be one with an output distribution centered around a copy of its input sentence. We refer to such a model as (lexically and syntactically) *unbiased*, and we propose to construct it by using a multilingual NMT system that treats paraphrasing as zero-shot translation (e.g., Czech to Czech). This is in contrast to the standard way of training such paraphrasers, which is to assemble and train from synthetic paraphrase data (Wieting and Gimpel, 2018), so as to produce models capable of *generating* diverse paraphrases. We show that our multilingual NMT approach is much closer to an ideal lexically/syntactically unbiased paraphraser than a generative paraphraser trained on synthetic paraphrases. It also allows a single model to work in many languages, and can be applied to the task of "Quality estimation (QE) as a metric", by conditioning on the source instead of the reference. Figure 1 illustrates our method, which we denote PRISM (Probability is the metric).

Trained in 39 languages, our single model:

- Outperforms or ties with prior metrics and several contrastive neural methods on the segment-level WMT 2019 MT metrics task in every language pair;
- Is able to discriminate between very strong neural systems at the system level, addressing a problem raised at WMT 2019; and
- Significantly outperforms all QE metrics submitted to the WMT 2019 QE shared task.

Finally, we contrast the effectiveness of our model when scoring MT output using the source, instead of the human reference. We observe that human references improve performance, and, crucially, allow our model to rank systems that are *substantially better than our model at the task of translation.* This is important because it establishes that our method does not require building a state-of-the-art multilingual NMT model in order to produce a state-of-the-art MT metric capable of evaluating state-of-the-art MT systems.

Nothing in our method is specific to sentence-level MT. In future work, we would like to extend Prism to paragraph- or document-level evaluation by training a paragraph- or document-level multilingual NMT system, as there is growing evidence that MT evaluation would be better conducted at the document level, rather than the sentence level (Läubli et al., 2018).

Our approach is described in detail in Thompson and Post (2020). The model, toolkit, and training data are publicly available.[1]

References

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

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

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the EMNLP*, pages 90–121, Online.

[1] `https://github.com/thompsonb/prism`

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.