Filtering Pseudo-References by Paraphrasing for Automatic Evaluation of Machine Translation

Ryoma Yoshimura **Hiroki Shimanaka** **Yukio Matsumura**
Hayahide Yamagishi **Mamoru Komachi**
Tokyo Metropolitan University, Tokyo, Japan
{yoshimura-ryoma, shimanaka-hiroki, matsumura-yukio
yamagishi-hayahide}@ed.tmu.ac.jp, komachi@tmu.ac.jp

Abstract

In this paper, we introduce our participation in the WMT 2019 Metric Shared Task. We propose a method to filter pseudo-references by paraphrasing for automatic evaluation of machine translation (MT). We use the outputs of off-the-shelf MT systems as pseudo-references filtered by paraphrasing in addition to a single human reference (gold reference). We use BERT fine-tuned with paraphrase corpus to filter pseudo-references by checking the paraphrasability with the gold reference. Our experimental results of the WMT 2016 and 2017 datasets show that our method achieved higher correlation with human evaluation than the sentence BLEU (SentBLEU) baselines with a single reference and with unfiltered pseudo-references.

1 Introduction

In general, automatic evaluation of MT is based on n-gram agreement between the system output and a manually translated reference of the source sentence. Therefore, automatic evaluation fails to evaluate a semantically correct sentence if the surface of the system output differs from that in the reference. To solve this problem, many automatic evaluation methods allow the use of multiple references that potentially cover various surfaces; in particular, Finch et al. (2004) reported that correlation between automatic evaluation results and human evaluation increases when multiple references are used for evaluation. However, owing to the time and costs involved in manually creating references, many datasets only include one reference per source sentence, which leads to improper translation evaluation, especially in the case of diverse machine translation systems.

In order to obtain cheap references without any human intervention, Albrecht and Hwa (2008) used the outputs of off-the-shelf MT systems as pseudo-references; They showed that using multiple references consisting of gold and pseudo-references may yield higher correlation with human evaluation than using a single gold reference. However, because they did not consider the quality of the pseudo-references, this may result in using poor references. Thus, in some cases the correlation becomes worse when using multiple references consisting of gold and pseudo-references relative to only using a gold reference.

To address the quality of pseudo-references, we filtered pseudo-references by checking the paraphrasability to the gold reference. Our approach can be applied to various MT evaluation metrics which can be evaluated with multiple references. The experimental results show that our method achieves higher correlation with human evaluation than the previous work.

2 Related Work

Albrecht and Hwa (2008) showed that using the outputs of off-the-shelf MT systems as pseudo-references in n-gram based metrics such as BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011) may yield higher correlation with human evaluation than using a gold reference. They use the outputs of off-the-shelf MT systems as they are, whereas we filter them by paraphrasing the gold reference.

Kauchak and Barzilay (2006) proposed a method to obtain a paraphrase of a gold reference that is closer in wording to the system output than the gold reference for MT evaluation. They evaluated an MT system using only the generated references, whereas we evaluated MT systems using multiple references, including those obtained by adding generated references to the gold reference. They generate a paraphrase of a gold reference, whereas we translate source sentences and identify whether the outputs are paraphrases of gold references. That is, they used only gold references whereas we used both source and gold

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 521–525
Florence, Italy, August 1-2, 2019.

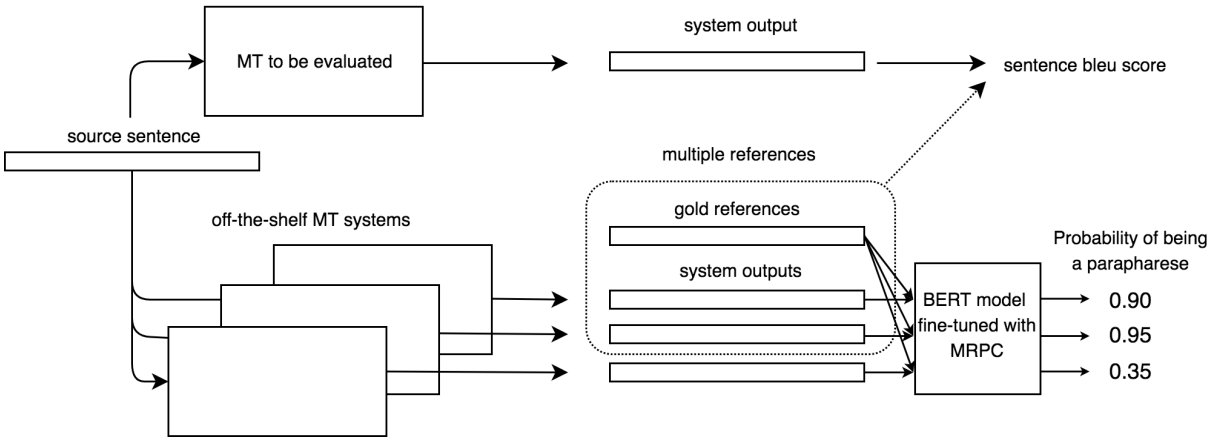

Figure 1: Overview of the proposed method.

reference information.

3 MT Evaluation Metric Using Filtered Multiple Pseudo-References

3.1 Overview

Figure 1 shows the overview of our proposed method. The procedure of our proposed method is as follows.

1. Prepare off-the-shelf MT systems for generating pseudo-references.

2. Translate the source sentence in the evaluation data using the abovementioned MT systems.

3. Filter the outputs of off-the-shelf MT systems by checking the paraphrasability of being a paraphrase to the single gold reference.

4. Calculate the sentence evaluation score with multiple references obtained by adding filtered pseudo-references to the single gold references.

3.2 Automatic pseudo-reference generation

Any MT system can be used as a pseudo-reference generation system except for the translation system to be evaluated. [1] There are no restrictions on the type of MT systems, such as neural machine translation (NMT) or statistical machine translation (SMT) systems, or the number of MT systems.

[1]If the system to be evaluated were used as a pseudo-reference generation system, the output would be used as a reference.

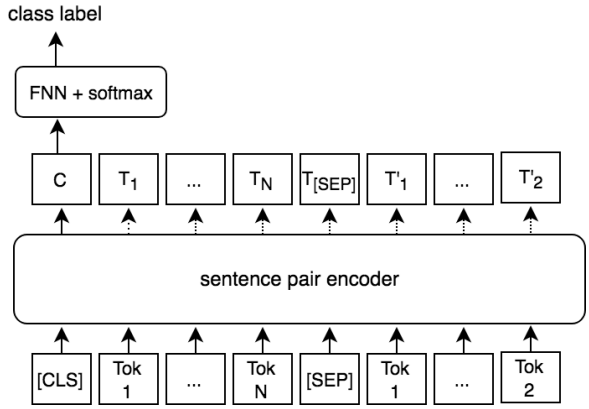

Figure 2: BERT model architecture for sentence pair classification.

3.3 Filtering by paraphrasing

We use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) to filter pseudo-references by checking the paraphrasability with a gold reference. BERT is a new approach to pre-train language representations, and it obtains state-of-the-art results on a wide variety of natural language processing (NLP) tasks, including question answering (QA), semantic textual similarity (STS), natural language inference (NLI). The key to pre-training BERT is the prediction of masked words and of the next sentence. Masking words allows bidirectional learning, which improves joint training of language context relative to Embeddings from Language Models (ELMo) (Peters et al., 2018), which combines forward and backward training. Prediction of the next sentence leads to capturing the relationship between two sentences.

Figure 2 shows the BERT model architec-

	cs-en	de-en	fi-en	ru-en
single reference	0.557	0.484	0.448	**0.502**
single reference + pseudo-references	0.565	0.499	0.543	0.456
single reference + filtered references (MAS)	0.576	0.473	0.517	0.469
single reference + filtered references (BERT)	**0.589**	**0.519**	**0.572**	0.490

Table 1: Segment-level Pearson correlation between SentBLEU and human evaluation scores in WMT 2016.

	cs-en	de-en	fi-en	ru-en
single reference	0.435	0.433	0.571	0.484
single reference + pseudo-references	0.515	0.565	0.653	0.519
single reference + filtered references (MAS)	0.524	**0.586**	0.650	0.517
single reference + filtered references (BERT)	**0.555**	0.580	**0.671**	**0.545**

Table 2: Segment-level Pearson correlation between SentBLEU and human evaluation scores in WMT 2017.

corpus	train	dev	test	Accuracy
MRPC	3,669	408	1726	0.845

Table 3: Numbers of sentences in each split of MRPC and accuracy of BERT.

ture for sentence pair classification. In classification tasks where labels are attached to sentence pairs, BERT encodes sentence pairs together with a [CLS] token for classification and a [SEP] token for sentence boundaries; The output of the [CLS] token is used for the input of classifier of a feedforward neural network with softmax. BERT achieves state-of-the-art performance in a paraphrase identification task on the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) with this architecture.

For that reason, we use BERT to estimate the paraphrasability between pseudo-references and the gold reference. We fine-tune BERT with MRPC. The output of the classifier is the probability of the paraphrase from 0 to 1. We use pseudo-references whose paraphrase probability is greater than 0.5.

4 Experiments

4.1 Data

We used the segment-level evaluation datasets of Czech-English (cs-en), German-English (de-en), Finnish-English (fi-en), Russian-English (ru-en) language pair from WMT 2016 (Bojar et al., 2016) and 2017 (Bojar et al., 2017). The datasets consist of 560 pairs of sources and references,

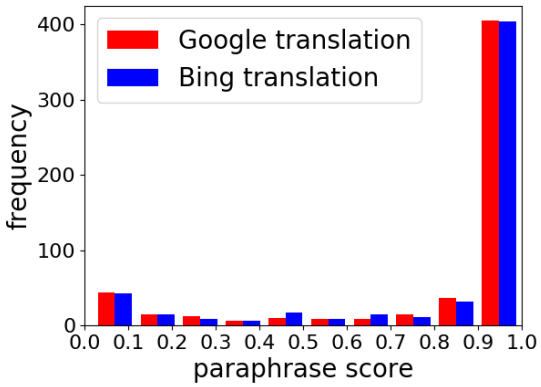

Figure 3: Histograms of paraphrase score of pseudo-references in the fi-en language pairs of WMT 2016.

along with the outputs of each system and human evaluation scores.

4.2 Off-the-shelf MT systems

We used Google Translation [2] and Bing Microsoft Translator [3] as MT systems to generate pseudo-references. We chose these two MT systems because they are widely used, easy to use, and well known to have good performance. We automatically translated source files using each translation API.

4.3 Fine-tuning BERT with MRPC

We use the pre-trained BERT-Base Uncased model [4], which has 12 layer, 768 hidden, 12 heads

[2]https://translate.google.com/
[3]https://www.bing.com/translator
[4]https://github.com/google-research/bert

system output	gymnastics and freestyle exercises - where bayles defends the title of world champion - lie in the veil .
gold reference	balance beam and floor exercise - where biles is the defending world champion - lay in wait .
pseudo-reference (Google)	gymnastics log and floor exercises - where biles defends the world champion title - lie in wait . (0.994)
pseudo-reference (Bing)	gymnastic log and freestyle exercises — where the bayles defends the title of world champion — lie in ambush . (0.215)
human score: -1.497; **SentBLEU**: single reference: -1.118, without filtering: -0.335, filtering: -1.662	

Table 4: Example of pseudo-references in ru-en language pair of WMT 2017; The value in parentheses at the end of each pseudo-reference indicates the paraphrase score by BERT. Each score is standardized according to the mean and standard deviation to compare human evaluation and each SentBLEU score.

and 110M parameters. We fine-tuned BERT with MRPC. MRPC is a dataset extracted from web news articles along with human annotations indicating whether each pair is a paraphrase. If the pair is paraphrase, the label is 1, if not, the label is 0. The original dataset consists of 4,077 sentences for training and 1,726 sentences for testing. We divided the test set in half and used it as development data. The numbers of sentences in each corpus and the accuracy of the fine-tuned BERT model are listed in Table 3.

Figure 3 shows the histogram of paraphrase score of pseudo-references in the fi-en language pair of WMT 2016. Due to the use of high quality MT systems, more than 50% of the pseudo-references have paraphrase scores between 0.9 and 1.0. The same trend was observed in all languages and years.

4.4 Evaluation

We calculated the SentBLEU score with system output and multiple references which consisted of a single gold reference and pseudo-references. The SentBLEU is computed using the sentence-bleu.cpp [5], a part of the Moses toolkit. It is a smoothed version of BLEU (Lin and Och, 2004). We followed the tokenization method for each year's dataset. We measured Pearson correlation identically to WMT 2016 and WMT 2017 between the automatic and human evaluation scores. In order to compare with our method, we also performed filtering by Maximum Alignment Similarity (MAS) (Song and Roth, 2015), which is one of the unsupervised sentence similarity measures based on alignments between word embeddings and is known to achieve good performance on Semantic Textual Similarity (STS) task. We used GloVe [6] (Pennington et al., 2014) as word embeddings. We used pseudo-references whose MAS score is higher than 0.8.

5 Results

Tables 1 and 2 show the segment-level Pearson correlation coefficients between automatic and human evaluation scores. The result shows that our proposed method outperforms the baselines except in the case of the ru-en language pair in WMT 2016 and filtering by MAS does not produce any consistent result.

6 Discussion

Table 4 shows an example of pseudo-references with BERT's paraphrase score for the ru-en language pair in WMT 2017. The pseudo-reference from Bing translation has a low paraphrase score because "biles" in the gold reference remains as "bayles" in the pseudo-reference, and "floor exercise" became "freestyle exercise" in Bing translation. In the unfiltered method, the BLEU score is unreasonably high because the surface of the pseudo-reference from Bing translation is similar to the output sentence. Filtering the pseudo-references prevents the problem. The pseudo-reference from Google translation has different surfaces but carry the same meaning as in the gold reference. Our filtering method correctly retains the sentence because BERT assigned high paraphrase score.

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

[6] https://nlp.stanford.edu/projects/glove/ Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors)

7 Conclusions

We proposed a method to filter pseudo-references in terms of paraphrasability with a gold reference that addresses the problem of using poor pseudo-references from previous work (Albrecht and Hwa, 2008). We use BERT fine-tuned with MRPC to filter pseudo-references. By filtering pseudo-references in terms of paraphrasability with a gold reference, we can keep the references having the same meaning with the gold reference but different surface and solve the problem of using poor pseudo-reference from previous work. The experimental results show that our method outperforms baselines.

Acknowledgement

We would like to thank Tomoyuki Kajiwara for providing a script to calculate the MAS score.

References

Joshua S. Albrecht and Rebecca Hwa. 2008. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT2008)*.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation (WMT2017)*.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation (WMT2016)*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP2011)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL2019)*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Andrew M Finch, Yasuhiro Akiba, and Eiichiro Sumita. 2004. How does automatic machine translation evaluation correlate with human scoring as the number of reference translations increases? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004)*.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (NAACL2006)*.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLLING2004)*.

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 (ACL2002)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP2014)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL2018)*.

Yangqiu Song and Dan Roth. 2015. Unsupervised sparse vector densification for short text similarity. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (NAACL2015)*.