Multilingual Machine Translation Evaluation Metrics Fine-tuned on Pseudo-Negative Examples for WMT 2021 Metrics Task

Kosuke Takahashi[1], **Yoichi Ishibashi**[1], **Katsuhito Sudoh**[1,2], **Satoshi Nakamura**[1]
[1] Nara Institute of Science and Technology
[2] PRESTO, Japan Science and Technology Agency
{takahashi.kosuke.th0, ishibashi.yoichi.ir3, sudoh, s-nakamura}@is.naist.jp

Abstract

This paper describes our submission to the WMT2021 shared metrics task. Our metric is operative to segment-level and system-level translations. Our belief toward a better metric is to detect a significant error that cannot be missed in the real practice cases of evaluation. For that reason, we used pseudo-negative examples in which attributes of some words are transferred to the reversed attribute words, and we build evaluation models to handle such serious mistakes of translations. We fine-tune a multilingual largely pre-trained model on the provided corpus of past years' metric task and fine-tune again further on the synthetic negative examples that are derived from the same fine-tune corpus. From the evaluation results of the WMT21's development corpus, fine-tuning on the pseudo-negatives using WMT15-17 and WMT18-20 metric corpus achieved a better Pearson's correlation score than the one fine-tuned without negative examples. Our submitted models are named C-SPEC (Cross-lingual Sentence Pair Embedding Concatenation) and C-SPECpn, are the plain model using WMT18-20 and the one additionally fine-tuned on negative samples, respectively.

1 Introduction

Recent studies of automatic evaluation is mostly based on the family models of BERT (Devlin et al., 2019). BERTscore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020) have shown a strong correlation with human judgement scores. However, we reported in our previous study (Takahashi et al., 2020), it is hard for BERT based metrics to correctly evaluate the translation errors that are annotated with low Direct Assessment (DA) score.

Upon the problems of evaluating poor quality translations, Sudoh et al. (2021) has attempted to solve the problem by creating a different human annotation set and corpus. Compared to DA, their idea is to make a clear definition of critical translation errors and let models learn the critical errors that can cause a serious misunderstanding.

Following the idea, we used pseudo-negative examples to train the evaluation model. Since, empirically, the cases of the evaluation failure happens frequently with the nouns translation errors, we generated pseudo-negative sentences by transferring the attribute of nouns with Word Attribute Transfer (Ishibashi et al., 2020). This system is based on our previous work (Takahashi et al., 2020), with an extension with fine-tuning with the pseudo-negative examples.

2 Related Work

BERTscore (Zhang et al., 2020), BERT regressor (Shimanaka et al., 2019), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020) are applications of BERT to the machine translation evaluation. BERTscore measures the similarity of reference and hypothesis translation by the cosine-similarity of the token embeddings for each token in the reference and hypothesis. It uses a pre-trained BERT model without fine-tuning on evaluation data. Instead, BERT regressor and BLEURT are fully parameterized and require a human annotated evaluation corpus to fine-tune the models. Both metrics have the same model architecture; a linear layer is attached on top of the BERT encoder. They encode a paired reference and hypothesis sentence with BERT and predict the human evaluation score. Additionally, BLEURT conducts warm-up training of BERT before fine-tuning on an evaluation corpus. The model architecture of our submission is similar to BERT regressor and BLEURT, but its uniqueness comes from using the synthetic negative data to fine-tune the models to evaluate poor translations better.

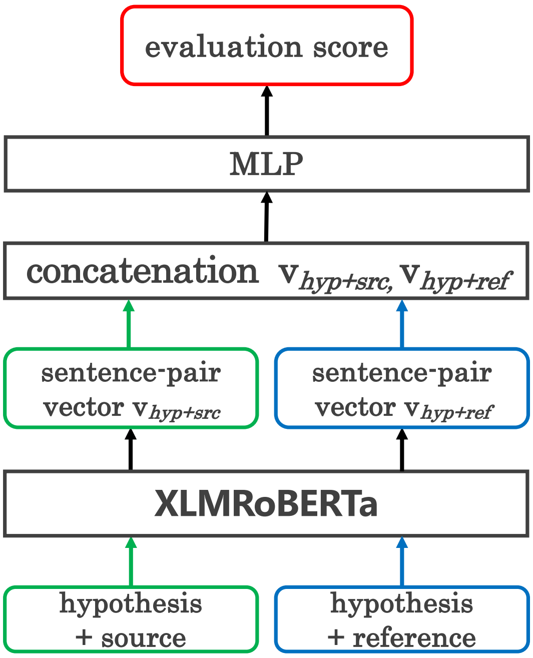
evaluation score
MLP
concatenation $v_{hyp+src}, v_{hyp+ref}$
sentence-pair vector $v_{hyp+src}$
sentence-pair vector $v_{hyp+ref}$
XLMRoBERTa
hypothesis + source
hypothesis + reference

Figure 1: Architecture of C-SPEC

3 Our system

3.1 Model architecture

We extend the BERT regressor (Shimanaka et al., 2019) and use a cross-lingual language models, XLMRoBERTa (Conneau et al., 2020), to utilize a source sentence as a pseudo reference. In order to obtain a sentence-pair vector from source language and target language sentences together, our model encodes input sentences with a cross-lingual language model instead of monolingual BERT.

The model procedure is illustrated in the Figure 1. Our metric, called C-SPEC (Cross-lingual Sentence Pair Embedding Concatenation), uses paired inputs of hypothesis-source and hypothesis-reference. It introduces another vector for hypothesis-source ($v_{hyp+src}$) in addition to the standard one for hypothesis-reference pair ($v_{hyp+ref}$) to make an ensemble evaluation. Both sentence vectors are concatenated and used to predict the evaluation score in multi-layer perceptron (MLP). At first of the evaluation process, the cross-lingual language model encodes an input sentence into a sentence-pair vector. Then, using the sentence-pair vector, a MLP outputs the final evaluation score in regression manner. In training, we used standardized z score of DA (Direct Assessment; Graham et al. (2013)) as the ground truth and updated the model parameters by backpropagation (Rumelhart et al., 1986) with Mean Squared Error (MSE) Loss.

Our model was trained by the following steps. Firstly, in order to speed up and stabilize the training procedure, our models were trained on the corpus of WMT2015-2016's DA. Secondly, the models were additionally trained on WMT2015-2017's DA, WMT2018-2020's DA, or WMT2015-2020's DA. Thirdly, they were fine-tuned with the pseudo negative examples. Lastly, they were fine-tuned again on the WMT20's MQM segment-level corpus.

In each step of fine-tuning, we initialize the output-layer and only inherit the parameters of XLMRoBERTa. We tried three different conditions in the second step because DA corpus after WMT2018 is relatively noisy, and removing those data may play out well.

In the system-level evaluation, we simply averaged the segment-level evaluation scores for each system.

3.2 Word Attribute Transfer

We used the reflection-based word attribute transfer Ishibashi et al. (2020) for data augmentation. This transfer can make conversion of words into a certain word attribute, such as *queen* to *king*, using parameterized mirrors composed of two multi-layer perceptrons.

For the pseudo-negative hypothesis generation, we used two types of word attribute transfer in gender (male/female) and antonym. The word attribute transfer was applied onto all the words in an input sentence, and words having a target attribute were rewritten into their transferred counterparts while those that were not related to the target attribute were kept unchanged. For example, a sentence *"It is our duty to remain at his sides", he said, to applause* is transferred into *"It is our duty to change at his sides", he said, to whisper*, by the antonym transfer. Note that the word attribute transfer may not make any changes on an input sentence when all the words were identified as non-related words. We eliminated such sentences from our pseudo-negative examples.

3.3 Fine-tuning using pseudo-negative examples

Our pseudo-negative examples were obtained from the reference sentences in the training corpus of all-English that was used to firstly fine-tune a model, because the word attribute transfer model works only in English. However, we did not have any DA scores on these pseudo-negative examples. So, we used them to fine-tune the evaluation models in classification manner. We introduced a different output layer on the top of the model illustrated in Figure 1 to classify an input example into the following categories:

1. A hypothesis is the same as its original system translation.
2. A hypothesis is the same as its reference.
3. A hypothesis is from the pseudo-negative examples

In the fine-tuning, we used three types of inputs corresponding to the classes above, and the models were trained to discriminate them. We expected such fine-tuned models to identify the serious word choice translation errors given in the pseudo-negative examples. We call the metric trained using the pseudo-negative examples C-SPECpn (pn:psuedo-negative).

4 Segment-level evaluation experiments

Our experiment was conducted on the development data for WMT21 metric task, which is randomly selected 10% of WMT20 MQM segement-level corpus. All the results were calculated by the Pearson's correlation with the MQM segment scores.

4.1 Results

The results of the WMT20 MQM segment-level corpus are shown in Table 1.

From the results, the models trained on negative examples of WMT15-17 and WMT18-20 overcame the plain models in Pearson's correlation. Among the models, the best score was archived by the one trained on WMT18-20 with the negative examples. Although WMT15-20 is a larger corpus than WMT18-20, the score of plain models was negligible at best, and the model trained on WMT18-20 and with negative examples did not overcome the plain one.

In order to figure out whether and how fine-tuning on the negative examples had impact on the evaluation performance, we calculated the Pearson's correlation for each small chunk of segment-level MQM scores and visualized the gap between models' outputs in Figure 2. Both the models trained on WMT15-17 and WMT18-20 with negative examples performed better in the MQM range of [-25.0, -5.0) and [-0.1, 0.0]. This suggests that using negative examples can improve the performance of evaluating high and critically low quality translations. However, the model trained on WMT15-20 with negative examples dropped its performance in the [-25.0, -5.0) range compared to the plain model. We assume the reason of the score drop is that the model was overly fine-tuned to the high quality translations, as it can be seen that the Pearson's correlation score in the [-0.1, 0.0] improved tremendously.

5 Conclusion

In this paper, we presented a BERT-based multilingual evaluation metric that is boosted by pseudo-negative examples to evaluate poor translations more precisely. Our model leverages our previous work Takahashi et al. (2020) and have shown an improvement of Pearson correlation when fine-tuning on the synthetic examples in the WMT15-17 and WMT18-20 corpus settings.

Acknowledgements

This work is supported by JST PRESTO (JPMJPR1856). With the support of JST, we used RIKEN's miniRAIDEN in a part of the experiments in this work.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yoichi Ishibashi, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. 2020. Reflection-based word attribute transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*,

Table 1: Pearson's correlation with MQM segment scores in WMT2020. C-SPEC stands for a plain model fine-tuned without negative examples. C-SPECpn is a model fine-tuned on negative examples.

model	en-de	zh-en	avg	all
C-SPEC w/ WMT15-17	0.609	0.773	0.691	0.787
C-SPEC w/ WMT18-20	0.612	0.805	0.708	0.813
C-SPEC w/ WMT15-20	0.603	0.798	0.700	0.808
C-SPECpn w/ WMT15-17	**0.626**	0.809	0.717	0.817
C-SPECpn w/ WMT18-20	0.619	**0.824**	**0.721**	**0.829**
C-SPECpn w/ WMT15-20	0.309	0.715	0.512	0.724

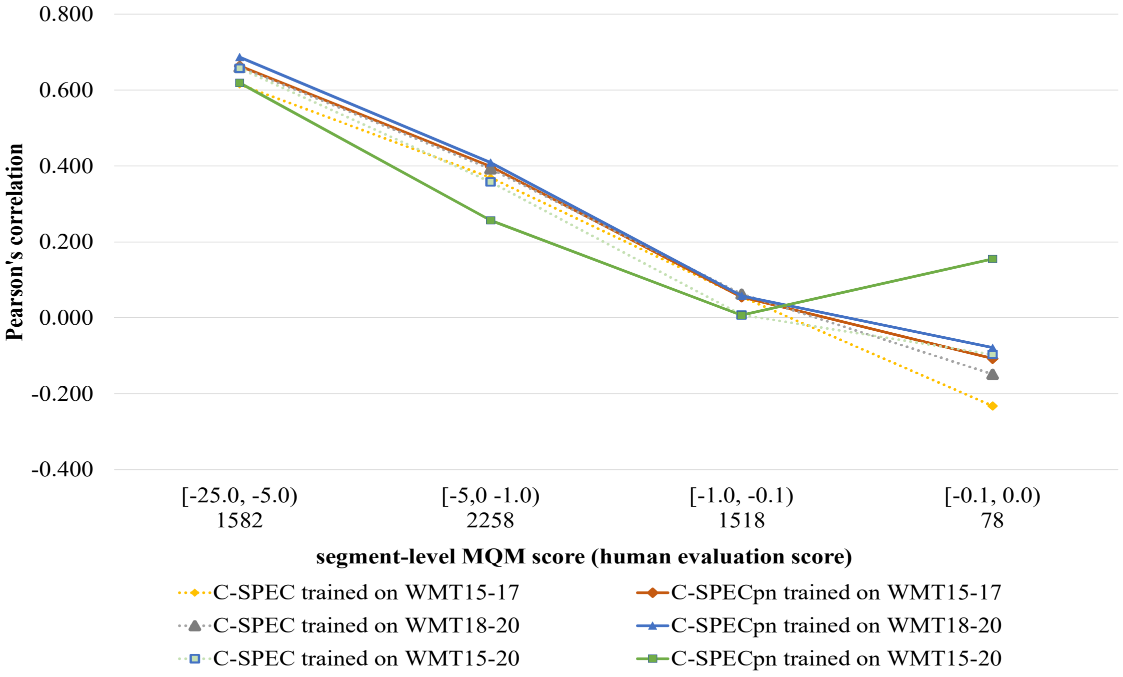

Figure 2: Pearson's correlation for each small segment-level MQM ranges. The amount of each segment is written below the range description.

pages 51–58, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

David Rumelhart, Geoffrey Hinton, and Ronald Williams. 1986. Learning Representations by Backpropagating Errors. *Nature*, 323(6088):533–536.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine Translation Evaluation with BERT Regressor. *arXiv preprint*, abs/1907.12679.

Katsuhito Sudoh, Kosuke Takahashi, and Satoshi Nakamura. 2021. Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 46–55, Online. Association for Computational Linguistics.

Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3553–3558, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.