Automatic Machine Translation Evaluation using Source Language Inputs and Cross-lingual Language Model

Kosuke Takahashi[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, sudoh, s-nakamura}@is.naist.jp

Abstract

We propose an automatic evaluation method of machine translation that uses source language sentences regarded as additional pseudo references. The proposed method evaluates a translation hypothesis in a regression model. The model takes the paired source, reference, and hypothesis sentence all together as an input. A pretrained large scale cross-lingual language model encodes the input to sentence-pair vectors, and the model predicts a human evaluation score with those vectors. Our experiments show that our proposed method using Cross-lingual Language Model (XLM) trained with a translation language modeling (TLM) objective achieves a higher correlation with human judgments than a baseline method that uses only hypothesis and reference sentences. Additionally, using source sentences in our proposed method is confirmed to improve the evaluation performance.

1 Introduction

Automatic machine translation evaluation (MTE) has been studied to substitute human evaluation in machine translation development because it is low-cost, handy, and stable to use. Popular automatic MTE metrics such as BLEU (Papineni et al., 2002) calculate the evaluation score based on a surface-level similarity of a paired 1-to-1 reference and translated hypothesis sentences. BLEU particularly evaluates the sentence similarity with the n-gram word matching rate between a reference and hypothesis. However, the evaluation score drops when a reference and hypothesis are dissimilar in the surface even if they share the same meaning.

To counter this problem, METEOR (Banerjee and Lavie, 2005) is proposed to mitigate the word matching of synonyms with a synonym dictionary. Yet still, with mitigation of word matching, surface-level similarity cannot fully compensate for semantics, thus word representation instead of word symbols is used in Word Mover's Distance (Kusner et al., 2015) and bleu2vec (Tättar and Fishel, 2017).

Besides, sentence representation is known to be an efficient feature instead of word representation because sentence vectors can represent more global meanings. RUSE (Shimanaka et al., 2018) and BERT (Devlin et al., 2019) based MTE, BERT regressor (Shimanaka et al., 2019), utilized sentence representation and performed well on WMT17 Metric Shared Task (Bojar et al., 2017). The metrics mentioned above compare a hypothesis translation to a reference. However, a reference translation represents only one possible translation and those MTE metrics are unlikely to correctly evaluate all candidates that share the same meanings of the reference or have fatally different meanings due to a few translation errors. This problem can be mitigated by the use of multiple reference translations as argued by Dreyer and Marcu (2012) and Qin and Specia (2015), but preparing such multiple references is costly.

Hereby, we propose a method to incorporate source sentence into MTE as another pseudo reference, since the source and reference sentences should be semantically equivalent. The proposed method uses Cross-lingual Language Model (XLM) (Lample and Conneau, 2019) to handle source and target languages in a shared sentence embedding space. The proposed method with XLM trained with a translation language modeling (TLM) objective showed a higher correlation with human judgments than a baseline method using hypothesis and reference sentences.

2 Related Work

Recent advances in sentence-level embedding have been used in MTE. Shimanaka et al. (2018) proposed an MTE framework called RUSE (Regressor Using Sentence Embeddings), which uses sentence-

Table 1: Available corpus size annotated with human judgments in WMT-2017 Metrics Shared Task (to-English)

	cs-en	de-en	fi-en	lv-en	ro-en	ru-en	tr-en	zh-en	{de,ru,tr,zh}-en	all-en
WMT-2015	500	500	500	-	-	500	-	-	1000	2000
WMT-2016	560	560	560	-	560	560	560	-	1680	3360
WMT-2017	560	560	560	560	-	560	560	560	2240	3920
ALL	1620	1620	1620	560	560	1620	1120	560	4920	9280

level embeddings obtained by a large-scale pre-trained model like InferSent (Conneau et al., 2017), Quick Thought (Logeswaran and Lee, 2018), and Universal Sentence Encoder (Cer et al., 2018). Its regressor takes sentence vectors for a reference and translation hypothesis as inputs and returns a score, which is trained to correlate well with human evaluation (Graham et al., 2015). RUSE achieved the best correlation score with human judgments in the WMT-2017 Metrics Shared Task (Bojar et al., 2017).

BERT regressor (Shimanaka et al., 2019) is a simple MTE metric based on BERT (Devlin et al., 2019) encoder. It is composed of BERT encoder and a multi-layer perceptron (MLP) regressor attached to the last layer of BERT. This BERT encoder is a 12 layers bi-directional language model, referring to $\text{BERT}_{\text{base}}$(uncased)[1], trained with masked language model (MLM) and next sentence prediction (NSP). BERT regressor surpassed RUSE on the WMT-2017 data.

3 Proposed method: Automatic evaluation using XLM

We propose an MTE method using source language sentences as additional *pseudo* references. We use cross-lingual language models called XLM (Lample and Conneau, 2019) to encode both source and target language sentences into an embedding vector.

XLM has three additional techniques to BERT: language independent subword based on Byte Pair Encoding (Sennrich et al., 2016), a language embedding layer, and a translation language modeling (TLM) objective that predicts masked words from surrounding words or a paired translation. The brief architecture of XLM is shown in Figure 1. (Lample and Conneau, 2019) reported that XLM trained with TLM objective obtains better performance than multilingual BERT(Devlin et al., 2019) on the XNLI cross-lingual classification task(Conneau et al., 2018).

The proposed method has two variants for the use of source language sentences, as illustrated in Figure 2. The first one called hyp+src/hyp+ref uses two sentence-pair vectors for hypothesis-source and hypothesis-reference, encoded by a cross-lingual language model independently. These sentence-pair vectors are given to an MLP-based regression model to predict the human evaluation scores. This can be regarded as an ensemble model using a monolingual vector based on the reference and a cross-lingual vector based on the source sentence. The other one called hyp+src+ref takes a concatenation of hypotheses, source, and reference sentences as an input to a cross-lingual language model to obtain a sentence-pair vector. This sentence-pair vector is expected to be directly learned to represent the quality of the translation hypothesis given two correct sentences aligned aside.

4 Experiments

We conducted experiments to evaluate the performance of the proposed method in MTE by comparing with some existing methods.

4.1 Setting

The experiments were conducted with a corpus of all language pairs to English translation from segment-level WMT2017 Metrics Shared Task (Bojar et al., 2017). We split sentences in WMT15 and WMT16 to training and development data with the ratio of 9:1 and whole sentences in WMT17 are used for evaluation of MTE methods. The corpus size for each language pair is shown in Table 1.

We used two different models from all available XLM family models[2]: XLM15 pretrained by MLM and TLM, and XLM100 pretrained only by MLM. XLM15 is expected to perform better by the paired bilingual training of TLM, but the number of available languages is limited. XLM15 is compatible with only German, Russian, Turkish, and Chinese in the corpus, which confines the model to partial access to the corpus. On the other hand, XLM100

[1] https://github.com/google-research/bert

[2] https://github.com/facebookresearch/XLM

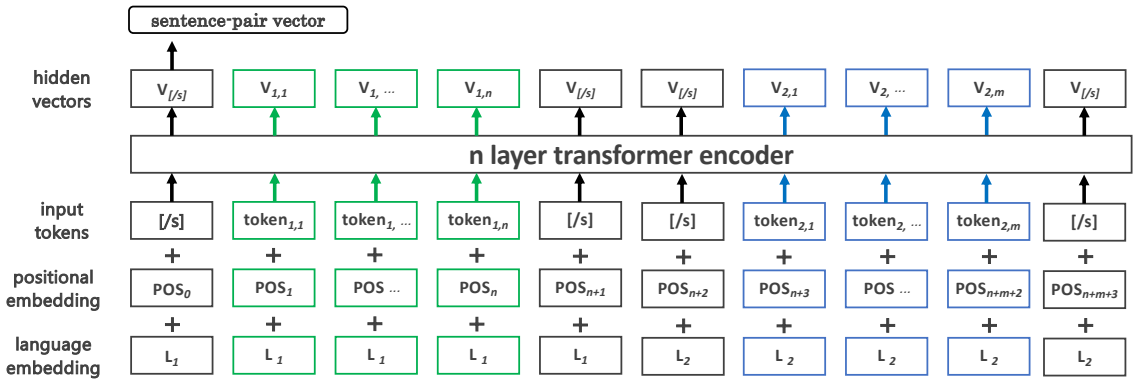
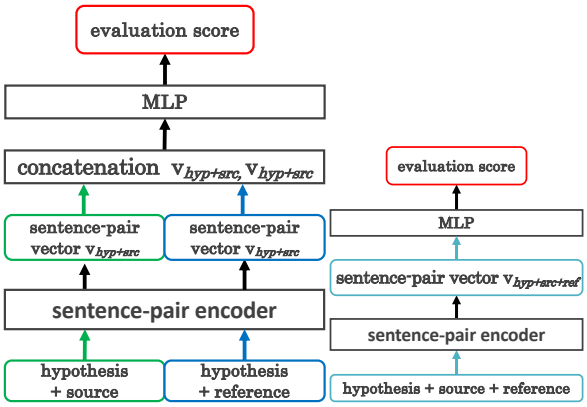


Figure 1: The architecture of XLM sentence-pair encoder



(a) hyp+src/hyp+ref (b) hyp+src+ref

Figure 2: Two variants in the proposed method

is compatible with all language pairs in the corpus, while it lacks supervised bilingual pretraining.

Thus the experiments had two corpus settings; One was a small corpus including {German (de), Russian (ru), Turkish (tr), and Chinese (zh)} to English (en) language pairs, and the other was a whole corpus including {Czech (cz), German (de), Finnish (fi), Latvian (lv), Romanian (ro), Russian (ru), Turkish (tr), and Chinese (zh)} to English language pairs. The evaluation was conducted with Pearson's correlation to human judgments in the test set.

We compared the proposed methods with SentBLEU (Bojar et al., 2017), BERT regressor (Shimanaka et al., 2019) by our implementation. We also conducted experiments using multilingual BERT, $\text{BERT}_{\text{multi}}$(cased), to contrast language models and experiments limiting the model's input into source-hypothesis only and reference-hypothesis only to study the impact of adding source sentences.

The fine-tuning on the proposed methods and BERT regressor was based on Mean Squared Error (MSE) loss in the training set, back-propagated to both MLP and XLM in order. The hyperparameters were selected through grid search for the following parameters. Since models are affected by randomness in training, we ran ten experiments for each of the settings and report results of the average scores.

- Optimizer : {Adam}
- Learning rate : {3e-5, 1e-5, 9e-6, 7e-6}
- Number of epochs : {1, ...,20}
- Dropout rate: {0.1}
- Batch size : {2, 4, 8, 16}

Table 2: Pearson's correlation scores in the small corpus ({de,ru,tr,zh}-en)

	de-en	ru-en	tr-en	zh-en	avg
SentBLEU	0.432	0.484	0.538	0.512	0.484
BERT regressor	0.729	0.757	0.770	0.702	0.740
multi-BERT					
hyp+src/hyp+ref	0.661	0.739	0.768	0.735	0.726
hyp+src+ref	0.625	0.713	0.725	0.691	0.689
hyp+src	0.520	0.558	0.601	0.559	0.559
hyp+ref	0.627	0.688	0.718	0.685	0.679
XLM15					
hyp+src/hyp+ref	**0.753**	**0.795**	**0.771**	**0.763**	**0.771**
hyp+src+ref	0.729	0.769	0.767	0.725	0.747
hyp+src	0.722	0.763	0.761	0.668	0.728
hyp+ref	0.716	0.787	0.746	0.714	0.741
XLM100					
hyp+src/hyp+ref	0.643	0.722	0.725	0.712	0.701
hyp+src+ref	0.635	0.695	0.715	0.661	0.677
hyp+src	0.464	0.450	0.557	0.449	0.480
hyp+ref	0.631	0.718	0.695	0.702	0.687

4.2 Results

The results of each small corpus and whole corpus experiments are shown in Tables 2 and 3, respectively. Note that XLM15 was not included in the

Table 3: Pearson's correlation scores in the whole corpus (all-en)

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg
SentBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
BERT regressor	**0.776**	**0.753**	**0.863**	**0.818**	**0.788**	**0.803**	**0.767**	**0.795**
multi-BERT								
hyp+src/hyp+ref	0.743	0.688	0.824	0.812	0.772	0.796	0.751	0.769
hyp+src+ref	0.714	0.670	0.802	0.774	0.754	0.758	0.722	0.742
hyp+src	0.599	0.525	0.699	0.681	0.586	0.633	0.571	0.613
hyp+ref	0.720	0.681	0.823	0.806	0.744	0.768	0.748	0.756
XLM100								
hyp+src/hyp+ref	0.712	0.681	0.822	0.810	0.756	0.773	0.745	0.757
hyp+src+ref	0.698	0.666	0.818	0.795	0.742	0.765	0.727	0.745
hyp+src	0.510	0.531	0.672	0.662	0.543	0.602	0.537	0.580
hyp+ref	0.692	0.666	0.813	0.788	0.743	0.746	0.714	0.738

whole corpus experiment due to its limited language coverage.

Performance of each language model As we can see from Table 2, the proposed method using XLM15 with hyp+src/hyp+ref structure surpassed BERT regressor in the small corpus. However, XLM100 did not work well in the experiments; its results were much worse than the others in the small corpus condition, and it did not compete with BERT regressor in the whole corpus condition as shown in Table 3. One possible reason is the lack of TLM objective pretraining in XLM100. Since the TLM task allows the model for learning semantically equivalent cross-lingual sentences directly, the TLM task can be concluded to be important for using source sentences in MTE. The results of multilingual BERT are worse than BERT regressor and XLM15, but close to XLM100 or slightly better in general. From this comparison of pretraining objectives and language models, we report that our proposed method is influenced by the multilingualism of a language model.

hyp+src/hyp+ref VS hyp+src+ref The results from Table 2 and Table 3 shows that hyp+src/hyp+ref structure is better than hyp+src+ref in most of the conditions, although we expected hyp+src+ref to perform better because it can access 2 translation answers as references at the same time. This is probably because both of XLM and multilingual BERT was not pretrained to handle 3 sentences in a sequence. However, it is perhaps possible that hyp+src+ref surpasses hyp+src/hyp+ref when a fine-tuning corpus is large enough.

Table 4: Pearson's correlation score in the halved small corpus {de,ru,tr,zh}-en

	de-en	ru-en	tr-en	zh-en	avg
BERT regressor	0.686	0.731	**0.753**	**0.691**	0.715
multi-BERT					
hyp+src/hyp+ref	0.583	0.670	0.720	0.675	0.662
hyp+src+ref	0.563	0.664	0.704	0.698	0.657
hyp+src	0.384	0.509	0.629	0.482	0.501
hyp+ref	0.574	0.651	0.722	0.693	0.660
XLM15					
hyp+src/hyp+ref	**0.712**	0.744	0.740	0.690	**0.722**
hyp+src+ref	0.679	**0.748**	0.706	0.666	0.700
hyp+src	0.570	0.635	0.654	0.616	0.619
hyp+ref	0.682	0.707	0.708	0.700	0.699
XLM100					
hyp+src/hyp+ref	0.594	0.676	0.706	0.686	0.666
hyp+src+ref	0.605	0.644	0.676	0.639	0.631
hyp+src	0.321	0.408	0.447	0.431	0.402
hyp+ref	0.559	0.631	0.675	0.668	0.633

Contribution of adding source sentences Every model with hyp+src/hyp+ref achieved a better score than both of hyp+src and hyp+ref, which indicates that source sentences contribute to the improvement of evaluation.

5 Analysis

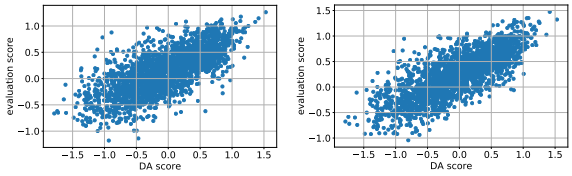
Training data size We conducted another experiment to see the effect of the training corpus size using randomly halved {de, ru, tr, zh}-en small corpus. From the results in Table 4, BERT regressor stably performed well even when the number of training data is about 1000 sentences, however, XLM15, XLM100, and multilingual BERT deteriorated their performances. Since our proposed hyp+src/hyp+ref is an ensemble model and has a more complex network structure than hyp+ref, the

Table 5: Pearson's correlation score for low and high human judgement score range in the small corpus ({de,ru,tr,zh}-en)

	All	DA ≥ 0.0	DA < 0.0	Reduction rate of Pearson's score from DA ≥ 0.0 to DA < 0.0 (%)
BERT regressor	0.728	0.553	0.464	16.10
multi-BERT hyp+src/hyp+ref	0.728	0.535	0.494	**7.77**
multi-BERT hyp+src+ref	0.686	0.512	0.423	17.51
multi-BERT hyp+src	0.539	0.339	0.316	**6.88**
multi-BERT hyp+ref	0.672	0.493	0.384	22.05
XLM15 hyp+src/hyp+ref	0.768	0.580	0.529	**8.68**
XLM15 hyp+src+ref	0.740	0.560	0.497	11.12
XLM15 hyp+src	0.679	0.469	0.430	**8.46**
XLM15 hyp+ref	0.735	0.534	0.458	14.20
XLM100 hyp+src/hyp+ref	0.703	0.535	0.419	21.75
XLM100 hyp+src+ref	0.662	0.501	0.389	22.29
XLM100 hyp+src	0.522	0.337	0.292	13.42
XLM100 hyp+ref	0.685	0.521	0.378	27.48

use of XLM and multi-BERT with the proposed method requires a certain amount of training data. Therefore, our proposed method deteriorated in the halved small corpus setting. On the other hand, monolingual BERT and hyp+ref benefits from the large corpus because it has no language limitation other than English.

Evaluation errors In order to see when models make errors to evaluate hypothesis sentences, we plot scatters of evaluation scores and human judgement scores (DA scores) in Figure 3(a), Figure 3(b), Figure 3(c), and Figure 3(d). Although in comparison, the evaluation scores of our best model XLM15 hyp+src/hyp+ref are set more linearly than the baseline BERT regressor, the scores of all models seem much dispersed in the low DA area (DA < 0.0). This indicates that all evaluation models listed here tend to miss-evaluate when a hypothesis is poor. Furthermore, we show Pearson's correlation score for each of high and low DA score range in Table 5. As we confirmed in the scatter figures, the correlation scores of low DA is low; evaluation models work poorly when hypotheses are poor. However, the reduction rate of Pearson's scores from high DA to low DA is small with XLM15 hyp+src/hyp+ref and hyp+src. Therefore adding source sentences has an impact to stabilize the evaluation performance when hypotheses are low-quality.



(a) BERT regressor r = 0.728 (b) XLM15 hyp+src/hyp+ref r = 0.768

(c) XLM15 hyp+src r = 0.679 (d) XLM15 hyp+ref r = 0.735

Figure 3: Scatter plots of human judgement scores (DA scores) and evaluation scores

6 Conclusion

In this paper, we proposed an MTE framework that utilizes source sentences using XLM. We show that the proposed method with TLM-trained XLM showed a higher correlation with human judgments than the baseline method in the small corpus condition and stabilize the evaluation performance regardless of the quality of translation sentences by using additional source sentences. We also investigated why our proposed method worked poorly in the other conditions and found the importance of TLM training. In future work, we will work around the problem of evaluation errors in the low DA range.

Acknowledgments

This work is supported by JST PRESTO (JPMJPR1856).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task". In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. pages 2475–2485. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 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.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate Evaluation of Segment-level Machine Translation Metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings to Document Distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pages 957–966.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint*, abs/1901.07291.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

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.

Ying Qin and Lucia Specia. 2015. Truly Exploring Multiple References for Machine Translation Evaluation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 113–120, Antalya, Turkey.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

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

Andre Tättar and Mark Fishel. 2017. bleu2vec: the Painfully Familiar Metric on Continuous Vector Space Steroids. In *Proceedings of the Second Conference on Machine Translation*, pages 619–622, Copenhagen, Denmark. Association for Computational Linguistics.