Machine Translation Reference-less Evaluation using YiSi-2 with Bilingual Mappings of Massive Multilingual Language Model

Chi-kiu Lo and **Samuel Larkin**
Multilingual Text Processing
Digital Technologies Research Centre
National Research Council Canada (NRC-CNRC)
1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
{chikiu.lo,samuel.larkin}@nrc-cnrc.gc.ca

Abstract

We present a study on using YiSi-2 with massive multilingual pretrained language models for machine translation (MT) reference-less evaluation. Aiming at finding better semantic representation for semantic MT evaluation, we first test YiSi-2 with contextual embeddings extracted from different layers of two different pretrained models, multilingual BERT and XLM-RoBERTa. We also experiment with learning bilingual mappings that transform the vector subspace of the source language to be closer to that of the target language in the pretrained model to obtain more accurate cross-lingual semantic similarity representations. Our results show that YiSi-2's correlation with human direct assessment on translation quality is greatly improved by replacing multilingual BERT with XLM-RoBERTa and projecting the source embeddings into the target embedding space using a cross-lingual linear projection (CLP) matrix learnt from a small development set.

1 Introduction

The machine translation quality estimation as a metric (QE as a metric) task was first introduced in WMT 2019 (Ma et al., 2019; Fonseca et al., 2019) to encourage the exploration of reference-less evaluation metrics. QE as a metric task shifts the use case of the QE systems from assisting professional translators to estimate post-editing efforts to assisting MT developers or general MT users to discriminate the translation quality of different MT systems without the presence of a human reference translation. YiSi-2, the reference-less variants of the YiSi metric (Lo, 2019), was the only metric who participated in evaluating all the translation directions in WMT 2019 QE as a metric shared task.

The QE as a metric task is very similar to Task 1 (Sentence-level direct assessment) of WMT20's quality estimation shared task where metric performance is evaluated in terms of correlation at the sentence-level with human direct assessment scores on translation quality. The subtle but crucial difference between the WMT20 QE Task 1 and the QE as a metric task is that QE systems for the former task is trained specifically to estimate the quality of a single MT system whereas QE metrics for the latter task is generalized for multiple machine translation systems. The QE systems for WMT20's QE Task 1 have access to the MT system that generate the translations while the reference-less metrics for the latter task have no information on the MT systems being evaluated.

In WMT 2019 metrics shared task, pretrained multilingual BERT (Devlin et al., 2018) was used in YiSi for both MT reference-based (YiSi-1) and reference-less (YiSi-2) evaluation in all tested translation directions where monolingual pretrained BERT model was not available for the target language (such as Czech, German, etc.). Since then, another massive multilingual pretrained language model, XLM-RoBERTa (Conneau et al., 2020), has been published. We evaluate the use of contextual embeddings extracted from each of the intermediate layers of the two models in MT reference-less evaluation.

In addition, despite using the same pretrained embedding model of last year, YiSi-2 showed a significant performance degradation when comparing to YiSi-1. For example, segment-level correlation with human direct assessment for evaluating English→Czech drops from 0.475 (YiSi-1) to 0.069 (YiSi-2). This shows that the cross-lingual semantic representation in pretrained multilingual BERT is not as accurate as the monolingual semantic representation for each language. In other words, we observed the language clustering effect where a clear segregation of vector subspace among different languages in the multilingual contextual em-

Proceedings of the 5th Conference on Machine Translation (WMT), pages 903–910
Online, November 19–20, 2020.

bedding model. Inspired by Zhao et al. (2020), we employ a weakly-supervised bilingual mapping learnt from a small development set that transforms the contextual embeddings of the source sentence to the target subspace for better cross-lingual semantic similarity evaluation.

In this paper, we show that YiSi-2's correlation with human direct assessment on translation quality is greatly improved by replacing multilingual BERT with XLM-RoBERTa$_{large}$ using the optimal intermediate layer (7^{th} layer count from the last) and projecting the source embeddings into the target embedding space using a cross-lingual linear projection matrix learnt from a small development set.

2 YiSi-2

YiSi (Lo, 2019) is a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. YiSi-1 measures the similarity between a machine translation and human references by aggregating weighted distributional (lexical) semantic similarities, and optionally incorporating shallow semantic structures. Improvements in YiSi-1 for WMT 2020 metrics shared task is detailed in (Lo, 2020).

YiSi-2 is the bilingual, reference-less version, which uses bilingual word embeddings to evaluate cross-lingual lexical semantic similarity between the input and MT output.

2.1 Massive Multilingual Pretrained Language Models

YiSi-2 relies on a cross-lingual language representation to evaluate the cross-lingual lexical semantic similarity. Previously, it used pretrained multilingual BERT (Devlin et al., 2018) for this purpose. BERT captures the sentence context in the embeddings, such that the embedding of the same subword unit in different sentences would be different from each other and be better represented in the embedding space. Since multilingual BERT is trained on the concatenation of non-parallel data from each language, the circular dependency deadlock between parallel resource and cross-lingual semantic similarity is broken (Lo and Simard, 2019). Multilingual BERT covers the 104 largest languages in Wikipedia.

XLM-RoBERTa (Conneau et al., 2020) (XLM-R) is also a massive multilingual pretrained language model. Similar to BERT, XLM-R is also trained with a masked language model task on the concatenation of non-parallel data. The differences between XLM-R and BERT are 1) XLM-R is trained on the CommonCrawl corpus which is significantly larger than the Wikipedia training data used by BERT; 2) instead of a uniform data sampling rate used in BERT, XLM-R uses a language sampling rate that is proportional to the amount of data available in the training set. Because of these differences, XLM-R performs better on low resource languages than multilingual BERT. XLM-R covers 100 languages. In this work, we use XLM-R$_{large}$ for the best performance on cross-lingual semantic similarity.

As suggested by Devlin et al. (2018); Peters et al. (2018); Zhang et al. (2020), we experimented using contextual embeddings extracted from different layers of the multilingual language encoder to find out the layer that best represents the semantic space of the language.

2.2 Inuktitut-English Cross-lingual Language Model

Since Inuktitut is neither covered by pretrained multilingual BERT nor XLM-RoBERTa, we trained our own Inuktitut-English XLM (Lample and Conneau, 2019) using the Nunavut Hansard 3.0 (NH) parallel corpus (Joanis et al., 2020). The model was trained with masked language model and translation language model tasks. The Inuktitut-English XLM model has 12 layers with 8 heads and embedding size of 512.

2.3 Cross-lingual Linear Projection

In the WMT 2019 metrics shared task (Ma et al., 2019), we saw a very significant performance degradation between YiSi-1 and YiSi-2. This shows that current multilingual language models construct a shared multilingual space in an unsupervised manner without any direct bilingual signal, in which representations of context in the same language are likely to cluster together in part of the subspace and there is a language segregation in the shared multilingual space. Inspired by Artetxe et al. (2016) and Zhao et al. (2020), we obtain subword token pairs from the news translation task development set for each language (each contains around 1k to 3k sentence pairs) aligned by maximum alignment of their semantic similarities. We then train a cross-lingual linear projection (Zhao et al., 2020) that transforms the source embeddings into the target embeddings subspace.

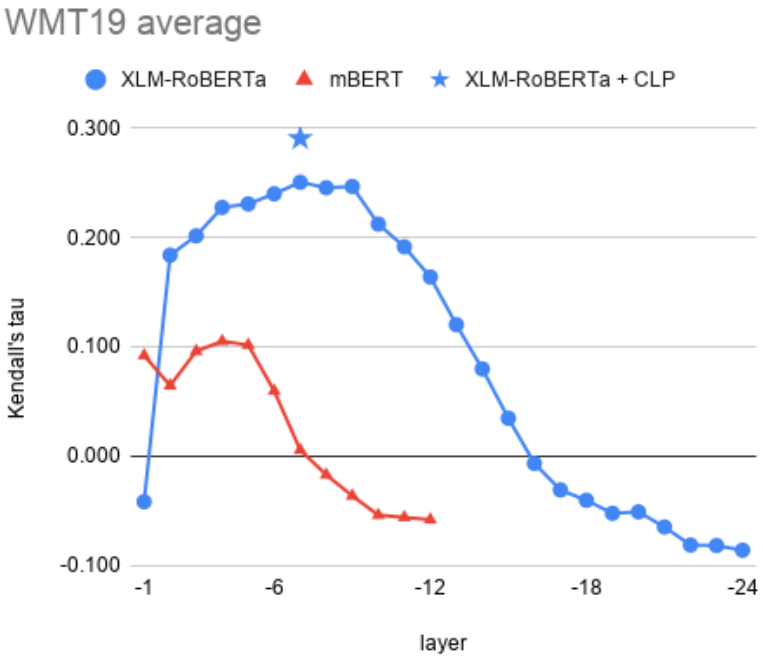

Figure 1: Segment-level Kendall's τ correlation with human direct assessment averaged over all WMT 2019 news translation test sets of YiSi-2 using contextual embeddings extracted from different layers of the multilingual pre-trained language models. On the x-axis, layer $-n$ means, YiSi-2 based on the embeddings of the n^{th} layer, counting from the last, of XLM-RoBERTa$_{large}$ (blue circles), multilingual BERT (red triangles) and layer -7 of of XLM-RoBERTa$_{large}$ with source embeddings projected to target language space using CLP (blue star).

Table 1: Segment-level Kendall's τ correlation of metric scores with the WMT 2019 official human direct assessment judgments.

input	de	fi	gu	kk	lt	ru	zh	en	en	en	en	en	en	en	en
output	en	en	en	en	en	en	en	cs	de	fi	gu	kk	lt	ru	zh
Reference-based evaluation metric															
YiSi-1 (2019)	.164	.347	.312	.440	.376	.217	.426	.475	.351	.537	.551	.546	.470	.585	.355
YiSi-0	.117	.271	.263	.402	.289	.178	.355	.406	.304	.483	.539	.494	.402	.535	.266
sentBLEU	.056	.233	.188	.377	.262	.125	.323	.367	.248	.396	.465	.392	.334	.469	.270
QE as a metric															
YiSi-2 (2020)	.116	.271	.249	.370	.281	.121	.340	.299	.329	.459	.512	.459	.314	.078	.158
YiSi-2 (2019)	.068	.126	-.001	.096	.075	.053	.253	.069	.212	.239	.147	.187	.003	-.155	.044

Table 2: Segment-level Kendall's τ correlation of metric scores with the WMT 2019 official human direct assessment judgments.

input	de	de	fr
output	cs	fr	de
Reference-based evaluation metric			
YiSi-1 (2019)	.376	.349	.310
YiSi-0	.331	.296	.277
sentBLEU	.203	.235	.179
QE as a metric			
YiSi-2 (2020)	.355	.294	.226
YiSi-2 (2019)	.199	.186	.066

3 Results

We use WMT 2019 metrics task evaluation set (Ma et al., 2019) for our development experiments. The official human judgments for translation quality of WMT 2019 were collected using reference-based direct assessment.

Since we use exactly the same correlation analysis as the official metrics shared evaluation and the 2019 version of YiSi performed consistently well among participants in WMT 2019, we only compare our results with the 2019 version of YiSi and BLEU. Our results are directly comparable with those reported in Ma et al. (2019).

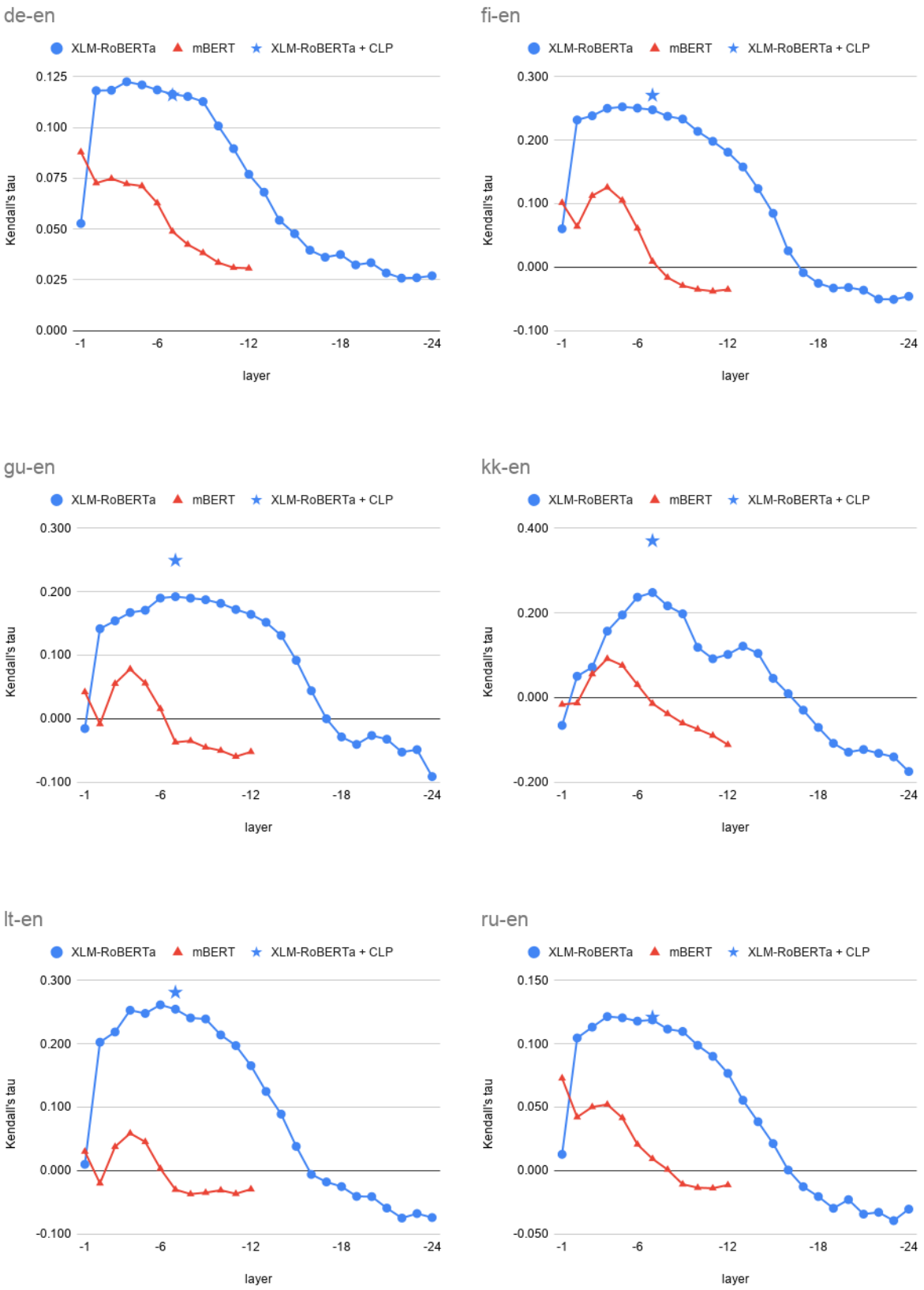

Figure 2: Segment-level Kendall's τ correlation with human direct assessment on WMT 2019 de-en, fi-en, gu-en, kk-en, lt-en and ru-en news translation test set of YiSi-2 using contextual embeddings extracted from different layers of the multilingual pretrained language models. On the x-axis, layer $-n$ means YiSi-2 based on the embeddings of the n^{th} layer, counting from the last, of XLM-RoBERTa$_{large}$ (blue circles), multilingual BERT (red triangles) and layer -7 of of XLM-RoBERTa$_{large}$ with source embeddings projected to target language space using CLP (blue star).

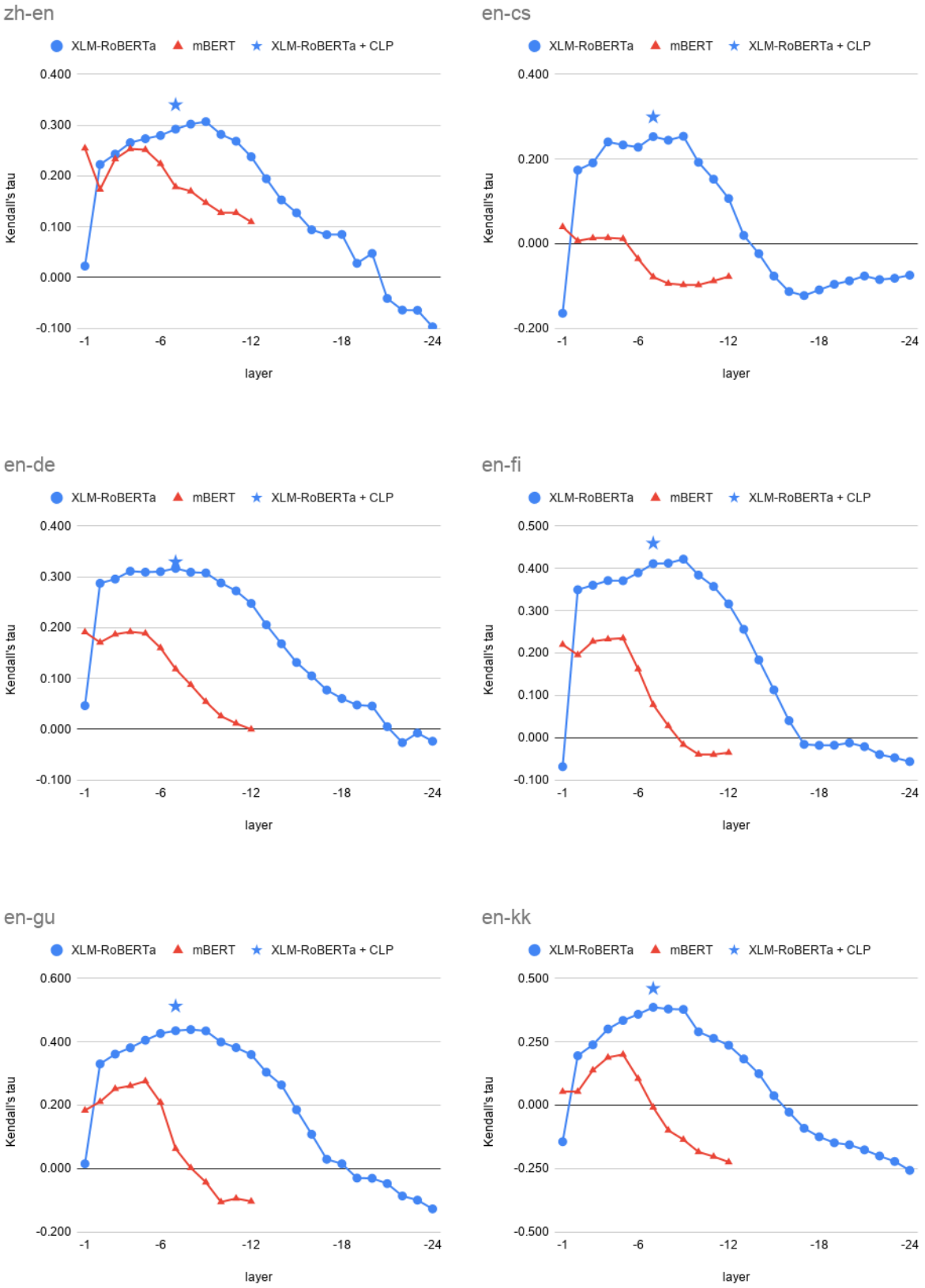

Figure 3: Segment-level Kendall's τ correlation with human direct assessment on WMT 2019 zh-en, en-cs, en-de, en-fi, en-gu and en-kk news translation test set of YiSi-2 using contextual embeddings extracted from different layers of the multilingual pretrained language models. On the x-axis, layer $-n$ means YiSi-2 based on the embeddings of the n^{th} layer, counting from the last, of XLM-RoBERTa$_{large}$ (blue circles), multilingual BERT (red triangles) and layer -7 of of XLM-RoBERTa$_{large}$ with source embeddings projected to target language space using CLP (blue star).

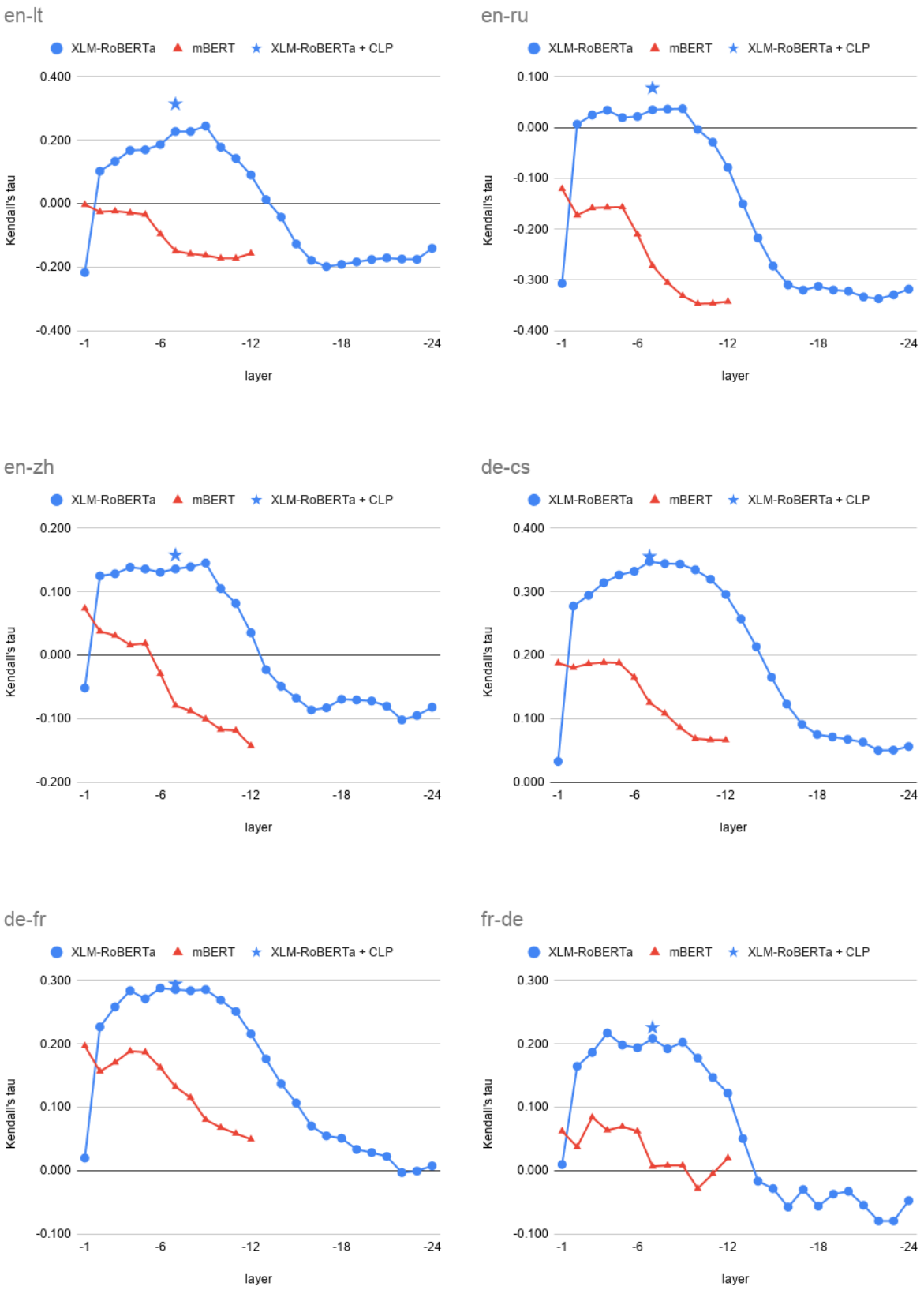

Figure 4: Segment-level Kendall's τ correlation with human direct assessment on WMT 2019 en-lt, en-ru, en-zh, de-cs, de-fr and fr-de news translation test set of YiSi-2 using contextual embeddings extracted from different layers of the multilingual pretrained language models. On the x-axis, layer $-n$ means YiSi-2 based on the embeddings of the n^{th} layer, counting from the last, of XLM-RoBERTa$_{large}$ (blue circles), multilingual BERT (red triangles) and layer -7 of of XLM-RoBERTa$_{large}$ with source embeddings projected to target language space using CLP (blue star).

Table 3: System-level Pearson's ρ correlation of metric scores with the WMT 2019 official human direct assessment judgments.

input	de	fi	gu	kk	lt	ru	zh	en	en	en	en	en	en	en	en
output	en	en	en	en	en	en	en	cs	de	fi	gu	kk	lt	ru	zh
Reference-based evaluation metric															
YiSi-1 (2019)	.949	.989	.924	.944	.981	.979	.979	.962	.991	.971	.909	.985	.963	.992	.951
YiSi-0	.902	.993	.993	.991	.927	.958	.937	.992	.985	.987	.863	.974	.974	.953	.861
BLEU	.849	.982	.834	.946	.961	.879	.899	.897	.921	.969	.737	.852	.989	.986	.901
QE as a metric															
YiSi-2 (2020)	.898	.959	.739	.981	.935	.461	.980	.773	.963	.906	.890	.977	.761	.473	.449
YiSi-2 (2019)	.796	.642	.566	.324	.442	.339	.940	.324	.924	.696	.314	.339	.055	.766	.097

Table 4: System-level Pearson's ρ correlation of metric scores with the WMT 2019 official human direct assessment judgments.

input	de	de	fr
output	cs	fr	de
Reference-based evaluation metric			
YiSi-1 (2019)	.973	.969	.908
YiSi-0	.978	.952	.820
BLEU	.941	.891	.864
QE as a metric			
YiSi-2 (2020)	.860	.853	.461
YiSi-2 (2019)	.606	.721	.530

3.1 Segment-level correlation with human judgment

In Figure 1, 2, 3 and 4, we plot the change of segment-level Kendall's τ correlation for YiSi-2 across different layers of XLM-R and multilingual BERT models. We identify a common trend, YiSi-2 using embeddings extracted from XLM-R significantly outperforms YiSi-2 using embeddings extracted from multilingual BERT. From figure 1, we see that, on average, on all translation directions, the optimal layer of representation in XLM-R for YiSi-2 is layer -7. Learning the cross-lingual linear projection matrix to transform the source embeddings into the target language subspace shows a greater improvement overall. This is our "YiSi-2 (2020)" submission to the QE as a metric task.

Table 1 and 2 show the Kendall's τ correlation with the segment-level human direct assessment relative ranking on the WMT 2019 evaluation set. YiSi-2 (2020) shows consistent and significant improvements when comparing to the previous version of YiSi-2 across all translation directions.

Although YiSi-2 (2020) still performs worse than YiSi-1, YiSi-2 (2020) correlates better with human judgment than the reference-based metric, sentBLEU, and its performances are comparable to those of the character-based YiSi variant, YiSi-0, on evaluating translation quality for most of the translation directions.

3.2 Correlation with human judgment at system level

Table 3 and 4 show the Person's ρ correlation with the system-level human direct assessment relative ranking on the WMT 2019 evaluation set.

Similar to the segment-level results, although YiSi-2 (2020) still performs significantly worse than YiSi-1, we observe significant improvements, compared to the previous version of YiSi-2, consistently across all translation directions. We also show that by replacing the multilingual BERT with XLM-R and using bilingual mappings to better align the source and target language subspaces in XLM-R, YiSi-2 (2020) correlates better with human judgment than the reference-based metric, BLEU, on evaluating translation quality for most of the translation directions.

4 Conclusion

We have presented an improved version of YiSi-2 that uses XLM-RoBERTa and a cross-lingual linear projection of the source embedding to the target language subspace to better capture the semantic representation across languages. Our results show that YiSi-2 correlates better with human judgement on evaluating translation quality than BLEU for most of the evaluation conditions. This improved version of YiSi-2 is submitted to the WMT 2020 Metrics shared task QE as a metric track. For evaluating Inuktitut↔English where one of the language (Inuktitut) is not covered by XLM-R, we build our own XLM cross-lingual language model with the parallel training data. Potential research directions definitely include improving massive multilingual pretrained language model to close the performance gap between YiSi-1 and YiSi-2 and expanding the language coverage of these models in post-hoc and unsupervised manner.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

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. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Chi-kiu Lo and Michel Simard. 2019. Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 206–215, Hong Kong, China. 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.

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: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. 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*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.