Extended Study on Using Pretrained Language Models and YiSi-1 for Machine Translation Evaluation

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

Abstract

We present an extended study on using pretrained language models and YiSi-1 for machine translation evaluation. Although the recently proposed contextual embedding based metrics, YiSi-1, significantly outperform BLEU and other metrics in correlating with human judgment on translation quality, we have yet to understand the full strength of using pretrained language models for machine translation evaluation. In this paper, we study YiSi-1's correlation with human translation quality judgment by varying three major attributes (which architecture; which intermediate layer; whether it is monolingual or multilingual) of the pretrained language models. Results of the study show further improvements over YiSi-1 on the WMT 2019 Metrics shared task. We also describe the pretrained language model we trained for evaluating Inuktitut machine translation output.

1 Introduction

Recent research on large-scale evaluation of automatic machine translation (MT) evaluation metrics (Ma et al., 2018, 2019; Mathur et al., 2020) showed that the newly proposed contextual embedding based metrics, like YiSi-1, BERTscore (Zhang et al., 2020) and ESIM (Mathur et al., 2019), significantly outperform BLEU (Papineni et al., 2002) and other metrics in correlating with human judgment on translation quality. YiSi-1 and BERTscore use contextual embeddings extracted from the pretrained language model, Devlin et al. (2018), as-is without further fine-tuning or fitting to existing labeled data predictions. Although fine-tuning the pretrained language models for specific downstream tasks show improvements in many cases, using the pretrained language models without fine-tuning makes the MT evaluation metrics more portable to languages without labeled data and the resulted metric scores are comparable to each other over time across systems. Thus, instead of spending efforts into fine-tuning the pretrained language models for MT evaluation, we focus on finding the most optimal way (which architecture; which intermediate layer; whether it is a monolingual or multilingual model) to utilize them as-is.

Zhang et al. (2020) investigated into a few aspects (architecture and layer) of the use of contextual embeddings in text generation evaluation. They evaluated several model architectures of different sizes, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and XLM (Lample and Conneau, 2019). As more pretrained language models that cover more languages are released since then, we extend the study on YiSi-1 to include more pretrained language models and also compare the effect of using monolingual pretrained models versus multilingual pretrained models.

In this paper, we experiment on different settings of YiSi-1 in the WMT 2019 metrics shared task, integrating it with different transformer-based (Vaswani et al., 2017) contextual language models in both monolingual or multilingual, such as BERT (Devlin et al., 2018), ALBERT (Lan et al., 2020), BART (Lewis et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020) and XLNET (Yang et al., 2019), using different intermediate layers. We show that YiSi-1's correlation with human judgment on translation quality is improved by using the results of this study.

2 YiSi

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

Proceedings of the 5th Conference on Machine Translation (WMT), pages 895–902
Online, November 19–20, 2020.

translation and human references by aggregating weighted distributional (lexical) semantic similarities, and optionally incorporating shallow semantic structures. YiSi-0 is the degenerate version of YiSi-1 that is ready-to-deploy to any languages by using longest common character substring, instead of cosine similarity of contextual embeddings, to measure lexical similarity.

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, and optionally incorporating shallow semantic structures. Improvements in YiSi-2 for WMT 2020 metrics shared task is detailed in (Lo and Larkin, 2020).

2.1 Pretrained Language Models

YiSi-1 relies on a language representation to evaluate the lexical semantic similarity between the reference translation and the MT output. In WMT 2019 metrics shared task, it used pretrained 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. Monolingual BERT pretrained model for English and Chinese and multilingual BERT pretrained that covers the 104 largest languages in Wikipedia were public released in 2019.

2.1.1 Monolingual models

Monolingual BERT in other languages After the success of using monolingual BERT models for downstream NLP tasks in Chinese and English, a number of monolingual BERT models in other languages have been publicly released, such as German (Chan et al., 2019), Finnish (Virtanen et al., 2019), French (Martin et al., 2020), Japanese Inui Laboratory (2019), Dutch (de Vries et al., 2019). In our experiments, we compare the performance of YiSi-1 using these monolingual models against that using multilingual language models.

Other monolingual models in English A number of modifications to BERT have been proposed to optimize the pretrained language models. Lan et al. (2020) proposed ALBERT to reduce the amount of parameters in BERT for lower memory consumption and faster training speed. BART (Lewis et al., 2019) is effective when fine tuned for text generation tasks. RoBERTa (Liu et al., 2019) is a more robustly trained version of BERT where the key hyperparameters are empirically chosen. XLNET (Yang et al., 2019) an autoregressive model that maximizes the expected likelihood over all permutations of the input sequence factorization order. We use these models in YiSi-1 for correlation analysis with human judgment on translation quality.

2.1.2 Multilingual models

In addition to multilingual BERT used in Lo (2019), 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 lexical semantic similarity.

2.2 Inuktitut-English Cross-lingual Language Model

Since Inuktitut is not covered by any publicly released pretrained language model, 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 Model size and intermediate layers

In this study, we are interested in achieving the best performance using the pretrained language models. Thus, if different sizes of the same model architecture are released, we only evaluate the largest one in out experiment. 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

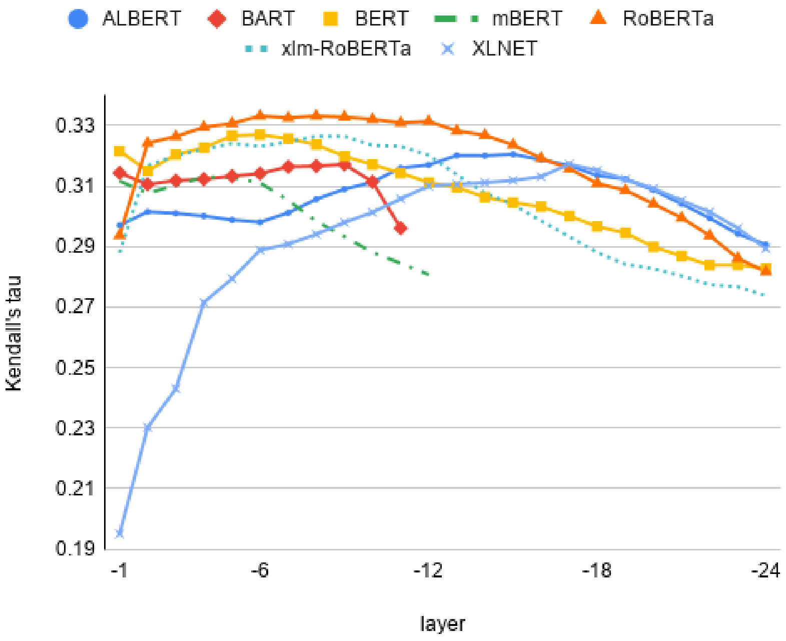


Figure 1: Average segment-level Kendall's τ correlation with human direct assessment on WMT19 *-en news translation test set of YiSi-1 using different pretrained language representation models. Solid lines represent the use of pretrained monolingual models. Dotted line represents the use of pretrained XLM-R and dashed line represents the use of pretrained multilingual BERT.

best represents the semantic space of the language.

3 Experiments and Results

We use WMT 2019 metrics shared 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.

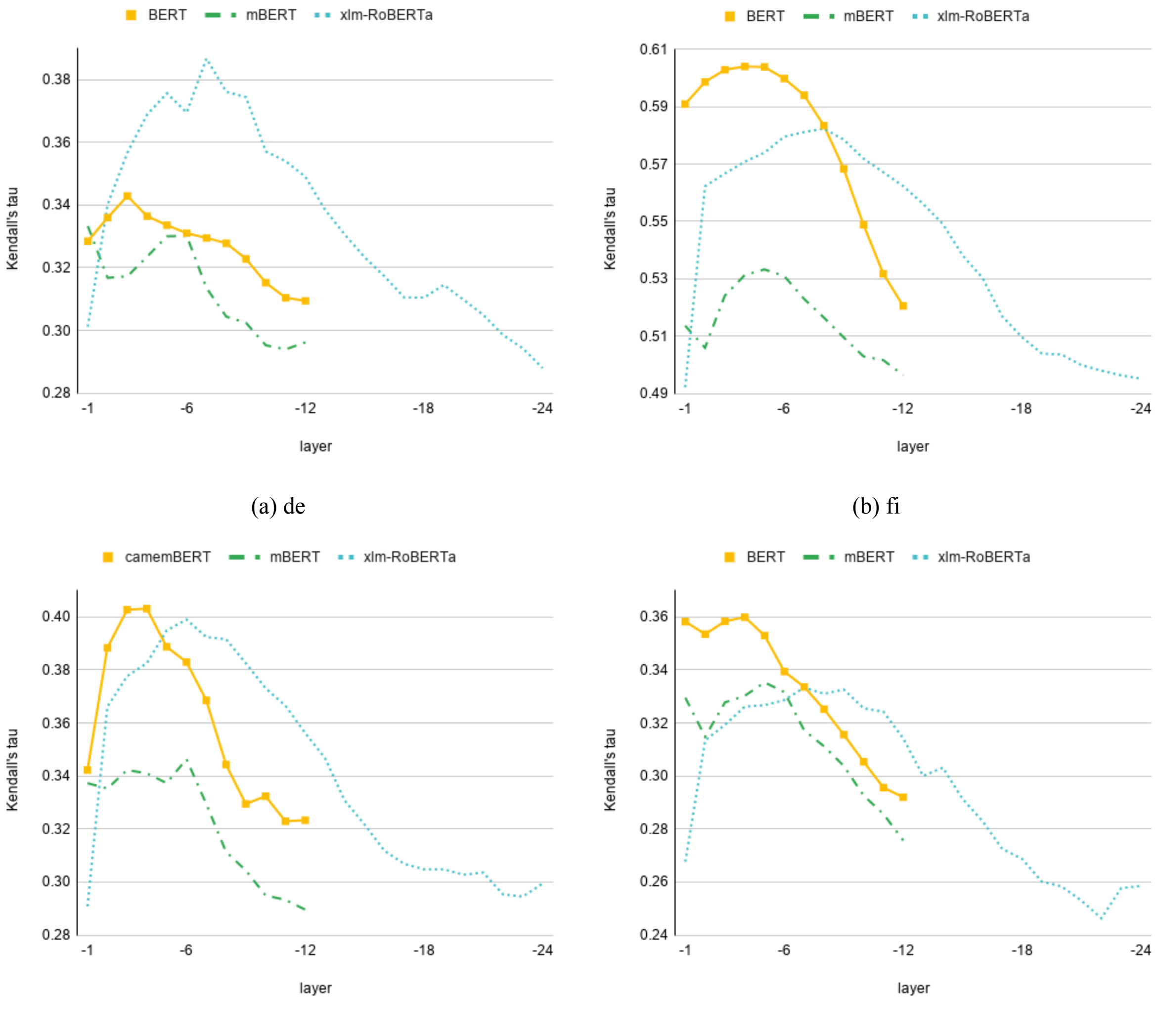
Since we use exactly the same correlation analysis as the official metrics shared evaluation and the 2019 version of YiSi performed consistently well among participants in WMT 2019, we only compare our results with the 2019 version of YiSi and BLEU. Our results are directly comparable with those reported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

3.1 Architectures of monolingual English models

In Figure 1, we plot the change of segment-level Kendall's τ correlation of YiSi-1 across different layers of the monolingual and multilingual pretrained language models for evaluating English MT output. We see that YiSi-1 using RoBERTa$_{large}$ at layer -6 achieved the correlation with human translation quality judgment; marginally better than that using BERT$_{large}$ and XLM-RoBERTa$_{large}$. Therefore, in WMT 2020 metrics shared task *-English MT output evaluation, we submit YiSi-1 scores based on embeddings extracted from the layer -6 of RoBERTa$_{large}$.

3.2 Monolingual models vs. multilingual models

In Figure 1 and 2, we identify a common pattern that for evaluating English, Finnish, French and Chinese, using monolingual models (RoBERTa for English, CamemBERT for French and BERT for Finnish and Chinese) in YiSi-1 achieved the best correlation with human translation quality judgment. The only exception is using German BERT in YiSi-1 for evaluating German MT output where YiSi-1 using XLM-RoBERTa significantly outperforms that using German BERT. One of the possible reasons is that there is a domain mismatch between the training data of the German BERT and the MT output in the evaluation. The data used in the German BERT included 20% of legal documents while the MT output of WMT 2019 metrics shared evaluation set belongs to the news domain. Since YiSi-1 using monolingual BERT model usually outperforms that using multilingual pretrained



(a) de

(b) fi

(c) fr

(d) zh

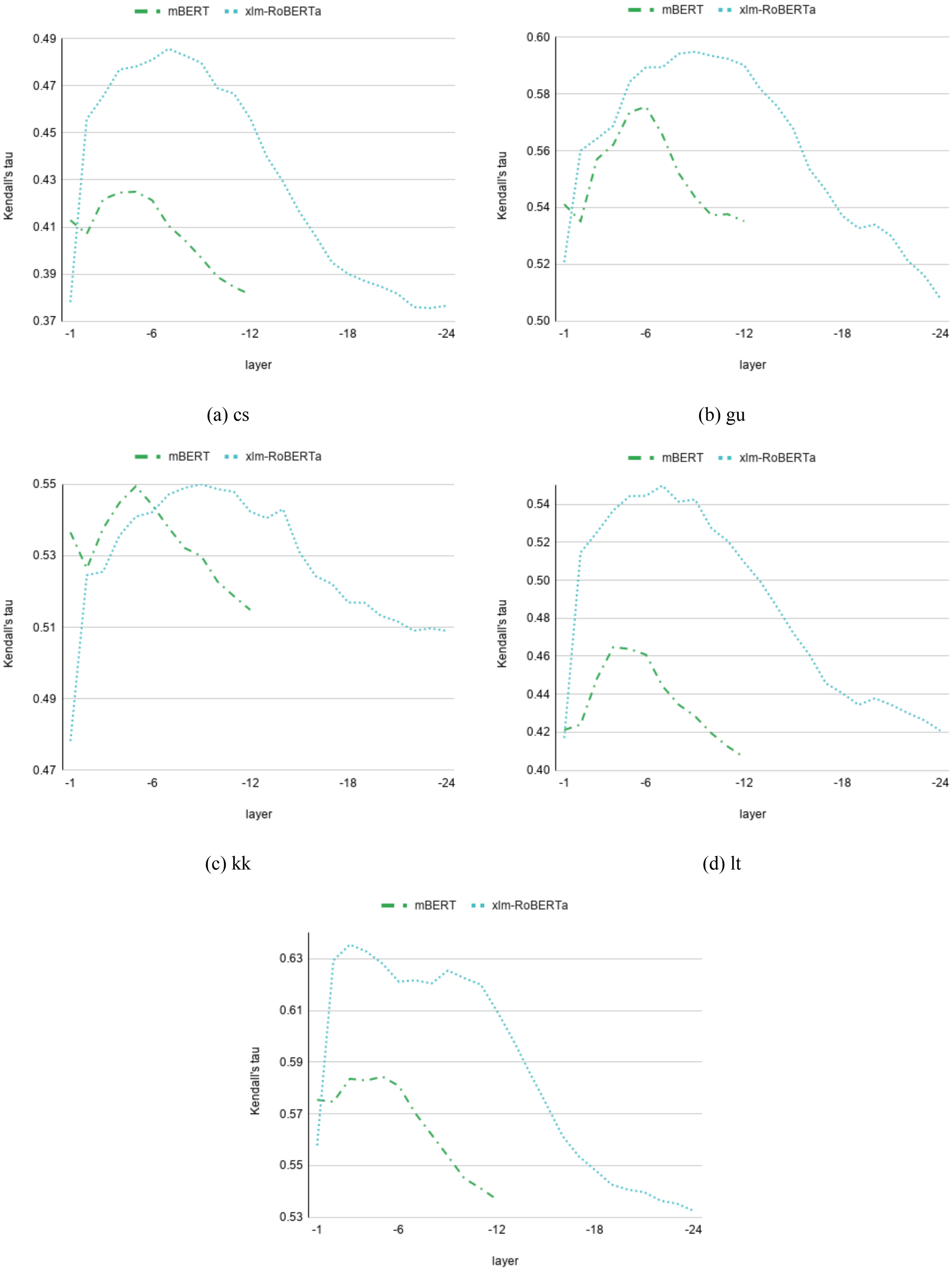
Figure 2: Average segment-level Kendall's τ correlation with human direct assessment on WMT19 (a) *-de, (b) en-fi, (c) de-fr and (d)en-zh news translation test set of YiSi-1 using different pretrained language representation models. Solid lines represent the use of pretrained monolingual models. Dotted line represents the use of pretrained XLM-R and dashed line represents the use of pretrained multilingual BERT.

language model, we believe that for evaluating MT output in WMT 2020 metrics shared task, using YiSi-1 with the monolingual BERT (while available, i.e. CamemBERT for French, BERT for Japanese and Chinese) would be a better model choice.

Another common pattern we see is that YiSi-1 using the monolingual BERT$_{base}$ model usually achieved the best correlation with human translation quality judgment at layer -4. Therefore, in WMT 2020 metrics shared task *-Chinese/French/Japanese MT output evaluation, we submit YiSi-1 scores based on embeddings extracted from the layer -4 of the corresponding monolingual BERT model.

3.3 Multilingual BERT vs. XLM-RoBERTa

In Figure 3, we plot the change of segment-level Kendall's τ correlation for YiSi-1 across different layers of XLM-R and multilingual BERT models for evaluating English-Czech/Gujarati/Kazakh/Lithuanian/Russian.
We identify a common trend, YiSi-1 using embeddings extracted from XLM-RoBERTa significantly outperforms YiSi-1 using embeddings extracted from multilingual BERT, except for evaluating Kazakh MT output where the gains of using XLM-RoBERTa is marginal. On average in all translation directions, the optimal layer of representation in XLM-R for YiSi-1 is layer -7.



(a) cs

(b) gu

(c) kk

(d) lt

(e) ru

Figure 3: Average segment-level Kendall's τ correlation with human direct assessment on WMT19 (a) *-cs, (b) en-gu, (c) en-kk, (d) en-lt and (e) en-ru news translation test set of YiSi-1 using different pretrained language representation models. Dotted line represents the use of pretrained XLM-R and dashed line represents the use of pretrained multilingual BERT.

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

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
YiSi-1 (2020)	**.172**	**.354**	**.328**	.425	**.385**	**.230**	**.438**	**.544**	**.384**	**.604**	**.589**	**.547**	**.550**	**.622**	**.360**
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

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

input	de	de	fr
output	cs	fr	de
YiSi-1 (2020)	**.427**	**.403**	**.389**
YiSi-1 (2019)	.376	.349	.310
YiSi-0	.331	.296	.277

4 Improvements over previous version of YiSi-1

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-1 (2020) shows consistent and significant improvements when comparing to the previous version of YiSi-1 across all translation directions.

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. Although the improvements at system-level correlation is less consistent across different translation directions, YiSi-1 (2020) outperforms YiSi-1(2019) in the evaluation of two-third of all the tested translation directions.

5 Conclusion

We have presented an extend study of the pretrained language models used in YiSi-1 for machine translation evaluation. From this study, we conclude that for the best performance of YiSi-1: 1) when evaluating MT output in English, it is recommended to use the contextual embeddings extracted from layer -6 of RoBERTa$_{large}$; 2) when evaluating MT output in languages where monolingual pretrained model in the same or general domain is available, it is recommended to use the contextual embeddings extracted from those models; and finally 3) when evaluating MT output in languages only covered by multilingual pretrained language models, it is recommended to use the contextual embeddings extracted from layer -7 of XLM-RoBERTa.

This improved version of YiSi-1 is submitted to the WMT 2020 metrics shared task. For evaluating Inuktitut↔English where one of the language (Inuktitut) is not covered by any released pretrained langauge model, we build our own XLM cross-lingual language model with the parallel training data.

References

Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. 2019. Open sourcing german bert.

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.

Tohoku University Inui Laboratory. 2019. Pretrained Japanese BERT models. https://github.com/cl-tohoku/bert-japanese.

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. Crosslingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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

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
YiSi-1 (2020)	**.953**	.987	**.998**	**.991**	.967	.929	**.986**	.971	**.993**	.979	**.945**	**.991**	**.979**	.980	.942
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

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

input	de	de	fr
output	cs	fr	de
YiSi-1 (2020)	**.981**	.953	**.924**
YiSi-1 (2019)	.973	**.969**	.908
YiSi-0	.978	.952	.820

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 and Samuel Larkin. 2020. Machine translation reference-less evaluation using yisi-2with bilingual mappings of massive multilingual language model. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

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.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, 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 *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, Pennsylvania.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

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*.