Translation Quality Estimation by Jointly Learning to Score and Rank

Jingyi Zhang[1] **and Josef van Genabith**[1,2]
[1]German Research Center for Artificial Intelligence (DFKI),
Saarbrucken, Germany
[2]Department of Language Science and Technology,
Saarland University, Germany
Jingyi.Zhang@dfki.de,Josef.Van_Genabith@dfki.de

Abstract

The translation quality estimation (QE) task, particularly the *QE as a Metric* task, aims to evaluate the general quality of a translation based on the translation and the source sentence without using reference translations. Supervised learning of this QE task requires human evaluation of translation quality as training data. Human evaluation of translation quality can be performed in different ways, including assigning an absolute score to a translation or ranking different translations. In order to make use of different types of human evaluation data for supervised learning, we present a multi-task learning QE model that jointly learns two tasks: score a translation and rank two translations. Our QE model exploits cross-lingual sentence embeddings from pretrained multilingual language models. We obtain new state-of-the-art results on the WMT 2019 *QE as a Metric* task and outperform sentBLEU on the WMT 2019 Metrics task.

1 Introduction

The translation quality estimation (QE) task (Fonseca et al., 2019) aims to evaluate the quality of a translation based on the translation and the source sentence without using reference translations. The QE task includes *word-level QE*, *sentence-level QE*, *document-level QE* and *QE as a Metric* tasks. The *QE as a Metric* task requires QE models to score a translation on the sentence level similar to the *sentence-level QE* task, but these two tasks are different as the goal of the *sentence-level QE* task (Martins et al., 2017) is to predict the percentage of edits needed to fix the translation for post-editing purposes while the goal of the *QE as a Metric* task is to estimate the general quality of the translation like machine translation (MT) evaluation metrics, such as BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2014), except without using reference translations.

Supervised learning of the *QE as a Metric* task requires human evaluation of translation quality as training data. Human evaluation of translation quality is generally very costly and can be performed in different ways, such as Direct Assessment (DA: requiring human assessors to assign an absolute score to a translation) (Barrault et al., 2019; Graham et al., 2013, 2014, 2017) or Relative Ranking (RR: requiring human assessors to rank different translations) (Bojar et al., 2015). Since the *QE as a Metric* task requires QE models to assign an absolute score to a translation, DA human evaluation data can be straightforwardly used as training data for the *QE as a Metric* task. In order to also make use of the RR human evaluation data, we propose a multi-task learning QE model that jointly learns two tasks, score a translation and rank two translations. Multi-task learning of these two closely related tasks enables us to use both DA and RR human evaluation data for training the QE model and improve performance compared to learning these two tasks separately. Our model performs translation quality estimation based on cross-lingual sentence embeddings from pretrained multilingual language models (Devlin et al., 2019; Conneau et al., 2019) and does not need reference translations. We obtain new state-of-the-art results on the WMT 2019 *QE as a Metric* task and outperform sentBLEU on the WMT 2019 Metrics task (Ma et al., 2019).

A number of previous works also used sentence embeddings for evaluating translation quality (Shimanaka et al., 2018; Guzmán et al., 2015; Gupta et al., 2015). However, Shimanaka et al. (2018); Gupta et al. (2015)'s models only learn to score a translation and Guzmán et al. (2015)'s model only learns to rank two translations while our model jointly learns to score a translation and rank two translations in order to make use of different types of human evaluation data for model training. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2592–2598,
November 16–20, 2020.

addition, Shimanaka et al. (2018); Guzmán et al. (2015); Gupta et al. (2015)'s models use the reference translation for evaluating translation quality while our QE model does not require reference translations. There are existing QE models (Lo, 2019; Yankovskaya et al., 2019) that do not need the reference translation and perform translation quality estimation based on cross-lingual word/sentence embeddings, but these QE models give relatively poor and unstable results for different language pairs (Ma et al., 2019) while our QE model achieves more robust and better results. In addition, Lo (2019); Yankovskaya et al. (2019)'s QE models only score a translation while our QE model jointly learns to score a translation and rank two translations via multi-task learning.

2 Our Approach

We propose a multi-task learning QE model that jointly learns two tasks: score a translation and rank two translations. Our QE model is based on cross-lingual sentence embeddings from multilingual BERT (M-BERT) (Devlin et al., 2019; Reimers and Gurevych, 2019). To compute the sentence embedding for a given sentence, we feed this sentence into M-BERT and then perform MEAN pooling over the output of M-BERT to obtain fixed-size sentence embedding. We fine-tune M-BERT for the QE tasks.

The scoring task To score a translation t given the source sentence s, we use the cosine similarity between the source sentence embedding $\vec{s}$ and the target sentence embedding $\vec{t}$ as the score of the translation.[1] Equation 1 gives the loss function for the scoring task, where Y_{human} ($0 \leq Y_{human} \leq 1$) is the normalized DA score of the translation assigned by human assessors.

$$L_{score} = \left(cos\_sim\left(\vec{s}, \vec{t}\right) - Y_{human}\right)^2 \tag{1}$$

The ranking task To rank two translations t_1 and t_2 given the source sentence s, we compute Euclidean distances between source and target sentence embeddings $Euc\_dis\left(\vec{s}, \vec{t_1}\right)$ and $Euc\_dis\left(\vec{s}, \vec{t_2}\right)$. The translation that has a smaller Euclidean distance with the source is predicted to be the better translation.[2] Equation 2 gives the loss function for the ranking task, where t_b is the better translation and t_w is the worse translation according to the human ranking. By minimizing L_{rank}, we want $Euc\_dis\left(\vec{s}, \vec{t_b}\right)$ to be at least ε less than $Euc\_dis\left(\vec{s}, \vec{t_w}\right)$. We tuned ε on the development set and finally set $\varepsilon = 1$ in our experiments.[3]

$$L_{rank} = ReLU\left(Euc\_dis\left(\vec{s}, \vec{t_b}\right) - Euc\_dis\left(\vec{s}, \vec{t_w}\right) + \varepsilon\right) \tag{2}$$

Multi-task learning We train our QE model to jointly learn the scoring task and the ranking task via multi-task learning. Each training step includes two training batches: one training batch for the scoring task and one training batch for the ranking task. Direct Assessment (DA) human evaluation data which requires human assessors to assign an absolute score to a translation is used as training data for the scoring task; Relative Ranking (RR) human evaluation data which requires human assessors to rank different translations is used as training data for the ranking task.

The main advantage of multi-task learning for these two closely related tasks is that we can use both DA and RR human evaluation data for training the QE model and improve performance compared to learning these two tasks separately. We test our method on the WMT 2019 *QE as a Metric* task which requires QE models to assign an absolute score to a translation. We show that, on the *QE as a Metric* task, our multi-task learning method can achieve significantly better results compared to only training the QE model to learn the scoring task with DA human evaluation data.

3 Experiments

3.1 Settings

We evaluated the performance of our QE model on the WMT 2019 *QE as a Metric* task[4]. For model training, we used human evaluation data of WMT NEWS translation tasks. Since 2016, WMT performed human evaluation for submissions of NEWS translation tasks via Direct Assessment

[1] Instead of using cosine similarity as the translation score, we also tried to use a 1-layer feed forward network (FFN) to compute the translation score and use $\left[\vec{s}, \vec{t}\right]$ (concatenation of the source and target sentence embeddings) as the input of the FFN. Compared to cosine similarity, the FFN achieved slightly worse results on the test sets.

[2] Note that we can also use cosine similarity instead of Euclidean distance to rank the two translations in the ranking task. However, we find that using Euclidean distance for the ranking task achieved better QE results in our experiments.

[3] For tuning ε, we tried $\varepsilon = 0.2, 0.5, 1, 2, 3$ and found $\varepsilon = 1$ achieved the highest Pearson correlation on the development set.

[4] http://ufallab.ms.mff.cuni.cz/ bojar/wmt19-metrics-task-package.tgz

			de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
Metrics	sentBLEU		0.056	0.233	0.188	0.377	0.262	0.125	0.323
	YiSi-1_srl (Lo, 2019)		0.199	0.346	0.306	0.442	0.380	0.222	0.431
QE as a Metric	UNI+ (Ma et al., 2019)		0.015	0.211	-	-	-	0.089	-
	YiSi-2 (Lo, 2019)		0.068	0.126	-0.001	0.096	0.075	0.053	0.253
	Ours	M_NO	0.021	0.102	0.001	-0.026	-0.001	0.072	0.226
		M_DA	0.075	0.261	0.220	0.285	0.284	0.109	0.304
		M_MU	0.082	0.260	0.246	**0.329**	0.289	0.118	0.319
		X_NO	0.022	0.208	0.104	0.067	0.151	0.073	0.251
		X_DA	0.081	0.286	0.215	0.247	0.287	0.106	0.299
		X_MU	**0.101**	**0.294**	**0.256**	0.316	**0.311**	**0.125**	**0.335**

Table 1: Segment-level metric results for to-English language pairs in newstest2019: Kendalls Tau formulation of segment-level metric scores with DA scores. **Bold**: best results for the *QE as a Metric* task.

			en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
Metrics	sentBLEU		0.367	0.248	0.396	0.465	0.392	0.334	0.469	0.270
	YiSi-1 (Lo, 2019)		0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355
QE as a Metric	UNI (Ma et al., 2019)		0.060	0.129	0.351	-	-	-	0.226	-
	YiSi-2 (Lo, 2019)		0.069	0.212	0.239	0.147	0.187	0.003	-0.155	0.044
	YiSi-2_srl (Lo, 2019)		-	0.236	-	-	-	-	-	0.034
	Ours	M_NO	0.056	0.171	0.251	0.214	0.239	0.076	-0.094	0.083
		M_DA	0.376	0.286	0.465	0.383	0.438	0.406	0.140	0.277
		M_MU	0.383	0.310	0.481	0.428	0.463	0.415	0.152	0.262
		X_NO	-0.108	0.035	0.161	0.113	0.147	-0.100	-0.171	-0.096
		X_DA	0.433	0.264	0.523	0.426	0.398	0.498	0.205	0.300
		X_MU	**0.502**	**0.339**	**0.556**	**0.493**	**0.485**	**0.546**	**0.228**	**0.317**

Table 2: Segment-level metric results for out-of-English language pairs in newstest2019: Kendalls Tau formulation of segment-level metric scores with DA scores. **Bold**: best results for the *QE as a Metric* task.

		Number
Direct Assessment	WMT 2016	141,905
	WMT 2018	228,409
Relative Ranking	WMT 2014	254,000
	WMT 2015	258,749

Table 3: Numbers of training examples.

			de-cs	de-fr	fr-de
Metrics	sentBLEU		0.203	0.235	0.179
	YiSi-1 (Lo, 2019)		0.376	0.349	0.310
QE as a Metric	YiSi-2 (Lo, 2019)		0.199	0.186	0.066
	Ours	M_NO	0.145	0.172	0.051
		M_DA	0.199	0.269	0.127
		M_MU	0.240	0.285	0.149
		X_NO	0.076	0.078	-0.005
		X_DA	0.266	0.204	**0.174**
		X_MU	**0.314**	**0.333**	0.123

Table 4: Segment-level metric results for language pairs not involving English in newstest2019: Kendalls Tau formulation of segment-level metric scores with DA scores. **Bold**: best results for the *QE as a Metric* task.

(DA) (Barrault et al., 2019; Graham et al., 2013, 2014, 2017). Direct Assessment requires human assessors to assign an absolute score (between 0 and 100) to a translation based on general translation quality. We normalize DA scores to $[0, 1]$ for training our model for the scoring task. Before 2016, WMT performed human evaluation for NEWS translation tasks via Relative Ranking (Bojar et al., 2015). Relative Ranking requires human assessors to rank different translations based on general translation quality. The rank results of any two translations that are not tied can be used to train our model for the ranking task. Table 3 gives numbers of training examples for our model.

We trained our model via multi-task learning.[5] Each training step includes one training batch from DA data and one training batch from RR data. Each training batch contains 8 training examples. We set the learning rate to 2e-7 and the number of training epochs to 20. The DA data (4,787 human scores) of WMT 2017 NEWS task was used as the development set. The development set does not include any RR data because the final goal of our model is to assign an absolute score to each translation as required by the *QE as a Metric* task. We evaluated our model on the development set after every 1000 training batches and saved the checkpoint that achieved the highest Pearson correlation on the development set.

As described in the previous section, our model

[5] Code for reproducing our results can be found here https://github.com/jingyiz/sentence-transformers

is based on cross-lingual sentence embeddings from M-BERT (Devlin et al., 2019). Other than M-BERT, we also tested another pretrained multilingual language model XLM-RoBERTa (Conneau et al., 2019) which achieves better results than M-BERT on various cross-lingual tasks. Finally, we trained six QE models for comparison,

1. QE_M-BERT_NO-TRAIN (M_NO)
2. QE_M-BERT_DA-ONLY (M_DA)
3. QE_M-BERT_MULTI-TASK (M_MU)
4. QE_XLM-RoBERTa_NO-TRAIN (X_NO)
5. QE_XLM-RoBERTa_DA-ONLY (X_DA)
6. QE_XLM-RoBERTa_MULTI-TASK (X_MU)

where models 1, 2 and 3 use M-BERT for sentence embedding; models 4, 5 and 6 use XLM-RoBERTa for sentence embedding; NO-TRAIN means we do not fine-tune M-BERT (XLM-RoBERTa) for the QE tasks and simply use the pretrained model for sentence embedding; DA-ONLY means we only train the QE model to learn the scoring task with DA data; MULTI-TASK means we train the QE model with both DA and RR data to jointly learn the scoring task and the ranking task via multi-task learning.

3.2 Segment-Level Results

Tables 1, 2 and 4 give results of our models and the winning systems of the WMT 2019 *QE as a Metric* task (segment-level). We also show results of sentBLEU and the winning systems of the WMT 2019 Metrics task. Compared to the *QE as a Metric* task, the Metrics task allows the usage of the reference translation for translation quality estimation.

In Tables 1, 2 and 4, M_NO and X_NO had bad results, which shows that pretrained multilingual language models without fine-tuning do not perform well on the QE task; X_MU (M_MU) generally outperformed X_DA (M_DA), which shows that training the QE model with both DA and RR data to jointly learn the scoring and ranking tasks via multi-task learning can achieve better quality estimation results than only training the QE model to learn the scoring task with the DA data. Results also show that XLM-RoBERTa outperformed M-BERT for the QE task. Our best model X_MU[6] achieved new state-of-the-art results for all language pairs on WMT 2019 *QE as a Metric* task and outperformed sentBLEU for 14 out of 18 language pairs on WMT 2019 Metrics task. Particularly, among all the languages in the test sets, Gujarati (gu) and Lithuanian (lt) do not occur in the training data of the QE task. Nevertheless, our model still got good results (outperforming sentBLEU) for gu-en, lt-en, en-gu and en-lt tasks. In contrast, UNI (Ma et al., 2019), UNI+ (Ma et al., 2019), YiSi-2 (Lo, 2019) and YiSi-2_srl (Lo, 2019) give significantly worse and unstable results for different language pairs.

	MAX	CLS	MEAN
M_MU	0.641	0.646	0.648
X_MU	0.647	0.667	0.694

Table 5: Results (segment-level Pearson correlation) on the development set by using different pooling strategies for sentence embedding.

	cos_sim	Euc_dis
M_DA	0.633	
X_DA	0.651	
M_MU	0.647	0.648
X_MU	0.690	0.694

Table 6: Results (segment-level Pearson correlation) on the development set by using different loss functions for the ranking task.

Pooling Strategy Other than performing MEAN pooling to obtain sentence embeddings, we also tested MAX pooling or using the CLS token representation as the sentence embedding (Devlin et al., 2019). Results in Table 5 show that MEAN pooling achieved the best results for the QE task.

Cosine Similarity for the Ranking Task We also tried to use cosine similarity instead of Euclidean distance for ranking the two translations in the ranking task. That is we used Equation 3 instead of Equation 2 as the loss function for the ranking task. ω was tuned to be 0.1.[7] Results are shown in Table 6. Our multi-task learning QE model X_MU (M_MU) achieved better results when using Euc_dis for the ranking task compared to using cos_sim for the ranking task; X_MU (M_MU) always outperformed X_DA (M_DA) no matter Euc_dis or cos_sim was used for the ranking task.

$$L_{rank} = ReLU\left(cos\_sim\left(\vec{s}, \vec{t_w}\right) - cos\_sim\left(\vec{s}, \vec{t_b}\right) + \omega\right) \tag{3}$$

3.3 System-Level Results

Tables 7, 8 and 9 give results of our best model (X_MU) and the winning systems of the WMT

[6] The training process of X_MU takes 3 days with 1 GPU.

[7] For tuning ω, we tried $\omega = 0.02, 0.05, 0.1, 0.2, 0.3$ and found $\omega = 0.1$ achieved the highest Pearson correlation on the development set.

		de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
Metrics	BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899
	YiSi-1_srl (Lo, 2019)	0.950	0.989	0.918	0.994	0.983	0.978	0.977
QE as a Metric	IBM1-morpheme (Popović, 2012)	-0.345	0.740	-	-	0.487	-	-
	UNI (Ma et al., 2019)	0.846	**0.930**	-	-	-	0.805	-
	UNI+ (Ma et al., 2019)	**0.850**	0.924	-	-	-	**0.808**	-
	YiSi-2 (Lo, 2019)	0.796	0.642	-0.566	-0.324	0.442	-0.339	0.940
	YiSi-2_srl (Lo, 2019)	0.804	-	-	-	-	-	0.947
	Ours (X_MU)	0.841	0.841	**-0.288**	**0.034**	**0.698**	-0.214	**0.965**

Table 7: Pearson correlation of to-English system-level metrics with DA human assessment in newstest2019. Best results for the *QE as a Metric* task are highlighted in bold.

		en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
Metrics	BLEU	0.897	0.921	0.969	0.737	0.852	0.989	0.986	0.901
	YiSi-1 (Lo, 2019)	0.962	0.991	0.971	0.909	0.985	0.963	0.992	0.951
QE as a Metric	UNI (Ma et al., 2019)	0.028	0.841	**0.907**	-	-	-	**0.919**	-
	YiSi-2 (Lo, 2019)	0.324	0.924	0.696	0.314	0.339	0.055	-0.766	-0.097
	YiSi-2_srl (Lo, 2019)	-	0.936	-	-	-	-	-	-0.118
	Ours (X_MU)	**0.586**	**0.942**	0.824	**0.549**	**0.911**	**0.499**	-0.700	**0.151**

Table 8: Pearson correlation of out-of-English system-level metrics with DA human assessment in newstest2019. Best results for the *QE as a Metric* task are highlighted in bold.

		de-cs	de-fr	fr-de
Metrics	BLEU	0.941	0.891	0.864
	YiSi-1 (Lo, 2019)	0.973	0.969	0.908
QE as a Metric	IBM1-pos4gram (Popović, 2012)	-	0.085	-0.478
	YiSi-2 (Lo, 2019)	0.606	0.721	-0.530
	Ours (X_MU)	**0.660**	**0.782**	**-0.371**

Table 9: Pearson correlation of system-level metrics for language pairs not involving English with DA human assessment in newstest2019. Best results for the *QE as a Metric* task are highlighted in bold.

2019 *QE as a Metric* task (system-level). We also show results of BLEU and the winning systems of the WMT 2019 Metrics task. For system-level evaluation, metrics which can use the reference translations for quality estimation, such as BLEU, generally achieved consistently high correlation with human evaluation for all language pairs. In contrast, QE models (including our QE model and submitted systems for the *QE as a Metric* task) are not allowed to use the reference translations for quality estimation and tend to generate more unstable results: high correlation with human evaluation for some language pairs but very low or even negative Pearson correlation with human evaluation for some other language pairs. For example, our QE model beat BLEU for zh-en, en-de and en-kk directions but got negative Pearson correlation with human evaluation for gu-en, ru-en, en-ru and fr-de directions. Among all QE models which do not use the reference translations, our model achieved the highest Pearson correlation with human evaluation for 13 out of 18 language pairs. Compared to Tables 1, 2 and 4, our model tends to produce more unstable results for system-level evaluation than segment-level evaluation, likely because the segment-level correlation is computed using about 2000 segments for a language pair while the system-level correlation is computed using only about 10 systems for a language pair, therefore the segment-level correlation is more stable.

4 Conclusion

This paper presents a multi-task leaning QE model that jointly learns two tasks, score a translation and rank two translations. The scoring and ranking results performed by human assessors can be used as training data for learning the scoring and ranking tasks respectively. Multi-task learning of these two closely related tasks enables us to make use of both types of human evaluation data for model training and improve performance compared to learning these two tasks separately. Our model obtains new state-of-the-art results on the WMT 2019 *QE as a Metric* task and outperforms sentBLEU on the WMT 2019 Metrics task.

Acknowledgments

This work is supported by the German Federal Ministry of Education and Research (BMBF) under the funding code 01IW17001 (Deeplee). Responsibility for the content of this publication is with the authors.

References

Loc Barrault, Ondej Bojar, Marta R. Costa-juss, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Mller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. 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. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

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.

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.

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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 805–814, Beijing, China. Association for Computational Linguistics.

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.

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.

André F. T. Martins, Marcin Junczys-Dowmunt, Fabio N. Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

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.

Maja Popović. 2012. Morpheme- and POS-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137, Montréal, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Metric for automatic machine translation evaluation based on universal sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 106–111, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy. Association for Computational Linguistics.