Quality Estimation and Translation Metrics via Pre-trained Word and Sentence Embeddings

Lisa Yankovskaya **Andre Tättar** **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{lisa.yankovskaya,andre.tattar,fishel}@ut.ee

Abstract

We propose the use of pre-trained embeddings as features of a regression model for sentence-level quality estimation of machine translation. In our work we combine freely available BERT and LASER multilingual embeddings to train a neural-based regression model. In the second proposed method we use as an input features not only pre-trained embeddings, but also log probability of any machine translation (MT) system. Both methods are applied to several language pairs and are evaluated both as a classical quality estimation system (predicting the HTER score) as well as an MT metric (predicting human judgements of translation quality).

1 Introduction

Quality estimation (Blatz et al., 2004; Specia et al., 2009) aims to predict the quality of machine translation (MT) outputs without human references, which is what sets it apart from translation metrics like BLEU (Papineni et al., 2002) or TER (Snover et al., 2006). Most approaches to quality estimation are trained to predict the post-editing effort, i.e. the number of corrections the translators have to make in order to get an adequate translation. The effort is measured by the HTER metric (Snover et al., 2006) applied to human post-edits.

In this paper, we introduce a light-weight neural method with pre-trained embeddings, that means it does not require any pre-training. The second proposed method is the extension of the first one: besides pre-trained embeddings, it takes log probability from any MT system as an input feature.

In addition to the official datasets provided for this year's WMT sentence level shared task, we analyze the performance of our methods against the extended datasets made from previous years data. Using the extended datasets allows to get a more reliable score and avoid skewed distributions of the predicted metrics.

Besides that we apply our method to predict direct human assessment (DA) (Graham et al., 2017). In direct human assessment humans compare the machine translation output with a reference translation not seeing a source translation. Usually MT metrics (Ma et al., 2018) are compared to DA, but we decided to compare our predictions as well, because there is a difference between a number of post-edits and a human assessment. For example, if everything in a translation is perfect except one thing: all indefinite articles are missed, the number of post-edits may be large enough and a score will be low whereas humans likely give it a high score. The main difference between MT metrics and quality estimation is that quality estimation is computing without reference sentences.

2 Architecture

Our method performs sentence-level quality estimation of machine translation. As other state-of-the-art methods (Kim et al., 2017; Fan et al., 2018), we use a neural-based architecture. However, compared to the other neural-based methods, we do not train embeddings from scratch, that usually takes a lot of data and computational resources. Instead of that, we use already well trained and freely available embeddings.

For our method we have picked BERT (Devlin et al., 2018) and LASER (Artetxe and Schwenk, 2018) multilingual embeddings toolkits. We extract both BERT and LASER embeddings and feed them into a feed-forward neural network. A sigmoid output layer produces the desirable score. In case of HTER prediction we can add log probability score obtained from a neural MT system as an additional feature to the described above feed-forward neural network. The whole architecture of our system is depicted in Fig.1.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 103–107
Florence, Italy, August 1-2, 2019.

BERT embeddings are extracted from a deep bidirectional transformer encoder, which is pre-trained on Wikipedia data, with the aim of generating a general-purpose "language understanding". LASER embeddings are extracted from bidirectional word-level recurrent encoder, where sentence embeddings are extracted from max-pooled word embeddings, trained on publicly available parallel corpora.

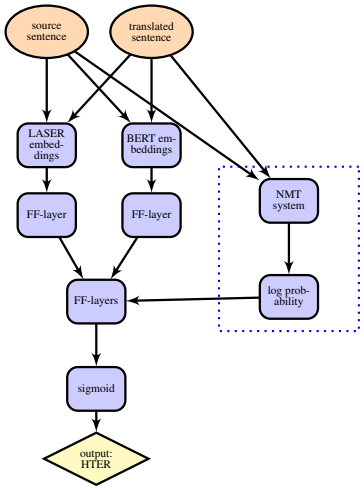

Figure 1: The proposed methods: LABEL: it requires LASER and BERT to get embeddings and NMT system to compute log probability and LABE: it requires only LASER and BERT to get embeddings.

3 Experimental Settings

In this section we analyze the performance of proposed methods on different prediction outputs (HTER and DA) and different datasets and compare them with another neural method DeepQuest (Ive et al., 2018) that does not require additional data.

To predict HTER we take a dataset that contains source sentences, their translated outputs and HTER scores. It is domain-specific: IT or pharmaceutical depending on the language pair. As there is no large enough corpus with DA labels, we use a dataset that consists only of source sentences and their machine translation output. The domain of this corpus is more general and source sentences have taken from the open resources.

3.1 Experiments

We have implemented our methods using the Keras toolkit. As a regression model we have used four-layered feed-forward neural network with sigmoid as a final activation function.

To obtain a log probability score, we trained neural MT systems using sockeye toolkit. We used Transformer (Vaswani et al., 2017) as a network architecture with six layers in encoder and decoder, word vectors of size 512, batch size 50, and Adam (Kingma and Ba, 2015) as optimizer with an initial learning rate of 0.0002.

We present two models with different set of features:

- **LABE**: embeddings extracted from LASER and BERT
- **LABEL**: embeddings extracted from LASER and BERT and log probability obtained from Transformer NMT model

BERT embeddings are extracted for multilingual cased BERT model. Only the last layer of embeddings is extracted. BERT gives 728-dimension embeddings for each word, source and target embeddings are separated by a special token and then average pooling is used to get sentence embeddings for source and target sentences.

3.2 Data and Results of HTER Prediction

Data

We gathered the data from WMT16 - WMT18 shared tasks on sentence-level quality estimation for English-German (En-De) (Bojar et al., 2016a, 2017a; Specia et al., 2018), from WMT17 - WMT18 German-English (De-En) and from WMT 18 English-Czech (En-Cs).

The En-De data contains translations from neural and statistical MT systems and De-En and En-Cs datasets contain outputs only from statistical MT. However, for our method there is no difference between neural and statistical MT output. En-De and En-Cs sentences on the IT domain and De-En — on the pharmaceutical domain.

We removed duplicated sentences and randomly split data into training, dev and test sets in the 70/20/10 ratio. As a result, we got the following number of sentences:

- **En-De**: $\approx$ 55K/16K/8K
- **De-En**: $\approx$ 37K/10K/5K
- **En-Cs**: $\approx$ 29K/8K/4K

We intentionally increased the size of the test sets to reduce the impact of skewed distributions towards high quality translations. These fluent

translation have the HTER score equalled zero and make up 70% of all data. Such distribution where we have 70% of zeros and other 30% of data is uniform from 0 to 1 is hard to learn with a regression model.

Results

Below we describe the results of our systems for two test datasets: the extended dataset is described above and the second one is the small dataset (around 1K sentences) provided by organizers of WMT19.

Results for extended datasets The resulting Pearson and Spearman coefficients for the all given language pairs are presented in Table 1. As one can see the highest values were obtained by applying the models `LABEL`, but the difference of the computing values is small. The obtained numbers for En-De and En-Cs are close to each other whereas the resulting coefficients for De-En are noticeably higher. Both our models showed the better performance than `deepQuest`.

	LABE		LABEL		deepQuest	
	PCC	SCC	PCC	SCC	PCC	SCC
DEEN	0.599	0.586	0.64	0.615	0.368	0.347
ENDE	0.533	0.566	0.542	0.57	0.294	0.305
ENCS	0.542	0.532	0.557	0.549	0.446	0.433

Table 1: Pearson and Spearman correlation coefficients for the monolingual models LABE and LABEL, and `deepQuest` . For models LABE and LABEL we show PCC and SCC between ensemble of five runs and HTER.

Results for WMT 2019 The results for the small WMT dataset do not look so impressive (Table 2) compared to the results of extended datasets. Without knowledge of data, it is difficult to say what the reason for it. We can assume that it may be due to the skewed distribution of the given dataset. It is worth noting that the same En-De (nmt) dataset was given also in WMT18 shared task and looking at the results[1], we can see a drop in performance for this dataset as well.

3.3 Data and Results for human assessment prediction

Data

We took data from News Translation Tasks 2015-2018 years (Bojar et al., 2015, 2016a, 2017a, 2018) for En-De, English-Finnish (En-Fi), English-Russian (En-Ru) (both directions for all three language pairs) and En-Cs. The data consists of source sentences and their translation. The number of unique source sentences (≈10-11K for each language pair) are significantly less than the number of translation, because every source sentence has several translations obtained from different systems. We randomly split the data into training and dev sets in the ratio 80/20:

- **En-De**: ≈ 141K/35K
- **De-En**: ≈ 111K/28K
- **En-Fi**: ≈ 100K/25K
- **Fi-En**: ≈ 73K/18K
- **En-Ru**: ≈ 95K/24K
- **Ru-En**: ≈ 94K/24K
- **En-Cs**: ≈ 113K/28K

As test sets we used DAseg-newstest2016 (Bojar et al., 2016b) that consists of 560 sentences for each language pair. As fine-tuning sets we took DAseg-newstest2015 (Stanojević et al., 2015) and DAseg-newstest2017 (Bojar et al., 2017b) that gave us around 1K sentences per each language pair.

	LABE		LABEL	
	PCC	SCC	PCC	SCC
ENDE	0.319	0.377	0.249	0.253
ENRU	0.401	0.336	-	-

Table 2: Pearson and Spearman correlation coefficients for the monolingual models LABE and LABEL. Test set: official test set of WMT19. We show PCC and SCC between ensemble of five runs and HTER.

Results

Below we describe the obtained results for newstest2016 (Bojar et al., 2016b) and compare them with results of metrics tasks. At the time of publication of the article, results of newstest2019 were not yet available.

Results for DAseg-newstest2016 The both proposed methods are supervised, so to train models we need labels. As DA data is scarce resource we trained models using chrF++ (Popović, 2017) (with default hyper-parameters) as labels.

To investigate how the number of language pairs affects the performance of models, we trained several models: with one language pair in the training

[1] `http://statmt.org/wmt18/quality-estimation-task.html#results`

set, with four (De-En, En-De, En-Cs, En-Ru) and with seven language pairs. As can be seen in the Figure 2, the best results were achieved with the mono language pair models, although the difference between mono- and multimodels is not large.

We also fine-tuned our models by using human assessment data. Fine-tuned models showed a little bit better results compared to the non-tuned models (Figure: 2).

We compared the obtained results to the metrics results. For De-En the best resulting Pearson correlation coefficient for metrics is 0.601 and for En-Ru is 0.666 (Bojar et al., 2016b), whereas the best scores of our models are 0.520 and 0.668 for De-En and En-Ru respectively. Our results are comparable to the metrics results, despite the fact that we did not use reference sentences in contrast to the metrics task.

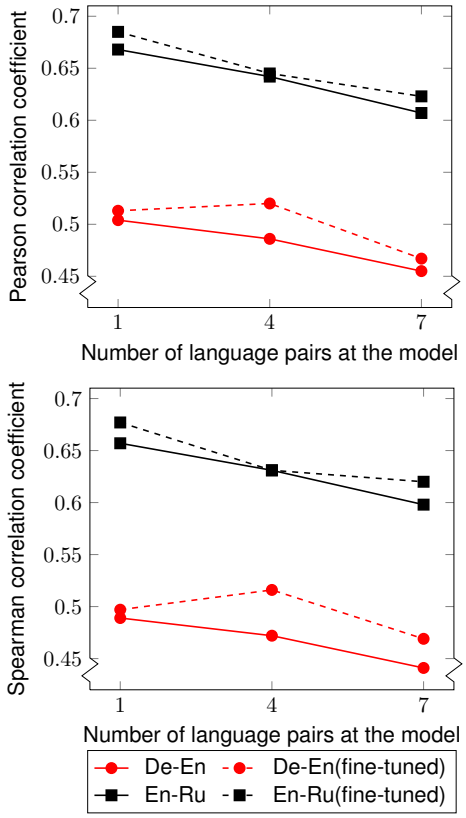

Figure 2: Pearson and Spearman correlation coefficients for LABE model and different number of language pairs in training dataset. We show average over three runs. Test dataset: newstest2016

Results for DAseg-newstest2019 We prepared scores for all language pairs described in 3.3 by using non-tuned models trained on seven language pairs and for De-En, En-Ru, Ru-En, Fi-En by using fine-tuned models. Results of this submission will be available (Fonseca et al., 2019).

4 Conclusions

We proposed neural-based models for quality estimation of machine translation. One of our models requires only freely available embeddings (LASER and BERT) and the second needs also log probability from any MT system (in our experiments, we use Transformer MT system).

We analyzed performance of both models on different language pairs and different prediction outputs and compared them to another neural quality estimation system. Both our methods showed better results compared to another light-weight approach deepQuest and we got comparable results with the metrics tasks even without using references.

Acknowledgments

This work was supported in part by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825303 as well as the Estonian Research Council grant no. 1226

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 315. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017a. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016a. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 131–198.

Ondrej 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 *WMT@EMNLP*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017b. 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.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 199–231.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kai Fan, Bo Li, Fengming Zhou, and Jiayi Wang. 2018. "bilingual expert" can find translation errors. *CoRR*, abs/1807.09433.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared task on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation*.

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.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. Deepquest: a framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568.

Diederik P Kingma and Lei Ba. 2015. J. adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

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.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the wmt15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273.

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.