Alibaba Submission for WMT18 Quality Estimation Task

Jiayi Wang*, **Kai Fan***, **Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, Luo Si**
Machine Intelligence Technology Lab, Alibaba Group
Hangzhou, China
{joanne.wjy, k.fan, shiji.lb, zfm104435, boxing.cbx, taiwu.syb, luo.si}@alibaba-inc.com

Abstract

The goal of WMT 2018 Shared Task on Translation Quality Estimation is to investigate automatic methods for estimating the quality of machine translation results without reference translations. This paper presents the QE Brain system, which proposes the neural Bilingual Expert model as a feature extractor based on conditional target language model with a bidirectional transformer and then processes the semantic representations of source and the translation output with a Bi-LSTM predictive model for automatic quality estimation. The system has been applied to the sentence-level scoring and ranking tasks as well as the word-level tasks for finding errors for each word in translations. An extensive set of experimental results have shown that our system outperformed the best results in WMT 2017 Quality Estimation tasks and obtained top results in WMT 2018.

1 Introduction

Quality Estimation (QE) is a task to estimate the quality of a Machine Translation (MT) system without the presence of any manually annotated reference translations. It can serve in a variety of computer-aided scenarios such as translation results screening before release or translation quality comparison between different MT systems. Currently, the classical and widely-used method to evaluate an MT system is measured by BLEU (Papineni et al., 2002), a statistical language-independent metric that requires human golden references for validation. What if we expect to efficiently get the detailed quality evaluation feedbacks (e.g. sentence or token-wise scoring) from an extremely large number of machine translation outputs? An automatic method with no access to any reference is highly appreciated.

The common approach to automatic translation quality estimation is to transform the problem into a supervised regression or classification task for sentence-level scoring and word-level labeling respectively. Traditional baseline models in WMT 12-17 have two modules: human-crafted rule-based feature extraction model via QuEst++ (Specia et al., 2015) (sentence-level task) or Marmot[1] (word-level task); and an SVM regression with an RBF kernel as well as grid search algorithms for predicting how much effort is needed to fix translations to acceptable results (sentence-level task) or a sequence-labeling model with CRFSuit toolkit to predict which word in the translation output needs to be edited (word-level task). A recently proposed predictor-estimator model with stack propagation (Kim et al., 2017) is a recurrent neural network (RNN) based feature extractor and quality prediction model that ranked first place in WMT17. Another novel method is to train an Automatic Post-Editing (APE) system and adapt it to predict sentence-level quality scores and word-level quality labels (Martins et al., 2017). A promising APE system can serve as a guidance to QE system by explicitly explaining errors in the translation output.

Our submitted system for sentence and word level QE tasks in WMT18, named ***QE Brain*** has two phases: feature extraction and quality estimation. In the phase of feature extraction, it extracts high-level latent joint semantics and alignment information between the source and the translation output, relying on the "neural Bilingual Expert model" introduced by Fan et al. (2018) as a prior knowledge model, which is trained on a large parallel corpus. The high-level latent semantic features and manually designed mis-matching features (Fan et al., 2018) exported from the prior

* indicates equal contribution.

[1] https://github.com/qe-team/marmot

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 822–828

https://doi.org/10.18653/v1/W18-6466

knowledge model are fed into a predictive model in the phase of quality estimation, with which the scoring prediction for the sentence-level task and erroneous or missing word predictions for the word-level task are targeted. This paper presents our submissions for the WMT18 Quality Estimation English-German and German-English Shared Tasks, namely, (i) a sentence-level QE scoring prediction system and (ii) a word-level QE labeling prediction system including word predictions and gap predictions. Since both systems are supposed to understand the complex semantic relationship between the source and the translation output, the features produced by a pre-trained neural Bilingual Expert model can be shared by the two level tasks per language direction.

In Section 3, we will discuss several techniques to boost our system's performance. We make use of extra human-crafted baseline features including basic descriptive statistics, language model (LM) probabilities and alignments information of the source and the translation output. They are combined with features from the neural Bilingual Expert model to predict the sentence-level scores. In addition, to make up the shortage of QE training data, we apply the round-trip translation technique to generate some artificial QE data that increases the error diversity and prevents overfitting. To further enhance our model's performance, we use a greedy algorithm based ensemble selection method to decrease the individual error among a bunch of single quality estimation models.

2 QE Brain Baseline Model

QE Brain base single model contains a feature extractor and a quality estimator. The feature extractor relies on the Bilingual Expert model to extract features representing latent semantic information of the source and translation pair. These features will be fed into a quality estimator to estimate the translation quality.

The Bilingual Expert model uses self-attention mechanism and transformer neural networks to construct a bidirectional transformer architecture (Fan et al., 2018), serving as a conditional language model. It is used to predict every single word in the target sentence given the entire source sentence and its context . The Bilingual Expert model consists of three modules: (i) transformer self-attention based encoder for the source sentence, (ii) forward and backward encoders for the target sentence with the masked self-attention in the transformer decoder module, (iii) reconstruction for the target sentence. Once the model is fully trained, we can use the prior knowledge learned from the Bilingual Expert model to extract the features for the subsequent translation quality estimator. There are two kinds of features upon the Bilingual Expert model defined by Fan et al. (2018): model derived features of latent representations and manually extracted mismatching features.

When we perform quality estimation on a source and translation pair, we need to obtain the semantics information of the source and the translation output and their alignment information. We can assume that it is more likely for the model to predict a correct target word if only few words around it are incorrect. Fan et al. (2018) claims that both the latent representations of the k-th word in the translation output and its mismatching features that reflect the error severity if it is a mistake are sufficiently beneficial to the downstream quality predictive model. Choices of the quality estimation models are compared as well. It is found that the bi-directional LSTM (Graves and Schmidhuber, 2005) will be appropriate in the QE situation. We treat the feature extraction model based on the neural Bilingual Expert model and the quality estimation based on Bi-LSTM model as our baseline system.

3 Boosting the QE Model Performance

3.1 Human-crafted Features

Along with the features produced by the Bilingual Expert model, we extract another 17 QE baseline features for the sentence-level task using QuEst++ and additional resources (source and target corpora, language models, ngram counts and lexical translation tables) provided on the WMT18 QE website[2]. Kozlova et al. (2016) verifies the significance of these features using Random Forest (Breiman, 2001). Four of them are the most crucial among all according to their degrees of importance.

- percentage of trigrams in quartile 4 of frequency of source words in a corpus of the source language

- LM probability of source sentence

[2] http://www.statmt.org/wmt18/quality-estimation-task.html

- percentage of bigrams in quartile 4 of frequency of source words in a corpus of the source language

- average number of translations per source word in the sentence

Language models (LM) assign probabilities to generate hypotheses in the target language informing lexical selection in statistical machine translation (SMT). It is reasonable that three of the above four baseline features are derived from the LM. Moreover, alignment models can essentially help SMTs determine translational correspondences between the N-grams in the source with those of the same meanings in the target. Particularly, a satisfying translation result can contain as many translated words as possible, according to an alignment model, IBM model 1 or 2. Consequently, average number of translations per source word in the sentence becomes large.

Fan et al. (2018) proposed to use the concatenation of the model derived and mis-matching features as input of a Bi-LSTM quality predictive model. The sentence-level score prediction can be formulated as a regression problem with the objective function,

$$\arg\min \left\| h - \text{sigmoid}\left(\mathbf{w}^\top [\overrightarrow{\mathbf{h}_T}; \overleftarrow{\mathbf{h}_T}]\right)\right\|_2^2 \quad (1)$$

where $\overrightarrow{\mathbf{h}_T}$ and $\overleftarrow{\mathbf{h}_T}$ are the hidden states of the last time stamps of the Bi-LSTM's output, h represents the translation score (HTER) and $\mathbf{w}$ is a vector. Alternatively, we introduce the human-crafted features as additional linear components for the predictive layer with a sigmoid activation function. Therefore, the objective function can be rewritten as,

$$\arg\min \left\| h - \text{sigmoid}\left(\mathbf{w}^\top [\overrightarrow{\mathbf{h}_T}; \overleftarrow{\mathbf{h}_T}; \mathbf{f}_h]\right)\right\|_2^2 \quad (2)$$

where $\mathbf{f}_h$ is the 17-dimensional QE baseline features.

3.2 Artificial QE Data Construction

Unlike stacking of an APE-based QE system and a "pure" QE system trained only on the provided QE training dataset (Martins et al., 2017), we came up with the idea to take advantage of the artificial training data augmentation technique (Junczys-Dowmunt and Grundkiewicz, 2016) in the APE task to provide more supplementary training data, aiming to increase the diversity of erroneous translations during the training process so that it can reduce the overfitting of our model. We trained two English-German quality estimation models with (i) the real QE training data alone or (ii) the real and artificial QE data, and evaluated them on the development data and the data made up with 1800 random samples from the real QE training data to investigate the robustness of them. As shown in Fig 1, the model trained with (ii) (Experiment 2) is more robust than the model trained with (i) (Experiment 1), but can achieve comparable performance on the development data.

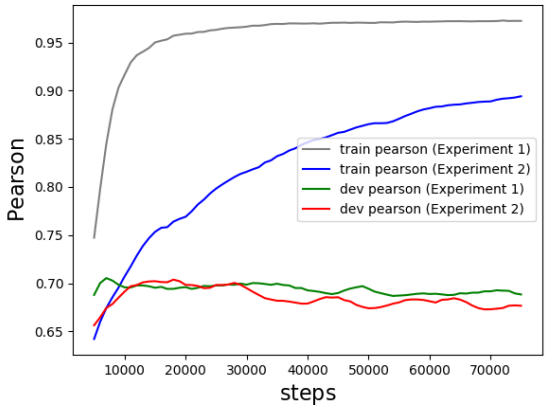

Figure 1: Robustness analysis on English-German QE model. Experiment 1: model trained with real QE data; Experiment 2: model trained with real and artificial QE data

The round-trip translation process can produce literal translations that may be similar to post-edited triplets including sources (SRC), translation outputs (MT) and post editions (PE). In order to mimic the QE data, we randomly pick triplets generated by the round-trip translation technique according to the distribution of HTERs in the real QE training and development data.

3.3 Greedy Ensemble Selection

To generate an ensemble of submissions for the WMT 18 QE task, the simplest methods are averaging the predictive scores for the sentence level and majority voting for the predictive labels for the word level from a number of single models. Homogeneous models can be derived from performing the same learning methodology but with different hyper-parameters of the model architecture including the neural Bilingual Expert model and Bi-LSTM quality predictive model.

In the sentence level, adding human-crafted features can be optional when we make different assumptions about the features of source and translation pairs. Under this situation, heterogeneous models can be derived from performing the same learning algorithm on different datasets. We can also use the Byte-Pair Encoding (BPE) tokenization as a substitution for a word tokenization in text pre-processing. Fan et al. (2018) compared the performance of the word and BPE tokenization on both sentence and word levels in WMT 18 and the results show that the models with BPE tokenization can produce comparable or better results than those with word tokenization.

In general, the ensemble output of K single models can be produced by the following objective function,

$$\arg\max_{t_k} \sum_{k=1}^{K} w_k m_k \left(X = x, T = t_k\right) \quad (3)$$

where m_k is the k-th single model that has probability distribution $m_k\left(x, t_k\right)$ with its corresponding weight w_k. X represents the feature instance of a single model, and T represents the HTER or the word label where t_k can be a continuous quality score or an OK/BAD label respectively. We assign equal weights to every single model in our case for simplicity.

Since not every single model in the ensemble is always needed for the optimized prediction, it is appropriate to select a subset from all candidate models. We follow the greedy ensemble selection algorithm, Focused Ensemble Selection (FES) (Partalas et al., 2008), to reduce the size of averaging ensembles but improve its efficiency and predictive performance.

In the sentence level, FES's output is averaging HTER scores of selected single models. However, in the word level, the ensemble can be made by majority voting of the binary predictions for selected single models or averaging their probabilities of predicting the word as OK. We use the development data for evaluation under the assumption that the development data and the test data are from the same distribution, even if it might be susceptible to overfitting. However, we did not observe this phenomena in results released for the test data in WMT18 QE task.

4 Experiments

4.1 Experimental Settings

4.1.1 Data for Bilingual Expert Model

We evaluated our system, QE Brain, for the WMT17/18 QE task for sentence/word-level in English-German and German-English. The followings are data resources that we used for training the neural Bilingual Expert model,

- parallel corpora released for the WMT17/18 News Machine Translation Task[3]
- UFAL Medical Corpus and Khresmoi development data release for the WMT18 Biomedical Translation Task[4]
- source and target corpora MT training data released in the additional resources for the WMT18 QE Task
- src-pe pairs for for the WMT17/18 QE Task

We filtered all the corpora except src-pe pairs with basic rules to guarantee the quality. A "high-quality" sentence pair should both start with a Unicode letter character, the lengths of them are equal to or less than 70, and the length ratio of the source sentence and the target one should be bounded by 1/3 and 3. The total resulting qualifying parallel corpora roughly include 13 million for WMT17 QE tasks and 29 million for WMT18 QE tasks.

4.1.2 Data for Quality Estimation Model

The data for quality estimation contains two parts: (i) real QE data provided by WMT QE organizers; (ii) artificial QE data generated by the round-trip translation technique (Junczys-Dowmunt and Grundkiewicz, 2016). We first combined the real QE data with the artificial QE data to train a baseline quality estimation model, then fine tuned the model with the real QE data alone. The English-German IT domain artificial QE data can be obtained directly from the additional resources of WMT18 Auto Post-Editing task[5] created by Junczys-Dowmunt and Grundkiewicz (2016). We applied the English-German artificial QE data on

[3] http://www.statmt.org/wmt18/translation-task.html

[4] http://www.statmt.org/wmt18/biomedical-translation-task.html

[5] http://www.statmt.org/wmt18/ape-task.html

Method	test 2017 en-de					test 2017 de-en				
	Pearson's r ↑	MAE ↓	RMSE ↓	**Spearman's ρ ↑**	DeltaAvg ↑	**Pearson's r ↑**	MAE ↓	RMSE ↓	**Spearman's ρ ↑**	DeltaAvg ↑
Baseline	0.397	0.136	0.175	0.425	0.0745	0.441	0.128	0.175	0.45	0.0681
Unbabel	0.641	0.128	0.169	0.652	0.1136	0.626	0.121	0.179	0.61	0.974
POSTECH Single-Ensemble	0.6731	0.1067	0.1412	0.7029	0.1198	0.7146	0.0942	0.1359	0.6327	0.1044
POSTECH Multi-Ensemble	0.6954	0.1019	0.1371	0.7253	0.1232	0.7280	0.0911	0.1332	0.6542	0.1064
QE Brain Base Single Model	0.6837	0.1001	0.1441	0.7091	0.1200	0.7099	0.0927	0.1394	0.6424	0.1018
+ HF	0.6842	0.1013	0.1449	0.7150	0.1213	0.7085	0.0901	0.1406	0.6551	0.1040
+ FT	0.6957	0.1001	0.1420	0.7205	0.1208	0.7128	0.0933	0.1394	0.6422	0.1013
+ HF/FT	0.6813	0.1021	0.1460	0.7070	0.1197	0.7149	0.0889	0.1385	0.6596	0.1026
QE Brain Ensemble	**0.7159**	0.0965	0.1384	**0.7402**	0.1247	**0.7338**	0.0882	0.1333	**0.6700**	0.105

Table 1: Results of sentence-level scoring and ranking on WMT17. HF: human features; FT: fine-tune strategy with artificial QE data.

the SMT QE task. For the neural machine translation (NMT) QE task, we followed the same procedure but trained two NMT models (German-English and English-German) instead.

Similarly, when generating German-English Pharmacy domain artificial QE data, we first applied domain data selection to the English monolingual corpus admissible for the WMT18 News and Biomedical Translation data with cross-entropy filtering method and seed data set – post-editing training data and the English biomedical data. In total, we got 5 million domain-like sentences for the round-trip translation. Afterwards, we created two phrase-based translation models, English-German and German-English, using the parallel bilingual corpora for the WMT18 News and Biomedial Translation tasks but with different language models. The 5 million domain-like sentences as PEs would be first translated to German as SRCs and the SRCs would be then translated to English as MTs. Finally, we would have 5 million artificial APE training data, leading to 5 million artificial QE training data with corresponding HTERs and word labels via the TER tool.

We filtered the English-German and German-English artificial QE data according to the HTER distribution of the combination of QE training and development data, and randomly pick 300,000 triplets per language pair.

Method	**Pearson's r ↑**	MAE ↓	RMSE ↓	**Spearman's ρ ↑**
	test 2018 en-de SMT			
Baseline	0.3653	0.1402	0.1772	0.3809
UNQE	0.7000	0.0962	0.1382	0.7244
QE Brain Ensemble 1	0.7308	0.0953	0.1383	0.7470
QE Brain Ensemble 2	**0.7397**	0.0937	0.1362	**0.7543**
Method	test 2018 en-de NMT			
Baseline	0.2874	0.1286	0.1886	0.4195
UNQE	**0.5129**	0.1114	0.1749	**0.6052**
QE Brain Ensemble 1	0.5005	0.1134	0.1734	0.6002
QE Brain Ensemble 2	0.5012	0.1131	0.1742	0.6049
Method	test 2018 de-en SMT			
Baseline	0.3323	0.1508	0.1928	0.3247
UNQE	**0.7667**	0.0945	0.1315	0.7261
QE Brain Ensemble 1	0.7539	0.0981	0.1355	0.7222
QE Brain Ensemble 2	0.7631	0.0962	0.1328	**0.7318**

Table 2: Results of sent level QE on WMT2018

Method	F1-BAD	F1-OK	**F1-Multi**
	test 2017 en-de		
Baseline	0.407	0.886	0.361
DCU	0.614	0.910	0.559
Unbabel	0.625	0.906	0.566
POSTECH Ensemble	0.628	0.904	0.568
QE Brain Base Single Model	0.6407	0.9045	0.5795
+ FT	0.6410	0.9083	0.5826
QE Brain Ensemble	0.6616	0.9128	**0.6039**
Method	test 2017 de-en		
Baseline	0.365	0.939	0.342
POSTECH Single-Ensemble	0.552	0.936	0.516
Unbabel	0.562	0.941	0.529
POSTECH Multi-Ensemble	0.569	0.940	0.535
QE Brain Base Single Model	0.5750	0.9471	0.5446
+ FT	0.5816	0.9470	0.5507
QE Brain Ensemble	0.5924	0.9475	**0.5613**
Method	test 2018 en-de SMT		
Baseline	0.4115	0.8821	0.3630
SHEF-PT	0.5080	0.8460	0.4298
QE Brain Ensemble 1	0.6616	0.9168	0.6066
QE Brain Ensemble 2	0.6808	0.9175	**0.6246**
Method	test 2018 en-de NMT		
Baseline	0.1973	0.9184	0.1812
SHEF-PT	0.3353	0.8691	0.2914
QE Brain Ensemble 1	0.4750	0.9152	**0.4361**
QE Brain Ensemble 2	0.4767	0.9149	0.4347
Method	test 2018 de-en SMT		
Baseline	0.4850	0.9015	0.4373
SHEF-PT	0.4853	0.8741	0.4242
QE Brain Ensemble 1	0.6475	0.9162	0.5932
QE Brain Ensemble 2	0.6523	0.9217	**0.6012**

Table 3: Results of word-level word prediction on WMT17/18

Method	F1-BAD	F1-OK	**F1-Multi**
UAlacante SBI	0.1997	0.9444	0.1886
SHEF-bRNN	0.2710	0.9552	0.2589
SHEF-PT	0.2937	0.9618	0.2824
QE Brain	0.5109	0.9783	**0.4999**

Table 4: Results of word-level gap prediction on WMT18 En-De SMT

4.1.3 Model Settings

The number of layers for the self-attention encoder and forward/backward self-attention decoder are all set as 2, where we use 8-head self-attention in practice. The number of hidden units for feed-forward sub-layer is 512. The bilingual expert model is trained on 8 Nvidia P-100 GPUs for about 3 days until convergence. For translation QE model, we use only one layer Bi-LSTM, and it is trained on a single GPU. Notice that for the QE task of WMT17, it is prohibited to use any data

from 2018, since the training data of 2018 includes some test data of 2017. The same setting is applied to all following experiments associated with 2017. We tuned all the hyper-parameters of our model on the development dataset to obtain the best single model, and report the corresponding results for test data.

We increased the model diversity from two perspectives. First, in terms of data resources, we experienced with three strategies: word/BPE tokenization, w/ or w/o artificial QE data and w/ or w/o human-crafted features for the sentence-level task. Secondly, we tuned the number of units for Bi-LSTM with 96 or 128 and training batch size with 32 or 64 from the model's perspective.

4.2 Evaluation Results

In this section, we will report the experimental results of our approach for WMT 2017 and 2018. For WMT17 QE task, we tried to verify our proposed strategies. For WMT18 QE task, we mainly participated in the sentence-level scoring and ranking tasks and the word-level word prediction tasks for English-German SMT, English-German NMT and German-English SMT. In addition, we also submitted results for the word-level gap predictions for English-German SMT. In Table 2, part of Table 3 and Table 4, results of WMT18 QE tasks are listed according to the WMT18 QE website.

4.2.1 Ablation Study on WMT17 QE Task

Since we can access the translation outputs of human post-editing for test data, it provides an ideal held-out test data to verify our proposed strategies. We illustrated our results in Table 1 and part of Table 3 on WMT17 QE Task. The competitors are POSTECH, DCU and Unbabel. Their results can be found in (Bojar et al., 2017) , Section 4.4 and Section 4.5. We also listed the WMT QE baseline results for reference. The QE Brain base single model follows the exact training scheme in (Fan et al., 2018) with model derived features and mismatching features. In sentence level, either incorporating human features or the use of artificial QE data will positively contribute to the metrics. For Pearson's r, the single fine-tuning strategy yields the improvement $+0.01$ on English-German and $+0.003$ on German-English. For Spearman's ρ, the single model with human features improves the performance by $+0.006$ in English-German and $+0.013$ in German-English.

In word level, we did not use any human features, but we found fine-tune strategy can always improve the performance. For F1-Multi, the single fine-tuning strategy yields the improvement $+0.003$ on English-German and $+0.006$ on German-English. In general, with all these strategies, our single models can be comparable or better than the state-of-the-art (SOTA) ensemble systems of WMT17 QE task. Our ensemble models significantly outperform all of the SOTA systems.

4.3 Ensemble Analysis on WMT18 QE Task

As we discussed previously, we tried both word and BPE tokenization for the data pre-processing. Thus, we submitted two types of ensemble models, where Ensemble 1 is referred to the model ensembles trained with word tokenization and Ensemble 2 is the model ensembles trained with both word and BPE tokenizations. Training with BPE tokenization can naturally increase the model diversity, so it makes sense that Ensemble 2 performs better than Ensemble 1, except for English-German NMT word-level task, which is very likely due to the small data size ($<$14000).

5 Conclusion

This paper introduces our machine translation quality estimation system, QE Brain, for both the sentence-level and word-level tasks in WMT 2018 Quality Estimation. The system proposes the neural Bilingual Expert model to extract semantic features from both the source and translation output for estimating translation quality with a bi-directional LSTM predictive model. In particular, three important strategies are utilized for obtaining positive results as incorporating human-crafted features, artificial QE data augmentation for more diversified training data and model ensemble with a greedy algorithm. The results of our system obtained No.1. in the English-German SMT scoring and ranking tasks as well as the German-English SMT ranking tasks. Furthermore, our system also produced the best results in all word-level English-German and German-English word and gap prediction tasks.

References

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt

Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 169–214.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Zhiming Chen, Yiming Tan, Chenlin Zhang, Qingyu Xiang, Lilin Zhang, Maoxi Li, and Mingwen Wang. 2017. Improving machine translation quality estimation with neural network features. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 551–555.

Kai Fan, Bo Li, Fengming Zhou, and Jiayi Wang. 2018. "bilingual expert" can find translation errors. *arXiv preprint arXiv:1807.09433*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. *CoRR*, abs/1605.04800.

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, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 562–568.

Anna Kozlova, Mariya Shmatova, and Anton Frolov. 2016. YSDA participation in the wmt'16 quality estimation shared task. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 793–799.

André FT Martins, Fabio Kepler, and Jose Monteiro. 2017. Unbabel's participation in the wmt17 translation quality estimation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 569–574.

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, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Ioannis Partalas, Grigorios Tsoumakas, and Ioannis P. Vlahavas. 2008. Focused ensemble selection: A diversity-based method for greedy ensemble selection. In *ECAI 2008 - 18th European Conference on Artificial Intelligence, Patras, Greece, July 21-25, 2008, Proceedings*, pages 117–121.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, System Demonstrations*, pages 115–120.