Semi-supervised Learning for Quality Estimation of Machine Translation

Tarun Bhatia tarunbhatia.ind@gmail.com
Technische Universität Berlin, Berlin, Germany and SAP SE

Martin Krämer martin.kraemer@sap.com
Eduardo Vellasques eduardo.vellasques@sap.com
SAP SE

Eleftherios Avramidis eleftherios.avramidis@dfki.de
German Research Center for Artificial Intelligence (DFKI), Berlin, Germany

Abstract

We investigate whether using semi-supervised learning (SSL) methods can be beneficial for the task of word-level Quality Estimation of Machine Translation in low-resource conditions. We show that the Mean Teacher network can provide equal or significantly better MCC scores (up to +12%) than supervised methods when a limited amount of labeled data is available. Additionally, following previous work on SSL, we investigate Pseudo-Labeling in combination with SSL, which nevertheless does not provide consistent improvements.

1 Introduction

Through the recent development of Machine Translation (MT), Quality Estimation (QE) has come to serve the need to predict the quality of translation provided by MT systems when no reference translations are available. QE has been mostly treated as a supervised learning problem, where supervised models can be trained on the source and translated text along with their respective quality labels. For example, QE at the word level includes the source and translated sentences as the data and their label sequence includes an OK or BAD label for each translated word in the sentence, which can determine if the word is correctly translated or not and potential errors in the translations can be flagged. In order to train supervised models for such problems, a large amount of labeled data is needed. However, such data is expensive to create as it involves human annotators to post-edit or generate labels for the given translations. Whereas the unavailability of labeled data is a problem, there is an abundance of unlabeled data for such a task, i.e. source sentences and the corresponding translations generated by MT systems. Semi-supervised learning (SSL) methods could be utilized to train QE models with few labeled data available along with unlabeled data that can be generated in abundance.

While the prominent SSL approach of Mean Teacher has shown good performance in computer vision (Tarvainen and Valpola, 2017), there has been little experimentation in NLP. Until now, no research has followed SSL to fine-tune pre-trained language models (LMs) for the task of QE of MT. This work focuses on implementing the aforementioned SSL strategies for word-level QE and tries to answer the following questions:

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 72–83
September 4–8, 2023, Macau SAR, China.

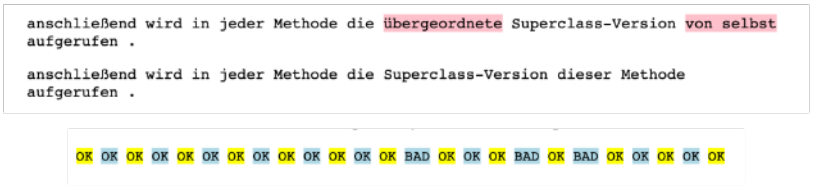

Figure 1: Example of German translation by a MT system and its human post-edited version with labels of gaps and target words for the example translated text. Tags for Gap tokens are highlighted in yellow and tags for words in sequence are highlighted in blue (Specia et al., 2021)

1. Can the SSL method of Mean Teacher perform equal or better than supervised methods on low-resource conditions?

2. Is it possible to utilize the Pseudo-Labeling approach on top of the Mean Teacher architecture to improve the results?

One should note that our research does not aim to achieve the highest MCC scores as compared to SoTA, but to test if SSL techniques can be useful in low resource conditions. Hence, our baseline here are the models created with a fully supervised setup under low resource conditions, which are then compared against our proposed models trained with SSL.

2 Related work

There has been few previous works on SSL methods for NLP. Liang et al. (2020) showed improvement on models trained with labeled data through Pseudo-Labeling for Named Entity Recognition. Wang et al. (2022) suggest a noise-injected consistency training with entropy-constrained pseudo labeling for labeling extractive summarization data. Such approaches have not been investigated for other NLP problems involving token level classification.

State-of-the art QE methods (Rei et al., 2020, 2022) employ fine-tuning of pre-trained LMs, but they don't take into consideration low-resource conditions. Concerning non-supervised methods, Fomicheva et al. (2020) perform unsupervised QE by utilizing internal decoding features of the MT models. With regards to low-resource conditions, Ranasinghe et al. (2021) demonstrate that it is possible to accurately predict word-level quality for any given new language pair from models trained on other language pairs. In an effort to address low resource conditions, Tuan et al. (2021) train off-the-shelf architectures for supervised QE using synthetic data from parallel corpora. To the best of our knowledge, none of the related work in QE of MT has used semi-supervised methods to address low resource conditions.

3 Methods

We focus on the task of QE of MT at the word level, as specified at the Shared Task of QE of WMT (Zerva et al., 2022), which aims at flagging potential errors in the translations generated by any MT system. The word-level task requires assigning binary tags of OK/BAD to determine the correctness of each word in source and target/translated sentences. The following types of labels are used:

Source side Each word in the source sentence is assigned a label (OK/BAD) which determines if the respective word is correctly translated in the target language or not.

Target side Each word in the target sentence is assigned a label (OK/BAD) which determines if the word is a correct translation for the respective word in the source sentence. Additionally,

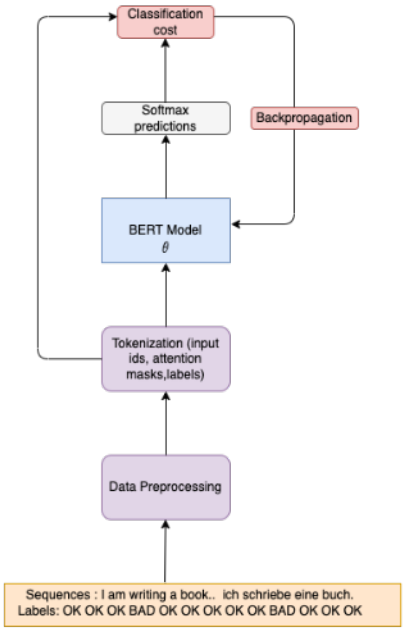

Figure 2: Training flow diagram of supervised fine-tuning methodology

gap tokens are also considered in the beginning, at the end, and between two words of the target sentence. Each gap token is assigned the label BAD if a word or more than a word is missing in the position of the gap token, and it is tagged OK otherwise. An example of the gap tokens can be seen in figure 1.

The proposed methods involve the training of models with Supervised and SSL methods. The fine-tuning of a large LM is done by utilizing three strategies i.e. 1) Supervised Learning, 2) SSL using the Mean Teacher approach., 3) SSL using a Mean Teacher with the Pseudo-Labeling approach.

3.1 Supervised Fine-tuning

Our baseline method is based on supervised fine tuning of a large language model. Here, the pre-trained LM is fine-tuned with only the labeled data for the problem to perform classification of the word sequence. The architecture for fine-tuning of the supervised model is shown in figure 2. As it can be seen, the data is first loaded from files preprocessed to remove the non-useful sequence from the train data. The simple fine-tuning involves utilizing the tokenized data as input to the model. As part of our problem, the input to the model is a sequence of two sentences. The first sentence is the sentence in the source language and the second is the sentence in the translated language.

3.2 Mean Teacher fine-tuning

The Mean Teacher approach (Tarvainen and Valpola, 2017) involves the usage of both labeled and unlabeled data to train the models in this setup. In this architecture (figure 3), two models are initialized, namely Teacher and Student, and weights for both the models are updated differently. The Student is trained using the mainstream method of minimizing the loss, whereas the Teacher is not trained but its weights are updated using an exponential moving average of the Student's weights after processing each batch of data. This behaves as an ensemble technique because eventually, Teacher model weights are the mean of Student model weights from

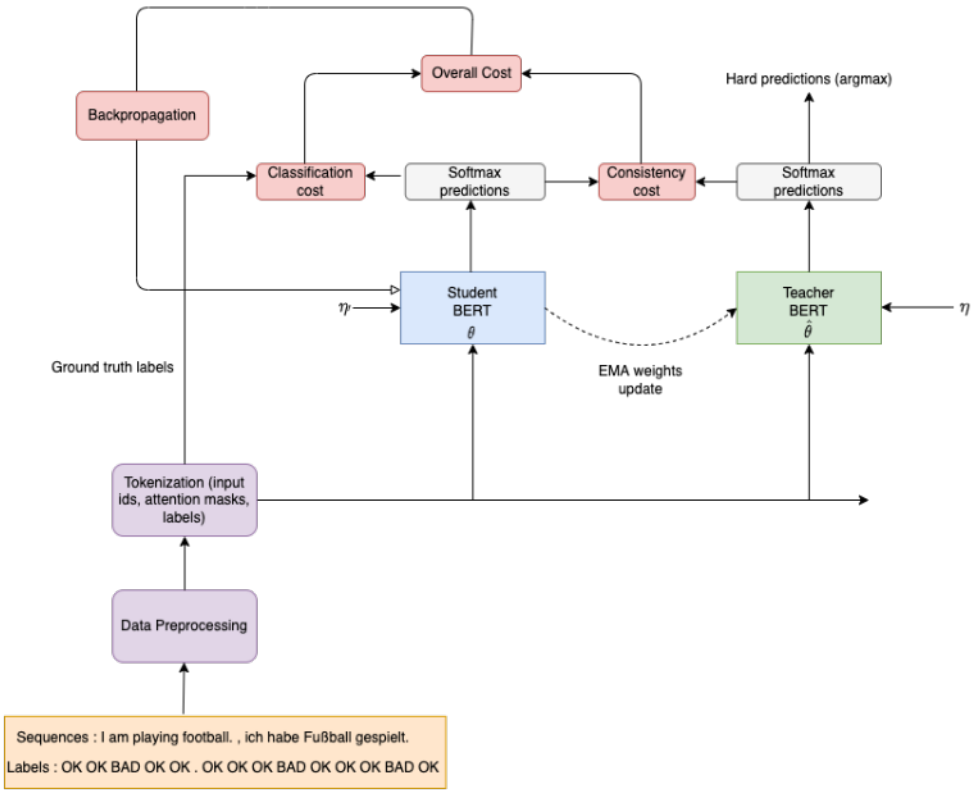

Figure 3: Training flow diagram of proposed fine-tuning methodology with Mean Teacher approach.

previous training iterations, therefore this method is known as *Mean Teacher*.

As shown in figure 3, the data is first preprocessed, and then tokenized using the BERT tokenizer or respective LM tokenizer, which converts the text data into numerical data by mapping each token to a numerical id. The tokenization also involves creating other tensors such as an attention mask, that is passed to convey the model information about the padded tokens. Also, another tensor for the label is passed that contains an actual label for the labeled data and default values for unlabeled data. The data loader, therefore, wraps both labeled and unlabeled data for each training batch, and then the data is passed through the network in batches. The input tensors for each batch are passed through both Teacher BERT and Student BERT models. The models utilize the same structure as defined for the supervised model in the previous section 3.1. An additional noising layer is added to the model which adds random Gaussian noise to the word embeddings generated by the LM. The noising strategy is based on one of the strategies of Zhang and Yang (2018). This noise is controlled by the standard deviation parameter while initializing the respective model, therefore this parameter has to be set to different values while initializing the Teacher and Student model. The noise is added to the models to ensure that both models' classifiers eventually receive a different perturbed version of the same input data. Figure 3 indicates how different loss functions play a crucial role while back-propagating.

The consistency cost ($C(\theta)$) is calculated between the soft predictions of Teacher and Student models so that models eventually learns to predict the same label for the tokens for two perturb version of the same data. Using this consistency cost, the model can effectively utilize

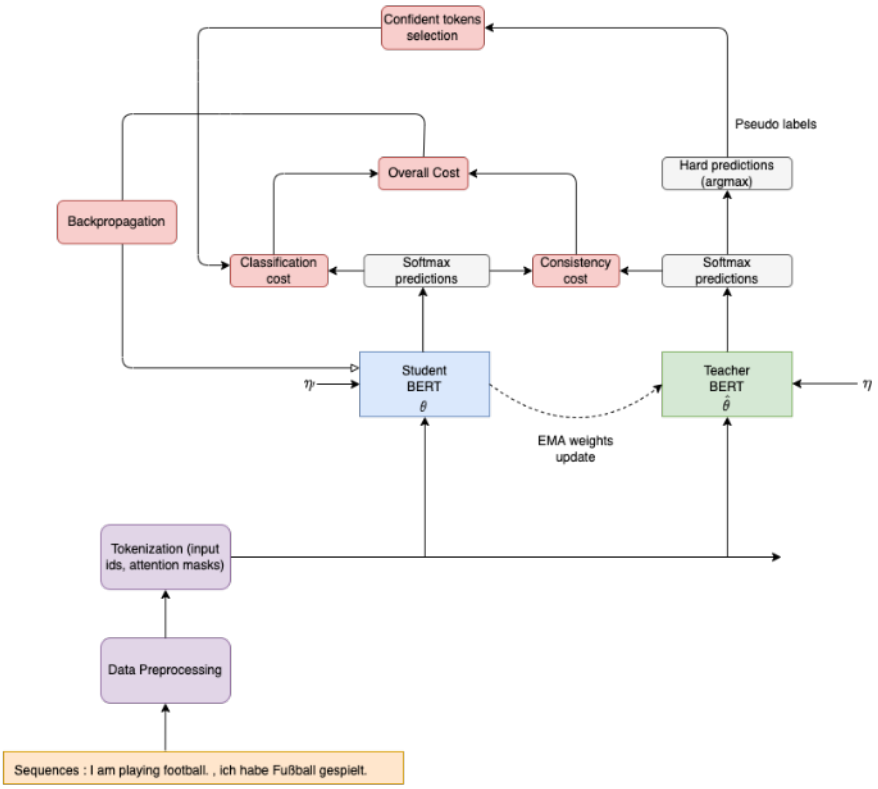

Figure 4: Training flow diagram of proposed PL fine-tuning methodology for stage II

unlabeled data as well to learn the patterns as this cost does not require ground truth labels.

3.3 Pseudo-Labeling fine-tuning

This proposed method calls for using the similar architecture as described in the previous section 3.2 for fine tuning the LMs but using a different methodology, called Pseudo-Labeling. The method is closely related to the work of Liang et al. (2020) and utilizes both labeled and unlabeled data during training. This SSL approach follows a two-stage framework, where in the first stage a baseline supervised model is trained using the limited labeled data and in the second stage, Pseudo-Labeling is used to improve the model fitting using unlabeled data.

In the first stage, the model is trained in supervised setup following the same strategy as for supervised fine-tuning (section 3.1). Figure 2 shows the model training in the initial step, on purely labeled data. The trained models using the first approach of supervised fine tuning could be utilized for implementation of this methodology. Model trained in this first stage is then used to initialize Teacher and Student models in second stage of the implementation.

The second stage (figure 4) is similar to the Mean Teacher fine-tuning (section 3.2) but here no labeled data is used, as all the data given to the Teacher and Student models is unlabeled. The Teacher provides predictions for its respective input sequence and the hard labels generated by Teacher model are then utilized as pseudo labels for the Student model to train on. The classification cost is only calculated for the tokens that are above a certain confidence threshold. In case of a two-class classification problem, the model will predict probabilities for both the classes for each token, the class with the higher probability is selected as the final prediction and the higher probability value is the confidence of prediction. Therefore confidence is the

Labeled	Unlabeled	Validation	Test
250	1750	500	1000
500	1500	500	1000
750	1250	500	1000
1000	1000	500	1000
1250	750	500	1000
1500	500	500	1000
1750	250	500	1000

Table 1: Labeled/Unlabeled split

probability with which the model predicted the label for the token. This high confidence tokens selection ensures the Student model to fit the tokens with high confidence better and thereby improves the robustness of the model for low confidence tokens.

4 Experiment Setup

4.1 Dataset

The experiments are performed using the dataset of Fomicheva et al. (2022) provided by the WMT 2021 shared task (Specia et al., 2021). The original dataset consisted of 7000 train, 1000 test and 1000 dev data for all language pairs. From that dataset, in order to simulate a low-resource setting, we sampled 2000 training sentence pairs along with 500 validation and 1000 test sentences to train the models for Mean Teacher, Pseudo-Labeling, and supervised set ups and evaluate their performances. The ratio of labeled and unlabeled data was varied keeping the amount of training sentences fixed at 2000 as shown in Table 1, in order to test the performance of SSL methods under different ratios, with the labeled data gradually increasing between 250 and 1750 sentences. For each given ratio in the table 1, a supervised model was trained on the number of labeled samples mentioned for the ratio, and SSL models were trained using the same labeled data and additional unlabeled data. The performance metrics for each model in the experiments were evaluated on the fixed 1000 test dataset provided in Fomicheva et al. (2022). In all cases, one joined model was trained including all language pairs of the dataset.

The supervised models are shown as a baseline for SSL methods using the same amount of labeled data. The performance of both the supervised and SSL models was compared in order to check if SSL algorithms provided better performance due to the presence of additional unlabeled data while training.

4.2 Model implementation

The experiments were performed with $\text{XLMRoBERTa}_{\text{Base}}$ by adding a feed-forward layer on top of the model.[1] For model training, AWS Sagemaker is used. The model is fine-tuned with early stopping on the evaluation metric on validation data. It is trained in batches and while training, the loss is calculated using weighted binary cross-entropy (Ho and Wookey, 2019) loss to tackle the issue of the imbalanced dataset in our case. The hyperparameters where initiated based on previous research involving LMs, and were optimized after multiple preliminary experiments to the ones shown in table 2. The Loss ratio (r) was found best to have the rampup value from 0 to 1 on steps. The ratio was kept very low in the begining of the training so that the models could adjust the weights according to actual labeled data provided and the loss of unlabeled

[1]The code and the data of the experiment are be publicly available with an open source license at `https://github.com/DFKI-NLP/semisupervised-mt-qe`

Hyper parameter	Values
Classification cost ($C(\theta)$)	Weighted Binary Cross Entropy
Batch Size	8
Learning rate	$2e-5$
Dropout	0.3
Optimizer	Adam
Consistency cost ($J(\theta)$)	Mean Squared Error
Max length	128
Epochs	25
Early stopping	8
Loss Ratio (r)	Rampup from 0 to 1.0 till 2 epochs (on steps)
Alpha (α)	0.99

Table 2: Hyperparameter Details

data have almost no contribution in the begining of the learning steps. This value is ramped up till the number of steps involved in two epochs. A reason for choosing the rampup period till two epochs was that LMs usually need around two epochs to fine tune for any problem. The maximum value of ratio after rampup is set to 1 as higher values resulted into large deviations of the learned weights and sudden increase in the validation errors. In order to determine the value of alpha (α), that controls the amount of weights being transferred to Teacher models from the Student models, various experiments were performed. The rampup of the EMA decay, as suggested in previous works related to computer vision (Tarvainen and Valpola, 2017; Laine and Aila, 2017) did not lead to good performance for our problem and hence we tried to determine the value of the parameter by testing the values from the set [0.99, 0.995, 0.999], concluding that the value of 0.99 performed relatively best amongst the values experimented and also gave consistent results.

4.3 Training strategies

For each given ratio of labeled/unlabeled data in table 1, models were trained with these strategies:

Supervised is the model trained on the amount of labeled data in a fully supervised fashion as described in 3.1. For example, for labeled data 250, the Supervised model is trained on 250 labeled data, and performance metrics of the model are calculated on the fixed 1000 test dataset. So, one supervised model was trained for each set of ratio labeled/unlabeled data mentioned in the table 1.

Mean Teacher: Teacher & Student are trained using the Mean Teacher network (Section 3.2). For each amount of labeled data, one Teacher and one Student model is trained. Apart from the labeled data in the given ratio, the rest of the data is utilized as unlabeled data, which is used in training the models with the Mean Teacher approach. The performance metrics for the models trained by utilizing the different ratios of labeled and unlabeled data are reported in the table with learning strategies as Mean Teacher Teacher and Mean Teacher Student. So, two models were generated for each ratio of labeled/unlabeled data by using this SSL strategy of fine-tuning.

Mean Teacher with Pseudo-Labeling: Teacher & Student are trained using the Pseudo-Labeling network (Section 3.3). For each amount of labeled data, one Teacher and one Student model is trained. Apart from the labeled data in the given ratio, the rest of the data is used as unlabeled data, for training the models with the Pseudo-Labeling approach. So, two models

lab'd	supervised	Student	Teacher	relative improvement (%)	
				Student	Teacher(%)
250	0.252	0.280	***0.283**	11.11	12.30
500	0.288	0.299	***0.300**	3.82	4.17
750	0.313	**0.317**	0.312	1.28	0.00
1000	0.320	0.344	***0.346**	7.50	8.13
1250	0.335	0.340	**0.344**	1.49	2.69
1500	0.333	***0.350**	***0.350**	5.11	5.11
1750	0.328	0.355	***0.361**	8.23	10.06

Table 3: MCC scores for Supervised and Mean Teacher experiments; * indicates significantly better scores based on bootstrap re-sampling, as compared to the supervised baseline

were generated for each ratio of labeled/unlabeled data by using this SSL strategy of fine-tuning. We repeated the experiments with confidence thresholds of 0,6 and 0,8, and the latter was chosen due to the higher performance. Additionally, we repeated the experiments without a consistency cost, but results are not reported, as no significant difference was observed.

4.4 Evaluation

For evaluating the systems generated with fully supervised approach or SSL approaches, the metric used is Matthews correlation coefficient (MCC; Matthews, 1975), as per WMT (Zerva et al., 2022) along with F1-scores for OK/BAD classes. In the first part of our experiments, contrary to WMT calculating MCC scores for source, target and gap tokens, we focused on the MCC score for the whole sequence, to ensure that our models can produce good labels for all the tokens of the sequence, as MCC for whole sequence consolidates classification and misclassification errors for all the tokens. In the second part of our experiments, we present disjoint MCC results, following the official WMT calculation.

In order to test the significance of the results with the Mean Teacher, we tested these models using paired bootstrap resampling method (Koehn, 2004). For this, 250 sentence sequences were sampled out of 1000 test dataset with replacement to form 100 virtual test sets of 250 sentences each.

5 Results

5.1 Mean Teacher fine-tuning

The performance of models trained with Mean Teacher vs. supervised learning are shown in table 3. Teacher models outperform the Student and supervised models significantly for every ratio of labeled to unlabeled data, apart from two cases where they don't show a significant improvement. In the best case, where a very little amount of training data is available, the Teacher model gives a relative improvement of 12.3% over the supervised baseline. It is also noticed that the average relative improvement for all experiments with different ratios of labeled/unlabeled data is approximately 6% for Teacher and 5.5% for Student models. Confirming previous work (Tarvainen and Valpola, 2017), the Teacher is more robust and performs better than the Student after certain iterations of training.

5.2 Pseudo-Labeling

As seen in Table 4, the approach of Pseudo-Labeling gave small improvements for some experiments but for most experiments it didn't perform as expected. There could be several reasons for this. One of them is models suffer from confirmation bias (Arazo et al., 2020), i.e mod-

lab'd	supervised	Student	Teacher	relative improvement (%)	
				Student	Teacher(%)
250	0.250	0.280	**0.283**	12.0	13.20
500	0.288	0.287	0.287	0.0	0.00
750	0.313	0.294	0.302	0.0	0.00
1000	0.320	0.306	0.310	0.0	0.00
1250	0.335	0.335	**0.337**	0.0	0.60
1500	0.333	0.314	0.331	0.0	0.00
1750	0.328	0.327	**0.332**	0.0	1.22

Table 4: MCC scores for Pseudo-Labeling

els relying on its own predictions. Additionally, despite experimenting with various confidence thresholds and the consistency cost, we generally used the same hyperparameters as in the Mean Teacher setup, so it is not possible to exclude the case that better results occur after a broader hyperparameter search.

5.3 Disjoint comparative analysis

More detailed results, following the disjoined calculations of all metrics as per WMT can be seen in Table 5. Here we present the MCC and the F1-scores for BAD/OK labels, measured for the source and target sentence with and without gaps, for every ratio of labeled/unlabeled data. It can be seen that in all cases, the MCC score and and the F1 score for BAD labels outperform the ones of the supervised baseline. In some cases there is no improvement shown for the F1 score for OK labels, but one should consider that the amount of OK labels in the dataset is overly high, and the F1 score is affected by the big amount of true positives.

6 Conclusion

This research focused on the Quality Estimation of Machine Translation at the word level. The goal is to generate a binary label of OK/BAD for each word and gap in the translations, by predicting if the word is correctly translated or not. We investigated two approaches of Semi-Supervised Learining that have not been explored yet for the given problem: The first utilized the well-known Mean Teacher approach that involves a Student and a Teacher model while training, initialized with the default weights of a pretrained LM. The second proposed architecture extends the former, by involving Pseudo-Labeling and follows a two-stage learning approach. In the first stage, the model is trained with limited labeled data available, through supervised learning. In the second stage, the Teacher and Student model are initialized with the model learned in the first stage, and are further trained using only unlabeled data.

It was experimentally shown that in low-resource settings the Mean Teacher architecture performed better or (in one case) comparably to the supervised models, achieving an improvement of up to 12%. The second proposed architecture of using Pseudo-Labeling with Mean Teacher framework did not behave as expected, when tested with various values of thresholds. Further work could focus on the implication of the improvements on various language pairs, as well as architectural improvements and data augmentation techniques.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the project TextQ, and by the German Federal Ministry of Education and Research (BMBF) through the project SocialWear (grant no. 01IW20002).

items	model_words	MCC	F1 BAD	F1 OK
250	supervised source	0.207	0.331	0.874
	supervised target and gaps	0.282	0.374	0.906
	supervised target	0.240	0.391	0.845
	Mean Teacher source	**0.240**	**0.373**	0.828
	Mean Teacher target and gaps	**0.309**	**0.384**	0.804
	Mean Teacher target	**0.248**	**0.411**	0.759
500	supervised source	0.240	0.373	0.811
	supervised target and gaps	0.319	0.401	0.840
	supervised target	0.252	0.411	0.703
	Mean Teacher source	**0.249**	**0.379**	0.816
	Mean Teacher target and gaps	**0.325**	**0.413**	0.909
	Mean Teacher target	**0.276**	**0.430**	0.768
750	supervised source	0.267	0.393	0.826
	supervised target and gaps	0.341	0.427	0.878
	supervised target	0.289	0.440	0.785
	Mean Teacher source	**0.276**	**0.399**	0.826
	Mean Teacher target and gaps	**0.343**	**0.427**	0.875
	Mean Teacher target	**0.291**	**0.440**	0.780
1000	supervised source	0.278	0.393	0.773
	supervised target and gaps	0.345	0.423	0.854
	supervised target	0.288	0.434	0.736
	Mean Teacher source	**0.300**	**0.418**	0.858
	Mean Teacher target and gaps	**0.375**	**0.458**	0.899
	Mean Teacher target	**0.336**	**0.473**	0.826
1250	supervised source	0.295	0.414	0.858
	supervised target and gaps	0.360	0.445	0.903
	supervised target	0.323	0.463	0.839
	Mean Teacher source	**0.304**	**0.421**	0.863
	Mean Teacher target and gaps	**0.369**	**0.453**	0.902
	Mean Teacher target	**0.330**	**0.469**	0.834
1500	supervised source	0.291	0.403	0.782
	supervised target and gaps	0.354	0.433	0.858
	supervised target	0.305	0.445	0.742
	Mean Teacher source	**0.312**	**0.427**	0.854
	Mean Teacher target and gaps	**0.374**	**0.456**	0.898
	Mean Teacher target	**0.334**	**0.472**	0.826
1750	supervised source	0.288	0.407	0.820
	supervised target and gaps	0.347	0.352	0.435
	supervised target	0.304	0.446	0.786
	Mean Teacher source	**0.320**	**0.433**	0.851
	Mean Teacher target and gaps	**0.387**	**0.466**	0.895
	Mean Teacher target	**0.347**	**0.480**	0.819

Table 5: Comparative analysis of Supervised and MT models on disjoint performance of tokens in source and target sentence.

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Fomicheva, M., Sun, S., Fonseca, E., Zerva, C., Blain, F., Chaudhary, V., Guzmán, F., Lopatina, N., Specia, L., and Martins, A. F. T. (2022). MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V., and Specia, L. (2020). Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Ho, Y. and Wookey, S. (2019). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, 8:4806–4813.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Laine, S. and Aila, T. (2017). Temporal ensembling for semi-supervised learning. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., and Zhang, C. (2020). Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1054–1064.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Ranasinghe, T., Orasan, C., and Mitkov, R. (2021). An Exploratory Analysis of Multilingual Word-Level Quality Estimation with Cross-Lingual Transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 434–440, Online. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., C. de Souza, J. G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., and Martins, A. F. T. (2022). CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645, Abu Dhabi. Association for Computational Linguistics.

Specia, L., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V., and Martins, A. (2021). Findings of the wmt 2021 shared task on quality estimation. Association for Computational Linguistics.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Tuan, Y.-L., El-Kishky, A., Renduchintala, A., Chaudhary, V., Guzmán, F., and Specia, L. (2021). Quality Estimation without Human-labeled Data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 619–625, Online. Association for Computational Linguistics.

Wang, Y., Mao, Q., Liu, J., Jiang, W., Zhu, H., and Li, J. (2022). Noise-injected consistency training and entropy-constrained pseudo labeling for semi-supervised extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6447–6456, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zerva, C., Blain, F., Rei, R., Lertvittayakumjorn, P., C. de Souza, J. G., Eger, S., Kanojia, D., Alves, D., Orăsan, C., Fomicheva, M., Martins, A. F. T., and Specia, L. (2022). Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhang, D. and Yang, Z. (2018). Word embedding perturbation for sentence classification. *arXiv preprint arXiv:1804.08166*.