Transfer Learning for Machine Translation Quality Estimation

Ganesh Murugappan Alexander Hobmeier David Gordon Cameron Benett
Georgia Institute of Technology
{ganeshm, ahobmeier3, dgordon48, cbennett49}@gatech.edu

Abstract

In a world of increasing smart technology, machine translation is a valuable task taken on to break the barriers between languages. There has been a great deal of progress in this task which can be seen at the Workshop for Machine Translation competition (WMT). With so many new architectures popping up and taking on the task of machine translation, there's been an increased concern for handling quality control of textual translations for models. Our paper seeks to develop a means for generating good quality estimations of machine translations and then gauging the effects of different pretraining combinations on it's performance. We do design a quality estimation system utilizing a transformer for feature extraction followed by a score generation model. We also demonstrate that translation specific data might be most optimal for Quality Estimation training and BERT reigns supreme as the best transformer for this task.

1. Introduction

The goal of this paper is to utilize prior knowledge of the pre-trained multilingual embedding model BERT in order to gain an understanding of whether quality estimation of machine translations between both English-German and English-Chinese performs better or worse when preceded by some training on similar data. We tested model architectures including simple LSTMs and linearly sequenced sets of dense layers with the objective of determining a single translation quality score ranging from 0 to 100. This score was then compared to the ground truth mean provided by the MultiLingual Quality Estimation dataset.

Until 2018, winners of the Workshop for Machine Translation (WMT) quality estimation competition have released papers describing their tactics employed to gain advantage in two tasks (sentence-level estimation and word-token scoring). We focus on the sentence-level estimation task. The winners of the 2018 WMT competition from Alibaba decomposed the task into two subtasks: feature extraction and score generation. The Alibaba submission utilized the features within a "Bilingual Expert Model" for the feature extraction portion of their submission. The quality scoring portion was handled through an ensemble of multi-layered perceptron (MLP) models. The primary existing limitation appears to be the size of the attention transformers within the Bilingual Expert Model.

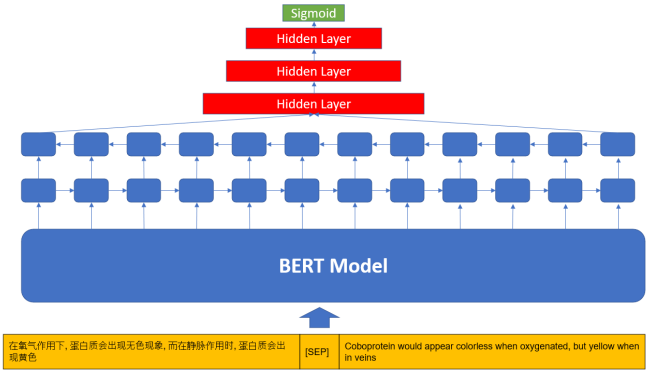
Training models with modern architectures in natural language is an extremely computationally demanding task. This introduces a barrier to entry and prevents individuals and teams without access to large amounts of computational power and data from trying their hand at quality estimation. Transfer learning alleviates this issue by reducing the need for large sets of labelled data and reducing the amount of computation needed to produce a working model. By showing that transfer learning using pre-trained embeddings from BERT is successful in quality estimation, we open the door for others to tackle the problem.

The data utilized is derived from the MultiLingual Quality Estimation dataset and is publicly accessible from Facebook Research. The dataset consists of several language pairs, including training and development sets for Estonian-English, Romanian-English, and English-Nepali and development and test sets for English-German and English-Chinese. The data is formatted as tab-separated values and includes the original and translated sentences along with information such as ground truth quality scores from at least 3 human annotators and log-probability scores associated with each output token.

2. Approach

The quality estimation model was broken down into two main parts: a predictor and an estimator. The predictor has the responsibility of extracting features from the source and translation, whereas the estimator is tasked with using the features from the predictor to estimate the quality of the translation. Our predictor begins with a tokenizer which encodes the source and target sentences with a separation token between them. The tokens are fed into a pre-trained version of multilingual BERT; the final hidden states are taken as embeddings. The parameters of the pre-trained multilingual BERT model were frozen during the training

process to preserve its role of producing embeddings of the input. The estimator consists of a bidirectional recurrent neural network as well as a series of four linear layers and activations between. The embeddings from BERT are taken as input to this estimator, and it in turn outputs a single number, which is normalized to be between 0 and 100 by applying a scaled sigmoid function.



When designing the architectures to be tested, we anticipated difficulty in selecting transformer-based models for feature extraction. Fortunately, our first choice of BERT showed promise in the initial experiments. Another potential source of issues that we anticipated was the method by which the input sentences would be fed into the model of our choice. This was made trivial by the BERTTokenizer module provided by the Hugging Face transformers package; the tokenized versions of the sentences were fed directly into the pre-trained language model itself.

One issue we did struggle with during the data processing stage was the several ways by which the data was malformed. The data was presented as tab-separated values, but after performing standard parsing on the raw files, there were several sentences that appeared to be incorrectly encoded. In order to mitigate this issue, we introduced several steps in order to make sure that data that was entered into the model was in the proper form.

Another problem we encountered was the unbalanced nature of the provided translation datasets. The sets are largely skewed towards high quality translations, particularly for the English-German and English-Ghinese datasets. This often caused the model to get stuck in local minima where it would guess the same relatively good score for all translation pairs. We attempted to avoid this tendency by including the other languages, which have notably worse translations on average. Between a more balanced distribution and a larger dataset to train on, adding the other datasets for pre-training improved performance dramatically. Another way we dealt with this issue was to increase the batch size, with the rationale that with imbalanced data, larger batches had better chances of including a more diverse range of scores. This helped increase performance but also caused other issues that we had to troubleshoot; at one point the combination of padding all sequences to the same length and the increased batch size, along with the large number of parameters from the LSTM and Dense Linear layers caused us to crash due to memory limits. We were able to resolve this by padding batches dynamically, reducing LSTM hidden size, and manually removing malformed extra-long sequences from the datasets. This allowed us to test larger batch sizes, which did perform marginally better on the test sets.

3. Experiments and Results

Because our model needs to approximate a single regression value, we were able to measure success utilizing a simple mean squared error loss function between our ground truth values and our output. The experiments we ran seek to understand what combination of training set pairs and transformers produce the most accurate quality estimation scores on our ground truth set. We broke our experimentation section into 2 sections in order to gauge whether quality estimation improved with the use of different training data or with the application of different transformer models.

3.1. Experimenting with different pretraining

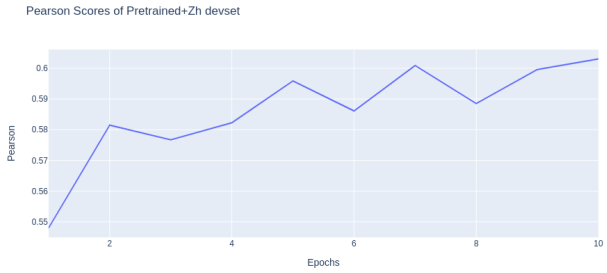
Our first set of experiments involved us utilizing one of 2 training dataframes and gauging performance on the test sets of 2 target language pairs (En-De or En-Zh). The first training dataframe consists of data from an English-Nepali, English-Romanian, English-Estonian and either the English-German or English-Chinese development set provided by the MultiLingual Quality Estimation dataset. The second dataframe consists of just the training dataset associated with either the English-German data or English-Chinese data. Our intention with this experiment was to see whether quality estimation of a single translation pair is greatly improved when utilizing translation specific data for training or if we are able to generalize our quality estimations between multiple languages by training on multiple unrelated language pairs. Below we have a table describing some of the results evident measured by our Pearson coefficient metric and RMSE loss function.

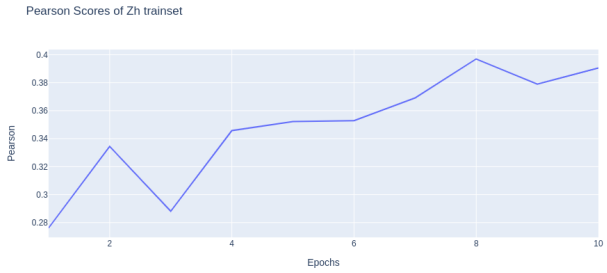
It should be noted that the Pearson coefficient generated in the table is a correlation score between the quality score outputs generated by our model and the ground truth labels we seek to replicate. Due to a lack of resources and time we were only able to get one trial per combination of variables though we did train with a large batch size of 128 for 10 epochs per experiment. Thus we feel confident that our correlation results are indicative of something meaningful. Upon initial glance it would appear that training on the English-Chinese training set produces the best results

	Varied Dev+En-De dev	En-De train	Varied Dev+En-Zh dev	En-Zh train
En-De Final Pearson Score	0.325	0.376	0.319	0.349
En-Zh Final Pearson Score	0.286	0.228	0.281	0.394
En-De Final RMSE Score	302.860	184.675	N/A	N/A
En-Zh Final RMSE Score	N/A	N/A	200.197	181.582

for any target language along with the second best score for any non-related language. These results oppose the intuition we initially developed which stated that we should expect the best Pearson score from a trial in which we train on a more varied dataset. Not only did training on the chinese training set generate the best Pearson score for a target pair but we also see it narrowly generated the best RMSE score as well. Why might this be?

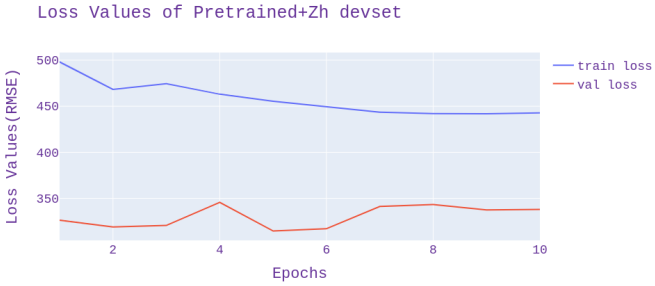
Upon further inspection, we noticed that the trials utilizing the varied development sets for training seemed to nearly overfit their data in a way that was not useful or indicative of performance on the desired target sets (English-German or English-Chinese). For example, the plots below display the Pearson score over time for two variants of training in preparation for the English-Chinese testing dataset.

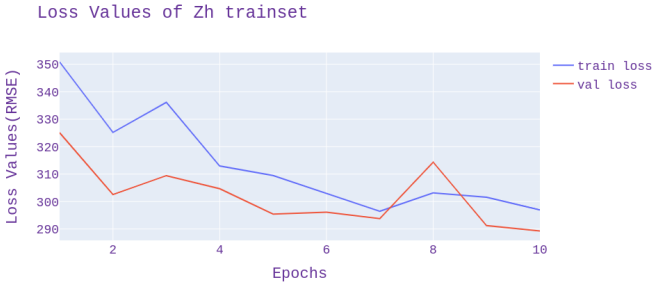




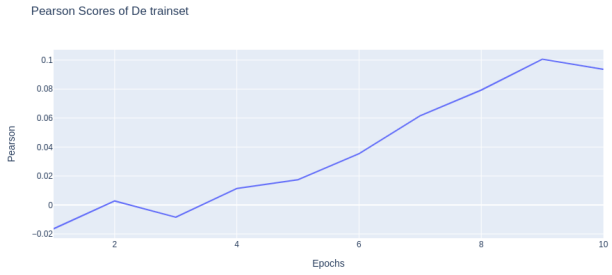
The final Pearson score generated for the Pretrained+Zh devset trial was barely half the value seen while training. This forced us to speculate on the possibilities of overfitting our data and possibly utilizing a larger combined set of data for a rerun. However, we then took a look at the plots for our loss values over time and saw that overfitting was not quite the issue.

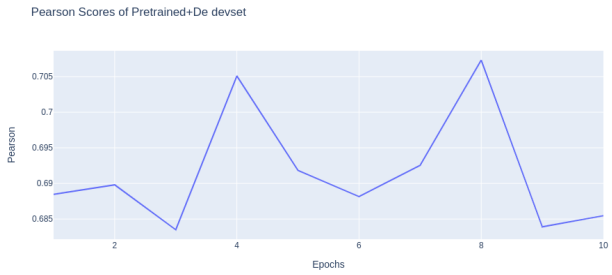
As we can see from the visualizations of our loss (below), our RMSE for the Pretrained+Zh devset was not nearly seeing as much convergence as the training on our Zh trainset. The stagnating loss curve indicated to us that overfitting was not the problem. The distance between the two loss curves does hint at the possibility for our experiment revealing that the chosen dataset is completely unrelated to the target language pair. We did experiment with varied combinations of English-Romania, English-Estonia and English-Nepali datasets but none produced Pearson scores above 0.15. If given more time it certainly would be of interest to identify more translation pairs to see if any can generate better loss curves than the original trainset.



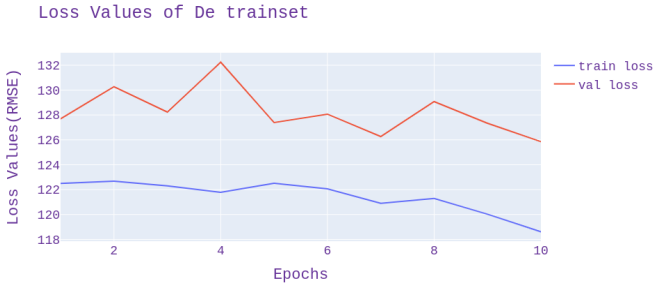


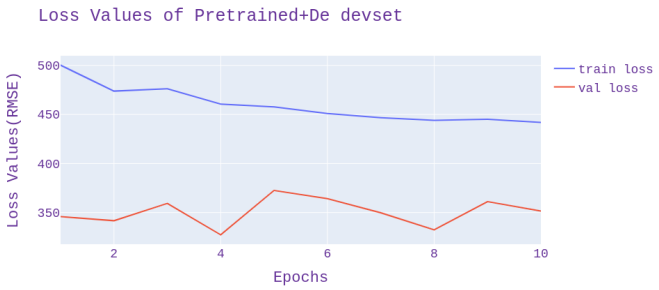
Some additionally interesting results that presented themselves can be found in the En-De trainset experiment. For some reason our validation accuracy for this experiment was consistently below 0.1 before arriving at a final accuracy above 0.3. This particular trial was ran 3 times to verify that this odd occurrence wasn't simply random chance. Subsequent repetitions did follow the same trend of having a validation accuracy consistently below 0.1 while generating a final Pearson score above 0.3. Below is the plot of our En-De Pearson score over time along with the loss plot for the same trial.





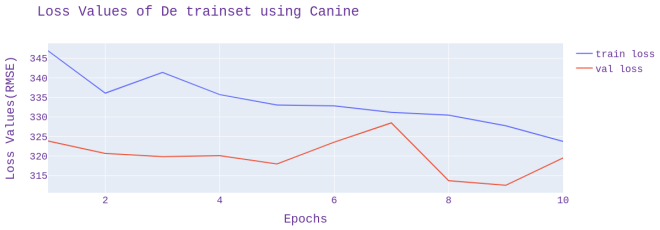
These results are not at all similar to the ones generated by the pretrained+De trials. The validation accuracy was consistently above 0.5 throughout the training process for the pretrain+En-De devset. A similar anomaly can be found in looking at the loss curves for each experiment trial. We notice that train loss is much lower for the experiment utilizing the desired trainset hinting at the notion that utilizing translation specific training datasets is more effective for generating accurate quality estimation than utilizing a mix.

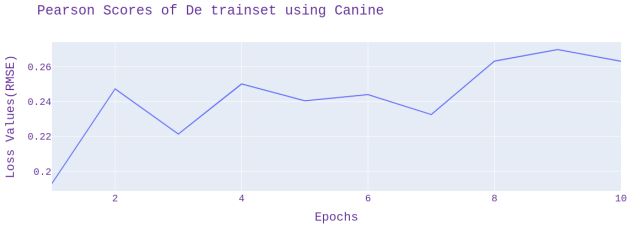




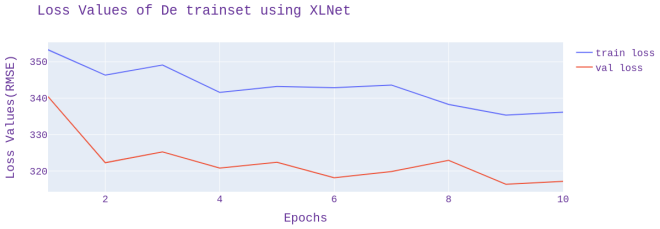
3.2. Experimenting with different transformers

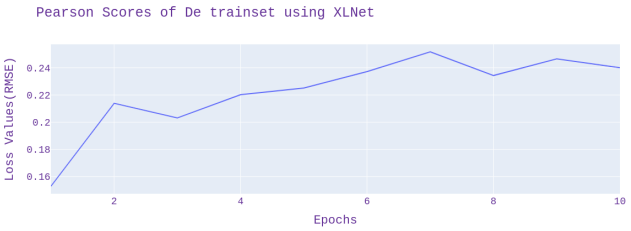
In an attempt to gauge whether our model could produce better Pearson scores given the base training set and test set, we ran 4 experiments in which we swapped out our default BERT model with a different transformer and trained to see how much better or worse our Pearson scores became. We utilized the English-Chinese translation set in order to avoid the anomaly found in training the English-German with BERT. The first model we tested with uses the Canine transformer which is similar to BERT except that it utilizes a character level tokenizer rather than word-level. It's introductory paper suggests that this tokenization process generates a finer-grained feature space that is more efficient for training. However upon application, it seems that this unicode tokenization doesn't provide many benefits to the final Pearson score output. It's final Pearson score ended up being 0.242 for the target data-set which is much less than the baseline score of 0.394 utilizing the base version of BERT.



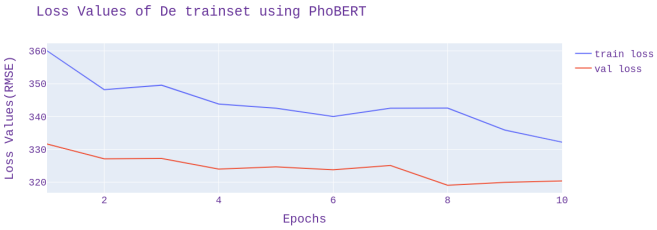


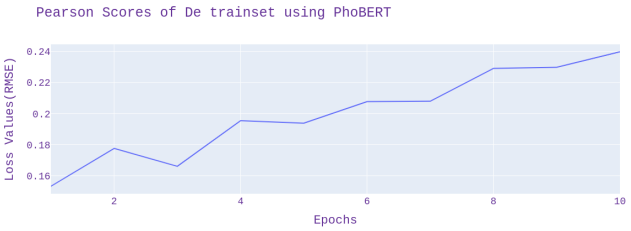
The second transformer we tried was the XLNet transformer commonly found in most of the same tasks as BERT. XLNet takes advantage of Autoregressive and AutoEncoding methods in order to overcome pretrain-finetune discrepancy making it a more effective transformer for most tasks including question answering, natural language inference, sentiment analysis, and document ranking. When placed into our existing model, the XLNet transformer surprised us by performing much worse than both the baseline BERT model and the Canine transformer after having consistently lower Pearson scores throughout the same 10 epochs as the other trials. Below we can see the plot of the Pearson scores and loss values. This model ended up with a final Pearson score of 0.217 for it's target test set.





Finally, we took a look at a recently popularized monolingual model PhoBERT which is pretrained on solely Vietnamese data. The reason for selecting this transformer is because the introductory paper associated with it claims that it outperformed a similarly popular multilingual model that we wished to utilize XLM-R. PhoBERT is supposed to be particularly powerful for Part-of-speech tagging, Dependency parsing, Named-entity recognition and Natural language inference. We attempted to see if Machine Translation Quality Estimation could be added to that list. The results we collected utilizing this transformer hint that it certainly should. PhoBERT was narrowly the best performing alternative to BERT with a final score of 0.245.





Though this is somewhat exciting, the majority of this experiment has indicated that BERT is by far the best performing transformer to utilize for Quality Estimation tasking. Additionally, none of the transformer models generated good scores for non-target translation pairs above 0.2.

4. Conclusion and Future Work

The work we've demonstrated has established some strong implications around desirable training sets and transformers to be used for effective Quality Estimation of Machine Translation. Our extensive training and large batch sizes affirms some of the ideas we've presented and we would like to further evolve some of these experiments in order to gain more understanding behind the trends that have presented themselves. Machine translation is certainly a complex task requiring many resources but if quality estimation of that task can be done with fewer resources then there would be a great deal of increased potential in evolving that space in a discernible direction.
Some tasks we could see ourselves improving upon in the future would be increasing our number of trials in order to validate some of the anomalous trends found in our experiments. Another thing we might adjust in future work would be the utilization of more translation pairs from east Asian languages. We were supplied with a great deal of English to European translation data which might have presented an implicit bias against our results measured with our English-Chinese data-set. We hoped that PhoBERT's success was an indicator of something along those lines. Finally, if given an opportunity to further this research we would attempt to design more intricate estimator systems which rely on more than an LSTM layer paired with some linear layers. It would be exciting to explore the space of Recurrent Layers that can be applied to this task.

5. References

Clark, J. H., Garrette, D., Turc, I., Wieting, J. (2021). CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. Arxiv.org. https://doi.org/10.48550/arXiv.2103.06874

Deng, Y., Cheng, S., Lu, J., Song, K., Wang, J., Wu, S., Yao, L., Zhang, G., Zhang, H., Zhang, P., Zhu, C., Chen, B. (2018, October 1). Alibaba's Neural Machine Translation Systems for WMT18. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6408

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv.org. https://arxiv.org/abs/1810.04805

Nguyen, D. Q., Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. ArXiv:2003.00744 [Cs], 3. https://arxiv.org/abs/2003.00744

Wang, J., Wang, K., Chen, B., Zhao, Y., Luo, W., Zhang, Y. (2021). QEMind: Alibaba's Submission to the WMT21 Quality Estimation Shared Task. Arxiv.org. https://doi.org/10.48550/arXiv.2112.14890

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. ArXiv.org. https://arxiv.org/abs/1906.08237