Transfer Learning for Neural Machine Translation in the Thai-English Language Pairing

Alex Kihiczak
University of California, Berkeley
akihiczak@berkeley.edu

Isaac Vernon
University of California, Berkeley
isaacvernon@berkeley.edu

Abstract

The abundance of robust pre-trained models has afforded the opportunity for low budget investigations of fine-tuning techniques in Neural Machine Translation (NMT). Because of this, we use a pre-trained model, which employs an encoder/decoder generative architecture (the current NMT standard), and fine-tune it in order to study transfer learning for Thai to English translation. Specifically, we explore domain adaptation as a strategy for more efficient fine-tuning, a novel strategy in Thai-to-English NMT. We find that in low resource training, we don't make huge gains in BLEU score from domain-specificity as our generally trained model performs comparably to the fitted models across domains with a little dip in performance in the more irregular domains (web). We also found that generalizing to new domains seemed to be helped by irregular examples. The highest performance on the product reviews and the verbal data was from Paracrawl, the most irregular dataset, followed by the generally-trained broader model (which included some irregularities from the Paracrawl examples). However, this may be due to syntactic structure rather than the irregularity. Finally, we propose a possible improvement for low-resource situations with weighted domain learning.

1 Introduction

Large pre-trained language models such as BERT, RoBERTa, BART, etc., are increasingly being used as the starting point for building and fine-tuning a state-of-the-art NLP model. However, this practice is still rather novel for sequence-to-sequence tasks (Rothe et al., 2020), in particular for Neural Machine Translation (NMT).

One area of research for fine-tuning and utilizing these large, pre-trained models is transfer learning. Transfer learning is a strategy at the forefront of NMT, and for good reason. It is more efficient to be able to fine-tune a model once and transfer that learning to other domains than to have to fine-tune for every new domain. Furthermore, due to limitations of budget and time, adapting a pre-trained model can provide a feasible means to get hands-on with exploring the practical improvements to current NMT base models. Due to this, we are able to investigate the possibilities and limitations of transfer learning.

Among the challenges currently present for NMT, transfer learning is best situated to help resolve the issue of domain specificity. As explicated by Koehn et al. in their 2017 paper, which explains the challenges faced by NMT at the time, "a known challenge in translation is that in different domains, words have different translations and meaning is expressed in different styles" (Koehn and Knowles, 2017). This challenge is still present four years after they published their paper, and therefore it is an important area for further study.

In order to explore the efficacy of transfer learning for NMT, we chose to utilize a Thai-English dataset of paired sentences, which is broken up by source into multiple distinct domains (Lowphansirikul et al., 2020). The domain divisions make the data well poised for investigating the category of domain adaptation transfer learning. Previous research has found cross-lingual transfer learning, using learning from related languages to improve performance on a low resourced language, to be a valuable strategy when translating from a high resourced language to a low resourced language. Although Thai is a relatively low resourced language (Lowphansirikul et al., 2021), we chose not to explore this area of transfer learning as Lowphansirikul, Lalita, et al., have recently pre-trained a model, using this dataset in conjunction with 78GB of other Thai-English data to mitigate the issue of low resourced Thai translation (Lowphansirikul et al., 2021). Thus, we assess the cross-domain efficacy of fine-tuning a related pre-trained multilingual model on specific domains of their dataset.

2 Background

2.1 Current Machine Translation Approaches

Until recently, the most popular method for performing machine translation was phrase-based statistical machine translation. This was eventually followed by sequence modeling approaches using Recurrent Neural Networks (RNN's) and then the RNN variant LSTMs (Long Short Term Memory model) (Vaswani et al., 2017). However, these models could not escape the constraints of a sequence model, specifically on the context window and sequential relevance. The field of machine translation welcomed a new paradigm with the creation of the transformer architecture, a model that does away with recurrence and focuses solely on attention mechanisms to model dependencies (Vaswani et al., 2017). This new type of model has become the gold standard for NMT, leading to large improvements in translation performance ever since it was introduced. Within a transformer model setup for machine translation there are two parts: an encoder and a decoder. Utilizing multi-head attention models as well as feed-forward layers, the model can be fit to produce a softmax output for a specific model input; for machine translation this will be the start sentence token or the softmax predicted output of the previous step.

2.2 Transfer Learning

One of the fields of work within the Neural Machine Translation scope is transfer learning. Transfer learning is the idea of extracting knowledge from a source setting and applying it to a different target setting (Ruder, 2019). Transfer learning can be roughly categorized into four types based on different combinations of tasks and labeled data: domain-adaptation, cross-lingual learning, multi-task learning, and sequential transfer learning. There are also variants in how the base models are trained. Some researchers use large general corpuses in their model's pre-training task, while others train an initial model on a much smaller training set from a specific domain before their fine-tuning (Cui et al., 2019). When choosing a pre-trained model, we also needed to consider the pre-training task and the target task. In our case, our target was translation. We also had to take into account cross-lingual pre-training as we needed a model trained on both English and Thai (Ruder, 2019).

2.3 Thai-English

Unlike many popular languages used in machine translation (English-French, English-German, and Chinese-English) with large corpuses and support (often referred to as "high-resource language pairs"), Thai-English is a low-resource language pair. This is a problem which can significantly impact performance of models as there are not a large number of training examples (Lowphansirikul et al., 2020). One of the challenges in creating the Thai-English translation pairs is that there are no clear sentence boundaries in Thai, so often the English delineations are used as a proxy. Before the creation of the dataset we used (2020), there were only a couple of sources for Thai-English pairs, with the largest and most prominent being OPUS (Open Parallel Corpus), though the examples from there are limited to subtitles, religious texts, and open source documentation. Overall, the field of English-Thai translations is very minimally covered (Vis, 2020 is the only example), and there has not been any prior work on transfer learning within this language pair.

2.4 Models, Architecture, and Evaluation

BERT and GPT are the models that are used as the backbone for our NMT architecture. The first piece of our translation architecture is an encoder via BERT (Bidirectional Encoder Representations from Transformers). BERT pretrains bidirectional representations from text by conditioning on both sides of the context in all the transformer layers. Essentially, BERT as an encoder can return context embedded word vectors for each input (Zhu et al., 2020) with improvements over the prior LSTMs that could only view context from a single direction – prior work had used a concatenation of left-to-right and right-to-left trained LSTMs (Devlin et al., 2019). The other framework that we use is GPT (Generative Pre Training). OpenAI's GPT has the same architecture as BERT, though it instead focuses on predicting the next word given the last input window (Radford et al., 2018). We used this GPT format as a decoder in our framework for the generative task of NMT. This is in lieu of configuring a BERT architecture to perform the generative decoding task. The combination of

BERT as an encoder and a GPT architecture as the decoder is the foundation of the BART infrastructure (Lewis et al., 2019), a variant of which we use as our model. However, due to resource constraints, we chose a pre-trained variant of the model that we can fine-tune for our task. Therefore, we use mBart50 (Tang et al., 2020) as the basis for our work as the 50 languages in the multilingual translation model include both English and Thai.

Finally, we use the classical measure of performance in machine translation to measure our models. This is BLEU score, an empirical precision-based measure correlated with human translators' judgements that balances the linguistics of the translation along with the context that remains (Papineni et al., 2002).

3 Methods

We are working on transfer learning in a Thai-English translation setting. Due to our resource constraints, we do not have the ability to pre-train and fine-tune a full model from scratch. Therefore, we have to use pre-trained models publicly available through HuggingFace and the transformers package, which we subsequently fine-tune. We attempted fine-tuning a handful of similarly structured models that shared the encoder-decoder architecture and BERT embeddings but ultimately found only one approach to be successful.

3.1 Initial Models and Approach

Our first approach to constructing this model architecture was to use multilingual BERT as both an encoder and decoder. One technique we used for achieving this was to establish the BERT encoder and decoder, with cross attention to pass the encoder hidden states to the decoder, and then fine-tune only the decoder parameters so that the decoder in effect learned how to translate the encoder's embedding of the Thai text. We explored this technique with a few pre-trained HuggingFace BERT models, and used different methods to configure similar architectures. In each case we would train on up to 10,000 samples (for two epochs) and up to 500 epochs (with eight samples). However, after training many variations of the models and training parameters, we were never able to achieve any reasonable performance, and frequently found generation tasks returning repeating punctuation after the first few tokens. This is likely due to the fact that there are a lot of new, untrained parameters being initialized in this architecture. The poor results can likely be attributed to a lack of training time, which was infeasible to increase given the sheer number of parameters to tune and our time and resource constraints. Therefore, our focus shifted from utilizing BERT as both an encoder and a decoder in our translation framework during fine-tuning for transfer learning to using a model suited for translation, which would require learning fewer parameters, and running the fine-tuning from there.

3.2 Final Model and Approach

We used a multilingual variant of the BART architecture (mBart50). This was pre- trained on a similar task (multilingual machine translation between 50 languages) to our target task (machine translation between English and Thai), which helped to facilitate the fine tuning of the model between domains for transfer learning. Our fine-tuning was fitting models on the following specific domains: Paracrawl, which contains around 1,000 web pages with both English and Thai, Wikipedia articles, the 500 most popular Thai websites, and government documents, both from the Asia Pacific Defense Forum online and from requested PDF's from the Thai Government (Lowphansirikul et al., 2020). We then evaluated the performances across domains. Specifically, we fine-tuned a pre-trained mBART model per domain for a total of 7 models including the baseline models: 5 fine-tuned domain fitted models (since we are reducing the dataset to those collections that were sentence alignment segments) along with the base mBart model and another generalized model (trained on equal parts of each domain) and evaluated performance across domains.

3.3 Training

We trained each of our models for the same number of training epochs and with the same number of training examples. Due to time and resource constraints, we ran 10,000 training examples for one epoch each, which resulted in a training time of 15-16 hours per model (4vCPU 26GB Virtual Machine). The pre-trained model we used was the HuggingFace implementation of the mBart-large-50-many-to-many-mmt model, a generalized model for multilingual machine translation (Tang et al., 2020). We also used the tokenizer from the mBart model and we trained our model with the Adam optimizer as well as the other HuggingFace base training parameters for an mBart Conditional

Model/Test	APDF	Assort Gov	Wiki	Paracrawl	Websites	Task	Review
Baseline	9.40	5.95	10.90	9.03	5.02	2.87	6.53
Mixed	**37.01**	**30.00**	**32.76**	30.00	23.24	8.63	18.25
APDF	36.44	6.75	22.63	19.74	9.55	7.69	14.51
Assort Gov	15.52	10.84	17.51	14.74	5.49	3.80	10.42
Wiki	20.14	16.15	31.74	19.28	7.42	7.24	13.73
Paracrawl	22.52	13.72	20.26	**33.97**	10.87	**10.15**	**18.88**
Websites	20.47	9.85	18.11	22.67	**26.05**	7.17	14.24

Table 1: Quality of models (BLEU), when trained on one domain (rows) and tested on another domain (columns).

Generation model. We then ran our evaluation over the same test sets for each model to allow us to generate apples-to-apples comparisons for the corresponding sacreBLEU (Post, 2018) scores. Again, due to time and resource constraints we ran our generative predictions over test sets of size 100 so that we could evaluate on a wider array of test sets. Each testing cycle took around a half hour to complete with a 4vCPU 26GB Virtual Machine through GCP.

4 Results and Discussion

4.1 Data Quality

As previously mentioned, Thai is a language without natural sentence boundaries so the method used to create the pairs is dependent upon the English sentence breakpoints. However, this difference in syntactic structure could possibly lead to some issues with the pairs, though the original creators of the dataset have worked to minimize this issue (Lowphansirikul et al., 2020). In addition, there were some Thai-English examples that, when tokenized, were longer than the maximum input length (1024) for our mBart model. This means that when we train on these examples, the labels for our predictions might no longer align to the actual meaning of the segment. However, this was a small minority of our examples, and should only minimally affect our translation results.

The other issue of data quality is the variance in translation quality between different domains (Lowphansirikul et al., 2020). Depending on the source, there may be different requirements for the level of translation (a popular travel website has more incentive to provide a strong translated version than a local business). However, this should only be pertinent within two domains: Paracrawl and Thai Websites, as they are scraped from sources on the internet. An issue that we found when inspecting the data from these domains was the lack of translation for a lot of the English proper nouns (Product names, Company names, etc) or English numerals (Thai language has its own numeric characters), which could impact translation quality if this is not consistent throughout the data.

4.2 Results and Analysis

We conducted work on transfer learning within a low-computing resource setting. If you have enough computing resources you will want to train on as many examples across domains that you can. Overall, fine-tuning a model on each domain gave us gains in performance over the baseline mBart translation model (by 27 points in the best situation and still reaching a 5 point improvement in BLEU scores in the worst case). In addition, domain-fitted models still have performance exceeding the baseline mBart when evaluated across domains (though the gains are not nearly as large). However, compared to our manufactured baseline, a mixed model trained on equal examples across domains (with the same amount of total training examples as the domain-specific models), the BLEU scores are roughly comparable across all domains. Overall, if the goal is general translation, we find that a mixed model trained across relevant domains is the best choice, though given a domain specific translation task we will see the best performance with a domain specific model. This will still provide translation improvements in other domains as well, just not to the extent of a general mixed model.

We can also look at specific results. One of these was the fact that the models fitted on APDF (Asian Pacific Defense Forum) websites and Assorted Government documents (PDF Documents from Thai Government Sources) did not seem to have much correlation at all. This was surprising to us as we assumed different forms of government documentation would have similar styles and language. We propose that some of the difference we find between these models is due to the format they came from. The authors of our dataset used different methods to process web data and PDFs, so there may be resulting discrepancies. Also, APDF is heavily skewed towards military and defense language while the Assorted Government documents covers civil domains. Overall, due to these models' resulting performances on the other web based datasets, we believe that the processing method was the more significant contributor to the discrepancy. For this dataset, document format also influences the domains created. The issue of domain adaptation is not just a matter of the same words being used in different ways but also the same language in different formats. Another result of note was the performance on new domains (Taskmaster is a dataset of professionally translated conversations and Product Reviews are professional translations of generated product reviews). We know from previous work that both conversations and reviews tend to have irregular syntactic structures. Following from this, the model that generalized the best (highest BLEU score) to these domains was our Paracrawl web model. We postulate that this is due to some of the irregularities of the dataset being well matched with these new domains.

4.3 Overall Thoughts and Future Work

Overall, we find that in low resource situations, a model trained across multiple domains gives comparable performance to domain-specific models. As we increase training and computing usage, this comparability is likely to fall off, but in a low resource setting it holds. From this, we propose an idea: weighted domain learning. With weighted domain learning, we create our set of training examples by taking weighted training subsets across relevant domains. We propose two paths: general weighted learning and targeted weighted learning. In general, for weighted learning, the weights of the dataset would be produced from actual or estimated domain weights. For example, if we assume (or measure) that government documents are 5% of our total documents, then 5% of our limited training set will be government example pairs. In targeted weighted learning, one would instead set the weights to correspond to their target translation task. If we want to translate product reviews well and a second domain acceptably, we could apportion 66% of the training set to reviews and 33% to the second domain. This idea is similar to the mixed fine tuning proposed by (Chu1 et al., 2017) but is a variant based on fine-tuning of a pre-trained model rather than initial training on a domain and subsequent fine-tuning. Based on our findings for low resource transfer learning, this option will likely increase overall translation performance.

5 Conclusion

We find that there is merit to fine-tuning on a specific domain in low computing-resource situations. These models perform as well or a little better than generally-trained models, though they do not extend to new domains as well. Furthermore, predicting the appropriate related domain for fine-tuning a model for an unseen domain can be tricky as seen with our unexpectedly successful Paracrawl performance on the conversations and product reviews test sets and the poor performance of the expectedly similar APDF and assorted government documents. The preferred method of training will depend on the specific translation tasks that you want your model to perform and exceed at. We also find that training on a wide variety of formats and irregularities helps to increase model performance across domains. This is most pertinent for transitioning to less structured domains (conversation, reviews, etc). Based on our results, we also presented the idea of weighted domain learning for future study and use. This is an attempt to improve our general mixed model for a resource-constrained situation, which will hopefully provide a more efficient method for training a generalized model than just increasing the number of total training examples and total computation. Finally, we looked at data that came from a variety of different translation methods (professional translation to Google translation). Future work could consider the impact of training solely on professionally-translated or machine-translated examples and how well those generalize across different domains.

References

2020. English-thai machine translation models.

Chenhui Chu1, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. aclweb:P17-2061.

Wanyun Cui, Guangyu Zheng, Zhiqiang Shen, Sihang Jiang, and Wei Wang. 2019. Transfer learning for sequences via learning to collocate. arXiv:1902.09092.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. arXiv:1706.03872.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. arXiv:2101.09635.

Lalita Lowphansirikul, Charin Polpanumas, Attapol T. Rutherford, and Sarana Nutanong. 2020. scb-mt-en-th-2020: A large english-thai parallel corpus. arXiv:2007.03541.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. aclweb:P02-1040.

Matt Post. 2018. A call for clarity in reporting bleu scores. arXiv:1804.08771.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Sebastian Ruder. 2019. The state of transfer learning in nlp.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv:2008.00401.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. arXiv:1706.03762.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. arXiv:2002.06823.

A Appendices

A.1 Translation Examples

The following is an example of the predicted translations compared to the ”correct” translation for the Paracrawl-Fitted model on an example from the Paracrawl test set.

Target Translation

Cards can not be played from here.

Actual Translation

The cards will not be able to move from here.

A.2 Materials and Code

The Thai segments and more sentence pairs can be found in our GitHub repo along with the code to generate the translations and the code for our models and training process.