Finnish to English Neural Machine Translation Models

Johanna Männistö
January 29, 2023

1 Introduction

For the Neural Machine Translation Course Project, I built 3 NMT models to translate from Finnish to English. I trained a simple Transformer model as my baseline and improved upon it using data cleaning methods, domain adaptation methods, and a multilingual approach. Section 1 details common features and evaluation metrics used across each model while Sections 2 through 4 detail model-specific approaches and results.

1.1 Default Model Architecture

For all models described in this report[1], I use the Transformer model following the base configuration presented in Vaswani et al., 2017[9]; implemented with OpenNMT-py version 2.3.0 [5]. 6 layers are used for the encoder and decoder, with 8 attention heads, a word embedding size of 512, and a hidden feed forward size of 2048. The batch size is 4096 tokens. Each model uses adam as the optimization method, with a beta2 parameter of 0.998. Label smoothing was set to 0.1, dropout probability to 0.1, and attention dropout probability to 0.1. noam is the decay method, with a starting learning rate of 2 and warm up steps set to 8000.

1.2 Evaluation and Tatoeba Benchmarks

All models are evaluated by translating the Tatoeba v2021-08-07 test set[2] and then evaluated by their BLEU and chrF scores. Notable Tateoba benchmarks that can be used to compare the results of my models to include the models included in Table1 below. The models are also evaluated on a test set of 5,000 sentences from the original Europarl Corpus data set.

Model	**Data**	**BLEU**	**chrF**
Fin-eng/opus	Tatoeba	53.4	0.697
fin-eng/opus+bt	Tatoeba	53.1	0.695
fin-eng/opusTCv20210807+bt	Tatoeba v2021 08-07	57.3	0.7232
fin-eng/opusTCv20210807+nopar+ft95-sepvoc_transformer-align	Tatoeba v2021 08-07	54.5	0.70539

Table 1: Tatoeba Test Dataset Benchmarks

Notably, these models were trained on the OPUS data set, which includes 46 million sentences between English and Finnish.

2 Baseline Model

2.1 Data

The baseline model was trained on the Europarl Corpus [8] parallel data for Finnish and English. Using OpusFilter [2], only sentences containing between 2 characters and 250 characters were pulled, resulting in a data set size of 1,680,038 parallel sentences. The only preprocessing done was the WhiteSpaceNormalizer to replace sequences of whitespace characters with a single space.

The sentences were encoded using BPE with a vocabulary size of 10,000. When building the vocabulary with onmt_build_vocab the default size was used. Table 2 the line counts for each file type for the model.

[1] Configuration files for each model, with detailed parameter options can be found at github.com/jmannisto/Finnish-EnglishNMT

[2] The data for the test set can be found at: github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/data/release/test/v2021-08-07

File Type	Line Count
Train	1,670,038
Validate	5000
Test	5000
Tatoeba Test	10,690

Table 2: Line Counts by File Type for Baseline Model

2.2 Model Architecture

The model's architecture corresponds to the default architecture explained in Section 1.1. The model trained for 150,000 steps, passing over 44.2 million tokens, or 26 epochs.

2.3 Results

At 150,000 steps, the model had iterated several times over the training set, however when reviewing the .err files it does appear that the model's accuracy or perplexity hadn't quite plateaued yet, so I used the final model checkpoint at 150,000 steps to translate the Tatoeba test set for evaluation.

The translation was evaluated using BLEU and chrF metrics from sacreBLEU. The model achieved a BLEU score of 28.7 and chrF score of 51.3 on the Tatoeba v2021-08-07 test data. Table 3 summarizes the Baseline's results on the Tatoeba data.

Model	Data	BLEU	chrF
Baseline	Tatoeba v2021 08-07	28.7	51.3
Baseline	Europarl Test Set	29	55.2

Table 3: Baseline Model Results

2.4 Discussion

The baseline model scored between 24.4 and 28.6 fewer BLEU points or between 18.2 and 21.02 fewer chrF points than the benchmarks noted in Table 1. While my model performed worse than the benchmarks, it is important to note the amount of data used to train each of these models. The listed benchmark models used OPUS data, which for the Finnish-English language paring holds approximately 46 million sentences, while my baseline model uses only 1.6 million sentences.

A limitation of using this baseline model for translating the Tatoeba test set is a domain mismatch. The baseline model is trained on the Europarl corpus, a corpus of the proceedings of the European Parliament. This corpus is characterized by formal language with government jargon along with longer sentences (averaging 13.85 words per sentence in the Finnish source) while the Tatoeba data set consists of generally shorter sentences (averaging 5.54 words per sentence in the Finnish source text) with a more informal register. However, notably the BLEU and chrF score between the in-domain test set and the Tatoeba test set are quite similar, with the largest difference being between the chrF scores. This indicates that the model isn't significantly better at translating in-domain data than this out of domain data.

3 Multilingual Model

To improve upon the baseline model, I turned to a multilingual model as multilingual NMT systems are a known way to improve translation quality, through transfer learning [3]. Aharoni et al., 2019 [1] found that in high resource settings when using a multilingual to translate into one particular language, a many-to-one model performs best. So, I decided to train a multilingual many-to-one using Estonian and Finnish as source languages to translate to English. I decided to use a closely related language (Estonian) as part of the multilingual system. Given that both of the source languages would have English as the target the artificial token is excluded as the model should be able to learn the source languages automatically[4].

3.1 Data

The model was trained on the Europarl Corpus parallel data for Finnish to English and Estonian to English. The same filters were used for Finnish and Estonian as described in Section 2.1. A summary of the data set sizes by file type and language pair is summarized in Table 4[3].

File Type	Language Pair	Line Count
Train	Finnish – English	1,670,038
	Estonian – English	492,086
Validate	Finnish – English	5000
	Estonian – English	5000
Test	Finnish – English	5000
	Estonian – English	0

Table 4: Summary of Multilingual Data Size by File Type

A single SentencePiece BPE model was made using a vocabulary size of 32000 on a concatenated set of the Finnish, English, and Estonian data. A vocabulary was built using onmt_build_vocab from OpenNMT-py on the full shared corpus.

3.2 Model Architecture

The model's architecture corresponds to the default architecture explained in Section 1.1. Given that the Estonian data set was approximately 3 times smaller than the Finnish to English set, it was given a weight of 3 while the Finnish to English set had a weight of 1, so that during training the data set for Estonian would be over sampled to approximately the same size as the Finnish to English set. The model trained for 100,000 steps, passing over 19.96 million tokens, or 6.35 epochs.

3.3 Results

The model had poorer results on the Tatoeba test set compared to the baseline model, with a BLEU score of 24.6 and chrF score of 45.5, however there is improvement on the Europarl test set by 4.6 BLEU and 5 chrF scores.

Model	**Data**	**BLEU**	**chrF**
Multilingual	Tatoeba v2021 08-07	24.6	45.5
Multilingual	Europarl Test Set	33.6	60.2

Table 5: Multilingual Model Results

3.4 Discussion

This multilingual model did not see improved performance on the Tatoeba task, however it did have a notable improvement in translating the Europarl test set. This may be unexpected given that we may expect across the board improvement multilingual models due transfer learning, however there are a few factors at play to consider.

While it seemed that domain issues may not be the primary concern when examining the baseline model results, given the similarity for in and out of domain data, the distinct difference in performance for the multilingual model indicates that domain adaptation is necessary for improvement. In the multilingual model's case we provided additional Europarl data which is very distinct from the Tatoeba data. The model, with now twice as much exposure to Europarl data as the baseline model may not be as adept at generalizing in out of domain situations.

We also assume that choosing related languages will automatically benefit a multilingual model, however what makes languages related is ambiguous. Kocmi Bojar, 2018 [6], also using Europarl corpus, as well as the Rapid corpus, found that when training a child Estonian model off of Finnish and vice versa the amount of gain on performance was less

[3] Excluded an Estonian language test set as performance was not under investigation

than that of Estonian with Czech or Russian, indicating considering factors outside of language family could be at play and relevant to transfer learning. It would be interesting to explore this further to see if it is possible to leverage even larger improvements with a two-source multilingual system in this set up.

4 Finetuned Model

As mentioned in Section 2.4 and 3.4, after examining the Tatoeba test data, it was evident that there was a domain mismatch between the model and what the model was expected to translate. To improve model performance, I finetuned the baseline model with a corpus which was more similar to the data set it was expected to translate.

The model was finetuned on data from the OpenSubtitles corpus[4] [7]. This corpus is much more colloquial, frequently with shorter and simpler sentences than the Europarl corpus. It appeared to be a better match in type and language register for the Tatoeba data, making it a good candidate to use for finetuning the model and adapting the domain.

4.1 Data

The baseline model was trained on the Europarl Corpus parallel data for Finnish and English. I performed additional filtering on the data to it clean up further. I added two additional filters to the collected and filtered Europarl data. The first, a LengthFilter which excludes sentences with fewer than 3 or more than 100 words. The second, an AlphabetRatioFilter with a threshold of 0.75, including white space. This excluded extremely short phrases from the corpus, previously included with the original filter settings described in Section 2.2 and also excluded phrases primarily consisted of punctuation, numbers, or other symbols rather than words. This processing reduced the data set size from 1,680,038 parallel sentences to 1,597,870 parallel sentences.

The data the model was finetuned on was from OpenSubtitles. This corpus was pulled using OpusFilter and the same filters used to re-filter the Europarl Corpus for this set. 1.5 million parallel sentences were randomly taken from the larger OpenSubtitles corpus for the finetuning task. From the set of 1,597,870 Europarl parallel sentences, I sampled 500,000 parallel sentences to include in the finetuning process as well to limit any forgetting by the model. This resulted in a data set size of 2 million parallel sentences. The model was finetuned on a training set size of 1,992,500 slightly larger than the baseline model by 322,462 parallel sentences.

File Type	Line Count
Train	1,992,500
Validate	5000
Test	2500
Tatoeba Test	10,690

Table 6: Line Counts by File Type for Finetuned Model

The SentencePiece BPE models of the baseline model were used but the vocabulary was updated to encompass new words.

4.2 Model Architecture

The model's architecture corresponds to the default architecture explained in Section 1.1. In addition to the 150,000 steps the baseline model had trained, this model trained for an additional 90,000 steps, a total of 10.75 epochs. An early stopping parameter was added, telling the model to stop if after 3 checkpoints perplexity did not improve.

4.3 Results

This model had a significant improvement in its results, with a BLEU score of 44.1 and chrF score of 65.0 at step 240,000 on the Tatoeba test set. This is a large improvement in the model and brings it more in line with the result from the other Tatoeba benchmarks (see Table 1).

[4] www.opensubtitles.org

After reviewing the .err files, I noticed that the perplexity and accuracy had been stagnant and appeared to have plateaued around the 200,000-step mark, as illustrated in Figures 1 and 2.

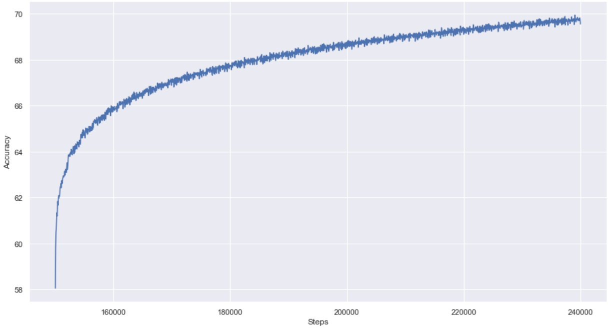

Figure 1: Accuracy Rate by Steps

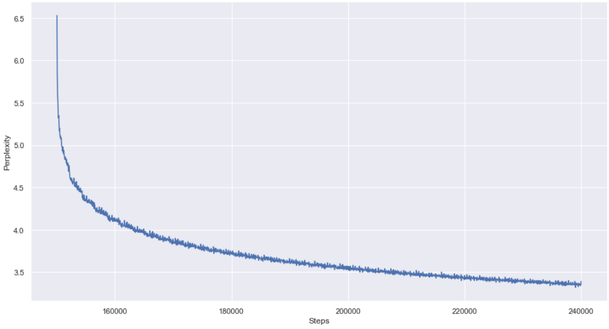

Figure 2: Perplexity by Steps

This indicates that the model may be over-fitting and would not continue to improve with additional training. Using the checkpoint model at step 210,000 to translate the test set again yielded a slight improvement on BLEU score to 44.2 while chrF remained the same.

Model	Data	BLEU	chrF
Finetune, step 210,000	Tatoeba v2021 08-07	44.2	65
Finetune, step 210,000	Europarl Test Set	29.1	57.2

Table 7: Results on Finetuned Model

4.4 Discussion

The finetuned model benefited significantly from the new, more in-domain data. After finetuning or an additional 60,000 steps, the BLEU and chrF scores increased 15.5 and 13.7 respectively, much closer to the baseline models on the Tatoeba data. It would be particularly interesting to have a baseline model trained on 1.5 million parallel sentences from the OpenSubtitles corpus to have a better understanding off the impact finetuning or the inclusion of the Europarl corpus had.

5 Conclusion

The finetuned model performed the best on the Tatoeba data, while the multilingual model performed the best on the Europarl data. This was likely due to the inclusion of in-domain data, as well as the addition of more data in general for the models to learn from. None of the models described in this report achieved the same performance as the models summarized in Table 1. Table 8 summarizes the results on each test set for each model.

Model	Tatoeba		Europarl	
	BLEU	chrF	BLEU	chrF
Baseline	28.7	51.3	29	55.2
Multilingual	24.6	45.5	33.6	60.2
Finetune	44.2	65.0	29.1	57.2

Table 8: Summary of Model Results on Tatoeba and Europarl test sets

This project highlighted the importance of data for NMT models. If Tatoeba was instead a medical text and I hadn't been able to find related in-domain data, achieving higher performance would be difficult due to the challenges of domain adaptation when the data or vocabulary is unavailable. While not the focus of this project, exploring different combinations of multilingual source and target pairs as well as examining different hyper parameters would likely yield improved results.

References

[1] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively Multilingual Neural Machine Translation, July 2019. arXiv:1903.00089 [cs].

[2] Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. OpusFilter: A configurable parallel corpus filtering toolbox. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 150–156. Association for Computational Linguistics, July 2020.

[3] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A Comprehensive Survey of Multilingual Neural Machine Translation, January 2020. arXiv:2001.01115 [cs].

[4] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. 2016.

[5] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[6] Tom Kocmi and Ondřej Bojar. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 244–252, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[7] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[8] Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. 2017.