Exploiting Multilingualism and Multistage Fine-Tuning for Low-Resource Neural Machine Translation.

Mohammed El Dor
Deep Natural Language Processing
Politecnico Di Torino
Turin, Italy
s289298@studenti.polito.it

Anaam Ur Rahman
Deep Natural Language Processing
Politecnico Di Torino
Turin, Italy
s283909@studenti.polito.it

Mohamad Mostafa
Deep Natural Language Processing
Politecnico Di Torino
Turin, Italy
s291385@studenti.polito.it

***Abstract*—This paper explores the potential of multi-stage finetuning methods using multilingual parallel corpus of low-resource target languages. Our work is a recreation of Dabre et al. (2019). We extend their work by using state-of-the-art tokenizers to tokenize Asian language sentences. In addition to that, we ran our experiments on 4 new Asian languages which were not tested before. The paper highlights the impressive utility of multi-parallel corpora for transfer learning in a one-to-many low-resource neural machine translation (NMT) setting. We report on a comparison of multistage fine-tuning of a pretrained model trained on two corpora using one language at a time, then a mix training step using all or one of the low resource languages and the dataset used for pretraining, followed by a pure fine-tuning process of one of the low resource languages. Our results show that multilingualism is a powerful tool to improve BLEU scores (up to 3-10 BLEU score gains).**

I. Introduction

Encoder-Decoder based Neural machine translation architectures (NMT) are state-of-the-art NMT architectures. Often they require rich-resource language pairs to train which introduces a limitation for resource-scarce languages. Hence, using low-resource language pairs (only) to train NMT models yields low performances for the task. To overcome this issue, transfer learning techniques are used to leverage from the knowledge learned during pre-training phase and finetuning the model on desired low-resource language pair.

It is possible to further improve the fine-tuning process by using multi-stage fine-tuning methods which uses multilingual parallel corpora to translate from English to a low-resource target language.

Our work focuses on (a) improving the performance of NMT for English to low-resource target language translations (b) via exploiting multilingualism (c) through transfer learning based on multistage fine-tuning. The power of multilingualism is verified by using multi-parallel corpora, i.e., the same text in different languages. Unlike previous studies that only apply single-stage fine-tuning for one-to-one translation, we systematically compare multistage fine-tuning that exploits one-to-many modeling. We show that this approach can significantly improve the quality of translations from English to nine Asian languages (Vietnamese (vi), Japanese (ja), Malay (ms), Hindi (hi), Indonesian (id), Filipino (fil), Khmer (khm), Myanmar (my), Thai (th)) in the Asian Language Treebank (ALT) corpus [3].

We extend Dabre et al. [1] work in 3 different directions:

- We used state-of-the-art tokenizers to tokenize Asian languages. These languages have diverse syntax and grammar rules. Using them in raw form or splitting with space and punctuation is clearly not an optimal solution.
- While mix-finetuning, we used all low resource

languages to train the pretrained NMT model, which yielded unbiased results for all the target languages.

- We added four new languages in our experiments which were not tested in previous work.

We also provide the source code that was missing in the previous work https://github.com/DeskDown/NMT . The Fine-tuned, best performing NMT models are available at https://huggingface.co/DeskDown

II. Methodology

A. *Multi-stage Finetuning*

Our work focuses on translation from English (En) to N different languages. In particular, we consider exploiting two types of corpora. One is a small-scale multi-parallel corpus, En-YY_1- $\cdots$ -YY_N , consisting of English and N target languages of interest. The other is a relatively larger helping parallel corpus, En-XX, where XX indicates the helping target language with different corpus than that in the multi-parallel ones, YY_k ($1 \leq k \leq N$).

The multi-stage fine-tuning techniques can be generalized as follows.

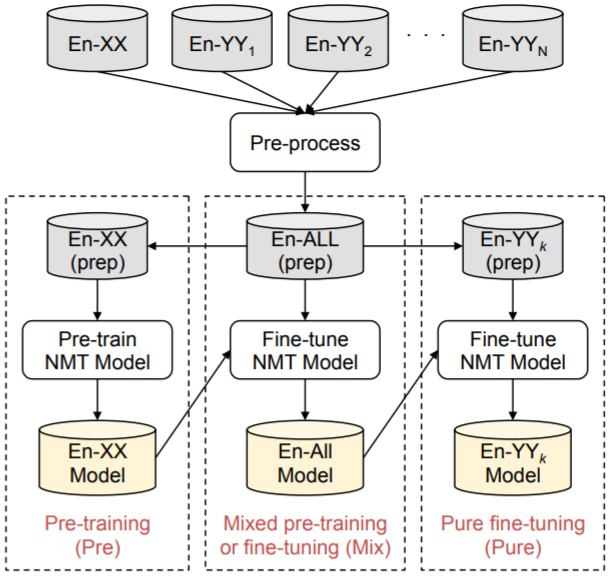

Fig. 1: The Multistage fine-tuning pipeline. Three dotted sections respectively indicate (i) the pre-training stage, (ii) the mixed pre-training/ fine-tuning stage, and (iii) the pure fine-tuning stage. Stage (i) and (iii) use one pair of translation sentences at a time. Instead, stage (ii) jointly trains on multilingual parallel corpora to exploit one-to-many modeling.

Pre-training (Pre): An initial NMT model is trained on a resource rich language pair, i.e., En-XX. Our choice of use were the Chinese and Japanese languages due to the close domains of them and the target resource-poor languages. At this stage, our main focus is to train the English encoder of NMT model.

Mixed pre-training / fine-tuning (Mix): Training of the NMT model is continued on a mixture of parallel corpora for En-XX and one or several low-resource En-YY pairs of low-resource target languages. This stage exploits one-to-many modeling using multiparallel corpora. In contrast to the original work, we performed under sampling to match helper dataset cardinality with low-resource Asian language sentences pairs.

Pure fine-tuning (Pure): This stage is similar to well-known and standard finetuning methods. A target language $EN-YY_k$ is selected and used for further fine-tuning the pretrained model of previous stages.

Figure 1 illustrates the training procedure with all of the above three stages.

To skip the pretraining stage, We used open source pretrained EN-XX NMT models available at HuggingFace hub (*Helsinki-NLP/opus-mt-en-zh* or *Helsinki-NLP/opus-tatoeba-en-ja*). These are Seq2Seq encode-decoder Transformer models that use the MarianMT model [2].

We continued the Mix-finetuning on pretrained models using 1 (or all 9) languages from ALT dataset and a subset of helper dataset. Finally, we concluded our training after pure-finetuing the mixed-finetuned model on $EN-YY_k$ language pair.

B. *Tokenization*

Pretrained NMT models have their own tokenizer which they used during pretraining phase. This tokenizer is capable of tokenizing the source and the helper languages but not the YY_k sentences. Asian targets languages have different characters, syntax and may not follow the same grammatical rules as English or helper language. One the other hand, the

previously contained knowledge of pretraining stage is also vital for efficient transfer learning.

We propose a Teacher-Student tokenization method. The Student tokenizer is the tokenizer used in the pretraining stage and it is perfectly capable of tokenizing the English and helper language. As the Teacher, we used state-of-the-art *facebook/mbart-large-50-one-to-many-mmt* tokenizer [4], which is capable of tokenizing Asian target languages. We created a list of unique tokens which were never seen by Student tokenizer before. Finally, these tokens were added in the vocabulary of Student tokenizer as special tokens. By Using this method we preserved the previous knowledge and extended the vocabulary space of the tokenizer and the NMT model. To ensure reproducible, we saved a local copy of added tokens and used it whenever it was necessary to hard-train the student tokenizer. The embedding matrix of NMT model is also extended to match new vocabulary size and it is randomly initialized for newly added tokens of Asian target languages.

III. Experiments

A. Datasets and Pre-processing

As the test-bed, we chose the ALT multilingual multi-parallel corpus (Riza et al., 2016) [3]. It offers the same English test sentences translated in different languages. ALT dataset contains a corpus of 18-20K sentences of each EN-YY pair. To indicate the target language of each sentence pair, we prepend an artificial token, for instance "2zz" for a target language "zz," onto source sentences. (Dataset Available on HF datasets hub *DeskDown/ALTDataset*)

As for the resource-rich En-XX data, we performed down-sampling and used 18K sentences of helper language from *opus100* dataset. This dataset was used to pretrain the NMT model. We use this dataset during mix-finetuning stage which exploits one-to-many modeling.

We split the ALT dataset into train-eval-test subsets using the partitions proposed by ALT project official documentation.

B. Training NMT Models

We used open source pretrained NMT EN-XX models. As the helper language, we choose Japanese and Chinese languages. For each helper languages, we perform training using three different configurations:

1) We set the baseline by performing standard (Pure)finetuning of one EN-YY pair. It is shown as #1 in table 1.

2) We used one $EN-YY_k$ pair alongside helper dataset to Mix-finetune the pretrained model and later performed Pure-finetuning using same $EN-YY_k$ corpus. It is shown as #2 in table 1. This settings helps us understand the effects of adding Mix-training stage for efficient domain adaptation.

3) To explore the maximum potential of multilingual corpora, we concatenated dataset containing 9 languages from EN-YY low-resource corpus and the helper language dataset. This dataset was used in Mix-Finetuning stage of the pretrained model. Later we Pure-finetuned this mix-trained model for each pair of EN-YY dataset. The process is shown as #3 in table 1.

The tokenizer is shared among all stages and training configurations. Hugging face Trainer API was used in all finetuning stages. We performed a basic hyper-parameter search for the learning rate and used the same learning rate for all stages of finetuning. Further details about training parameters are available in the source code. We continued the training cycles until convergence was achieved or the gain in the BLEU score on the development set was less than 0.2 for 5 consecutive epochs.

C. Results

Table 1 gives the BLEU scores for all the configurations. Among the three configurations, irrespective of the external parallel corpus for En-XX, the three-stage fine-tuned model (#3) achieved the highest BLEU scores for all low-resource target languages. #2 demonstrate that even without whole multilingual corpora, we can achieve a performance increase of upto 8 BLEU score by adding the Mix-Finetuning stage. It clearly highlights the benefits of training the pretrained model with low-resource Asian language and helper language sentences side

#	XX	N	Model Capacity	Training configuration			YY Test Set								
				Pre	Mix	Pure	Vi	Ja	Ms	HI	Id	Fil	Khm	My	Th
1	Zh	1	1-2	X	–	X	33.01	18.77	39.6	28.05	31.92	25.24	23.92	20.97	21.43
2	Zh	1	1-2	X	X	X	35.7	17.99	43.38	31.83	36.9	33.39	24.8	25.61	22.76
3	Zh	10	1-10	X	X	X	**37.37**	**19.09**	**44.92**	**33.82**	**40.28**	**35.05**	**27.23**	**28.14**	**24.26**
*	Zh	7	1-8	X	X	X	*35.34*	*20.08*	*33.19*	–	*27.24*	–	*28.66*	–	–
1	Ja	1	1-2	X	–	X	26.44	18.45	35.24	18.64	27.51	18.46	11.47	4.85	18.47
2	Ja	1	1-2	X	X	X	32.01	18.45	38.17	28.24	33.84	30.19	21.75	23.65	19.99
3	Ja	10	1-10	X	X	X	**34.99**	**19.07**	**40.66**	**30.85**	**36.29**	**30.55**	**25.24**	**26.01**	**22.01**
*	Ja	7	1-8	X	X	X	*37.06*	*22.60*	*34.75*	–	*28.89*	–	*30.03*	–	–

TABLE I: BLEU scores for all the tested configurations. The "XX" column indicates the external (helping) parallel corpus if used, where "Zh" and "Ja" stand for using the Opus English–Chinese or English–Japanese corpora, respectively. The columns under "YY test set" indicate the the target languages in our multi-parallel corpus, where **bold** marks our best scores for each target language with each external parallel corpus. * represents the results achieved by Dabre et al. [1]

by side. The presence of helper dataset allows better domain adaptation environment to NMT model.

In #3 To test the usefulness of one-to-many modeling. we expand the model's potential to learn EN-YY translations, all at once, during Mix-finetuning stage. Although, the English sentences are same for all $EN-YY_k$ pairs, the model learns to translate in the particular target language due to appended *"2zz"* token in source sentences (see section III-A). It is evident that the performance with respect to the baseline is improved by 3-10 Bleu scores. When compared with #2, we were able to achieve 2-4 Bleu score gain. Since the Pure-finetuning stages used the same datasets in #2 and #3, the gain in performance is only due to multilingual parallel corpus used for one-to-many modeling in Mix-training stage.

IV. Conclusion

We explored the use of small multi-parallel corpora and subset of helper language dataset for training one-to-many NMT models. Jointly training the pretrained model on helper and resource-scarce target language improves the models ability to adopt the new target domain. We tested multistage fine-tuning methods and confirmed that it can achieve better performances with respect to the models trained via fewer fine-tuning stages.

Multilingual corpora also plays a significant role when we use all low-resource ALT languages in Mix-training stage. Decoder learns the target representations all at once and it is possible that representations of one $EN-YY_i$ might help another $EN-YY_j$ pair in Pure-finetuning stages.

A possible explanation for the fact that En-Zh pretrained model performs better w.r.t En-Ja can be the their difference in training corpus size. They also have different benchmarks on Opus dataset. En-Zh was able to achieve 31.4 Bleu score on test sentences. En-Ja instead achieved only 15.2 Bleu score while pre-training.

We conclude our work by marking the usefulness of multilingual corpora. For both helper languages, It played a significant role when we used all low-resource ALT languages in Mix-training stage. Decoder learns the target representations all at once and it is possible that representations of one $EN-YY_i$ might help another $EN-YY_j$ pair in Pure-finetuning stages.

References

[1] Raj Dabre, Atsushi Fujita, and Chenhui Chu. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416, Hong Kong, China, November 2019. Association for Computational Linguistics.

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

[3] Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. Introduction of the asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6, 2016.

[4] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401, 2020.