Mattia Di Gangi · Follow

Oct 18, 2021 · 7 min read ★ · Listen

mBART50: Multilingual Fine-Tuning of Extensible Multilingual Pretraining

Train multilingual machine translation for 50 languages with pretraining on monolingual data

Photo by Soner Eker on Unsplash

Note: this is the second article in a series. Some concepts are described in the previous article Massive Pretraining for Bilingual Machine Translation.

encoder-decoder models:

1. Can the set of pretraining languages be extended to open more downstream fine-tuning possibilities?
2. Since the pretraining is multilingual, why isn't the fine-tuning also multilingual?

These are the questions answered in [1], which presents the following contributions:

1. Extend mBART to 50 languages with no loss of accuracy on the bilingual fine-tuning
2. Propose multilingual fine-tuning for Many-to-English, English-to-Many and Many-to-Many (through pivoting) with gains over all baselines. The gains are particularly large for the Many-to-English direction
3. Provide the ML50 benchmark with a standardized train/dev/test split, which covers low-, mid-, and high-resourced languages for a total of 230M parallel sentences.

Multilingual Fine-tuning

The multilingual fine-tuning process is not particularly complicated.
First of all, the data for all the language directions **s->t** are collected in a single training set. The amount of text available for each language direction is extremely diverse, covering the full spectrum from 4K to more than 10M sentence pairs per language pair. Thus, the data for each batch is chosen from the different sets using temperature sampling on the language directions themselves. As described in a previous paper [2], temperature sampling produces a more balanced distribution so that low-resourced language pairs are chosen with a higher probability.

Multilingual fine-tuning is experimented in three different settings: Many-to-English, English-to-Many, and Many-to-Many. The last one is trained by combining the data of the other two settings, so that English is a kind of pivot language.

Results on 25 Languages

A first experiment aims to compare multilingual fine-tuning with bilingual fine-tuning and multilingual from scratch on the 25 languages used to train mBART.

The results show that, on the Many-to-English direction, **multilingual models clearly outperform all the bilingual models**. Additionally, **multilingual fine-tuning is significantly better than multilingual from scratch,** except for language directions with more than 1M sentence pairs, where the advantage is quite small. On the other side, for the less-resourced languages (7K-30K sentence pairs), the advantage of multilingual fine-tuning over bilingual from scratch is an outstanding **18.03 average BLEU points**, which makes a huge difference between a toy and a usable model. For a comparison, the second best model is multilingual from scratch with an improvement of 14.63, while bilingual fine-tuning stops after 10.80 BLEU points of improvement over the bilingual from scratch baseline.

The English-to-Many setting presents a completely different scenario.
Firstly, it is important to say that here the results described in the paper do not fully match the table, which received less care than the previous ones. In fact, unless I missed something, the values marked in **bold** as the best ones in a row do not appear to be so.

Then, by reading only the numbers and not the text, we can see that multilingual fine-tuning achieves the best result (by a strict margin) only on the >10M sentence pairs group, while for the other groups multilingual from scratch English-to-Many is consistently better. Unfortunately, *the authors fail to explain why this happens* but it represents an interesting result for further research.

Moreover, on the least-resourced group, the Many-to-Many model with **multilingual fine-tuning** is only 0.90 better than the **bilingual from scratch** baseline, while **multilingual from scratch** has an advantage of 7.9 BLEU points over the baseline. The improvement for the multilingual fine-tuning model is much better on the one-to-Many scenario, but still slightly worse than its counterpart trained from scratch (7.6 vs 8.1).

While the authors fail to explain the counterintuitive results on One-to-Many, where training from scratch is better than fine-tuning, they are very clear in explaining their best scenario.

The Many-to-English direction is so strong because the model observes the English target side of 49 language directions, **and the**

experiments if mBART solves that problem, and if yes, why.

mBART50: Extending a Pretrained Model

In the previous article, we have seen that mBART is useful also to be fine-tuned on language pairs with at least one **unseen language.** The results, though, are not as good as for the pretraining languages. The degradation is particularly bad when the unseen language is on the source side.

Thus, the pretraining languages constrain the possible downstream tasks to those languages. Then, in order to use mBART effectively with new languages, it is important to have more monolingual data that can be added to the pretrained model. However, given the computational and time resources needed for training it, it would be nice not to start it from scratch.

To overcome this problem, the authors propose to extend the final mBART25 checkpoint with 25 additional languages, effetively doubling their amount. The procedure is conceptually simple but effective. They start by adding 25 randomly-initialized language tokens in the embedding layers, one for each new language. Then, they concatenate the monolingual training data of the original 25 languages with the data of the 25 new languages and use the resulting training set to continue the training of the saved checkpoint. They also reuse the same sentencepiece model for word segmentation, as it was trained on 100 languages and not only on the original 25. Also, the authors perform some data filtering for both preventing any overlap between train and dev/test sets, and filtering out sentences that are not recognized as in the correct language by fasttext.

Multilingual fine-tuning of this new pretrained model leads to huge improvements, particularly for low-resourced languages. However, the improvements are, again, mostly in the many-to-one direction, while the one-to-many model is generally not as performant as biligual fine-tuning, except for very low-resourced language pairs.

Additionally, when performing bilingual fine-tuning of mBART50 on the original 25 languages, the results are really similar to the ones obtained by fine-tuning mBART25. *The higher number of languages doesn't lead to performance degradation on any language pair*.

Open Questions

The work about multilingual fine-tuning of mBART50 is really interesting and the results are exciting, but it leaves room for further research.

The first question, which is already object of study, concerns how to better extend the pretrained model. In the paper, old and new data are merged into a new training set to obtain a new model that doesn't degrade on the old languages, but it assumes access to the old training data. How can we achieve similar results when we don't have access to the original data?

Then, the authors show that mBART25 and mBART50 obtain the same results on bilingual fine-tuning for the original 25 languages and thus deduce that the model quality doesn't degrade on those languages. However, based only on the mBART papers, we do not know how the translation quality on the downstream task is related to any quality metrics of the pretrained model. It would be interesting to explore this more, but running the experiments is too expensive for many labs.

Finally, given the number of languages, the results showed in the paper are averaged among many languages. From the averages we can see large improvements for Many-to-1 and slight degradation for 1-to-Many compared to bilingual fine-tuning. However, the results on individual languages show a more complex picture. While the improvements are significant for basically all languages with the Many-to-1 model, the slightly-below-zero average degradation for the 1-to-Many model is the result of averaging among highly varying results. For the single languages, the difference with the baseline can be as high as+- 4 BLEU points! It is an interesting question to understand the reason of such high fluctuations and if it's possible to move more languages on the positive side.

Conclusions

Multilingual fine-tuning of mBART50 is another important chapter that shows how to train strong multilingual models that can be used also for very low-resourced langauge pairs.

Multilingual models offer maintainance advantages over having a plethora of models in production. This advantage has always been traded-off with a lower quality for the highest-resourced languages. *With mBART50 it seems that the degradation does not occur anymore.* The experiments were run only on a single test for each language direction, but it would be great if confirmed on more test data.

obviously mBART for the translation side.

References

[1] Tang, Yuqing, et al. “Multilingual translation with extensible multilingual pretraining and finetuning.” *arXiv preprint arXiv:2008.00401* (2020).

[2] Arivazhagan, Naveen, et al. “Massively multilingual neural machine translation in the wild: Findings and challenges.” *arXiv preprint arXiv:1907.05019* (2019).[link]

Medium Membership

Do you like my writing and are considering subscribing for a Medium Membership for having unlimited access to the articles?

If you subscribe through this link you will support me through your subscription with no additional cost for you https://medium.com/@mattiadigangi/membership