PIVOT PRE-FINETUNING FOR LOW RESOURCE MACHINE TRANSLATION

Stephen Kiilu [1] Machel Reid [2]

[1]African Master's in Machine Intelligence (AMMI)-AIMS [2]Google DeepMind

TL;DR

We look to investigate the usage of pivot languages (i.e. intermediate languages) to bridge the gap from a pre-trained multilingual sequence-to-sequence model and extremely low-resource machine translation, an approach we refer to as pivot pre-finetuning.

Abstract

Current approaches to performant machine translation often require large amounts of data. However, a majority of the 7,000+ languages in the world often have a relative lack of digitized/organized text available and are considered low-resource. In practical terms, this often means that there is a substantial drop in quality in translation performance between high and low-resource language pairs. We look to explore the intersection of rapid NMT adaptation techniques and pre-trained sequence-to-sequence models to better leverage multilingual models, performing a case study on Kikamba.

Pivot pre-finetuning

We look to investigate the usage of pivot languages (i.e. intermediate languages) to bridge the gap from a pre-trained multilingual sequence-to-sequence model (e.g. mT5 and mBART) and extremely low-resource machine translation (e.g. English to Kikamba). We follow the approach of related language-pair pretraining approach for rapid adaptation of NMT systems to low-resource language with pre-trained language models in the loop. This takes advantage of the data efficiency of pre-trained LMs, while also leveraging the rapid NMT approach without having to pre-train a language model with machine translation data from scratch for increased performance. We refer to this approach as **pivot pre-finetuning**, as we leverage a pivot language pair for a pre- finetuning procedure to enable data-efficient machine translation performance.

Approach

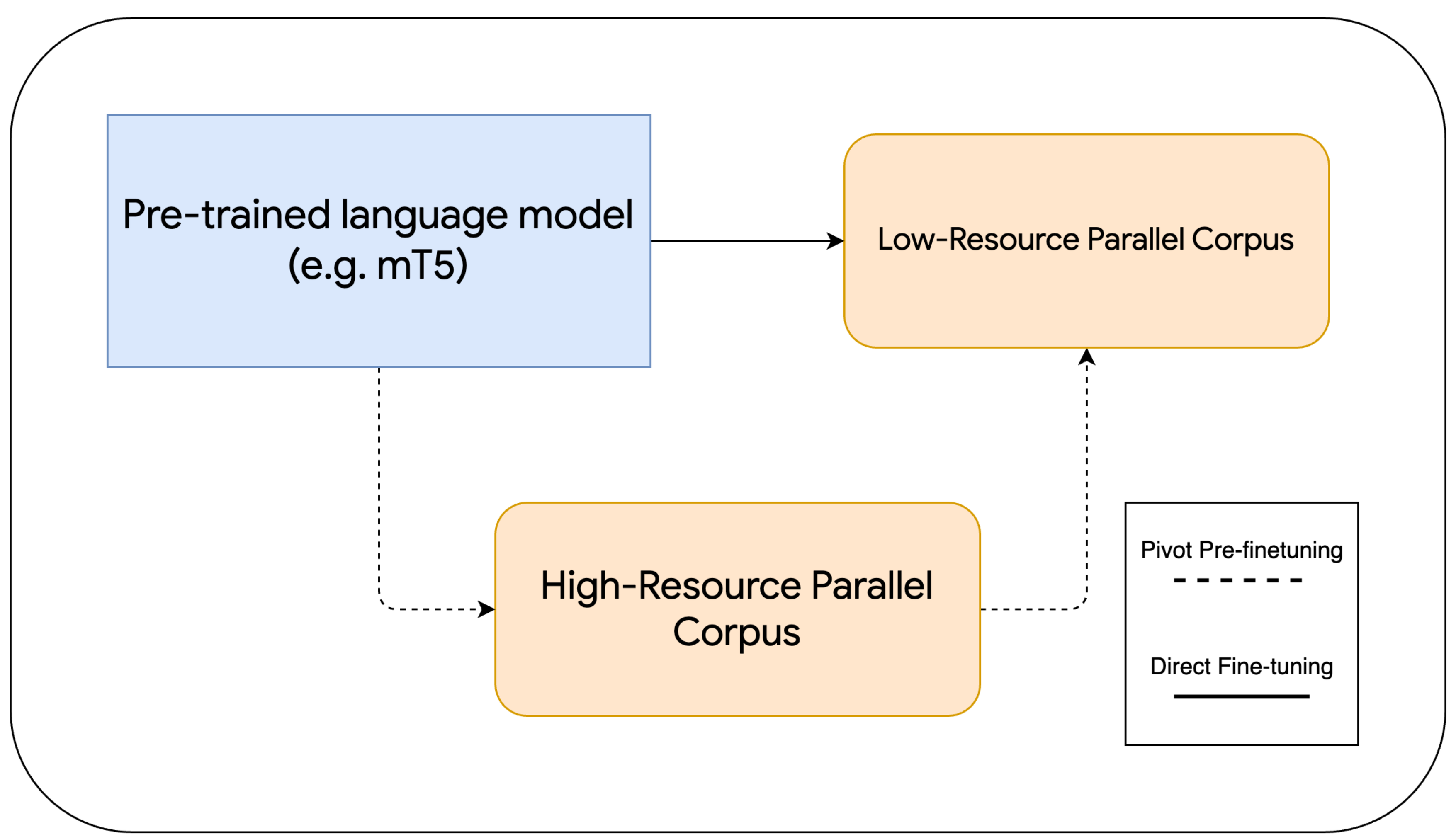

Figure 1. Comparison between pivot-prefinetuning and direct finetuning methods

Experimental Setup

To validate the use of pivot pre-finetuning, we compare direct Kikamba fine-tuning of a pre-trained massively multilingual sequence-to-sequence model (mT5; 7) and various intermediate pre-finetuning languages spanning different resource levels and dataset sizes.

Model We choose the 300M parameter mT5-small [7] to be the backbone for our experiments. Leveraging pre-trained multilingual sequence-to-sequence models for machine translation has shown not only improved performance but also improved data efficiency [1].

Datasets We use three pivot languages: Kikuyu (another extremely low-resource language, but most similar to Kikamba), Kiswahili/Kinyarwanda (relatively mid-resourced languages, however more distant than Kikuyu), and French (an extremely high-resource European language). For each language, we compare using varying amounts of *pivot pre-finetuning* and direct fine-tuning data pairs for training.

Evaluation For evaluation, we use a subset of 500 pairs from the FLoRes-200 English-Kikamba `devtest` data. We evaluate using SacreBLEU [5, 2] and ChrF [4] given the morphologically-rich nature of Kikamba. Finally, we conduct a small human evaluation of machine translation output using our technique.

Results

Target	Pivot lang	Resource-level	# Examples	Backbone	BLEU	ChrF
Kikamba	None	-	0.5k	mT5-small	**0.0858**	**4.429**
Kikamba	Kikuyu	low	1k	mT5-small	0.0034	3.261
Kikamba	Kinywaranda	mid	1k	mT5-small	0.0035	4.394
Kikamba	Kiswahili	mid	1k	mT5-small	0.0065	3.906
Kikamba	French	high	1k	mT5-small	0.0022	4.212
Kikamba	None	-	25k	mT5-small	0.1296	7.805
Kikamba	Kikuyu	low	50k	mT5-small	0.4662	11.143
Kikamba	Kinywaranda	mid	50k	mT5-small	0.6487	10.735
Kikamba	Kiswahili	mid	50k	mT5-small	**0.7806**	**11.151**
Kikamba	French	high	50k	mT5-small	0.5730	10.665
Kikamba	None	-	50k	mT5-small	0.0243	7.982
Kikamba	Kikuyu	low	100k	mT5-small	0.086	**11.625**
Kikamba	Kinywaranda	mid	100k	mT5-small	0.2823	10.993
Kikamba	Kiswahili	mid	100k	mT5-small	**0.3517**	11.042
Kikamba	French	high	100k	mT5-small	0.1705	9.798

Table 1. Comparison on various pivot pre-finetuning settings for Kikamba translations.

Target	Pivot lang	# Examples	Backbone	# 1-star	# 2-star	# 3-star
Kikamba	–	0.5k	mT5-small	100	0	0
Kikamba	–	25k	mT5-small	100	0	0
Kikamba	Kiswahili	50k	mT5-small	80	20	0
Kikamba	Kiswahili	100k	mT5-small	80	19	1

Table 2. Human evaluation results. We sample 100 Kikamba sentences and perform human evaluation. 1-star means almost no fluency (only $< 30\%$ of the concepts are translated, 2-star – some fluency in the translation (there is context and about half of the concepts are translated) and 3-star means almost fluency (about 70% correct translation).

Discussion

We show automatic evaluation results in Table 1. For all languages, we show improvements from introducing pivot pre-finetuning after 50k examples, however, these results seem to plateau with 100k pivot pairs maintaining similar performance. Given this, we can assert that within our experimental setup, pivot pre-finetuning is indeed helpful, boosting performance by 40% (measured by chrF) over a non-finetuned baseline. Despite the lesser quality of pairs in Kiswahili and Kinyarwanda, we find consistent improvements over using French data, suggesting that language similarity is important.

Our human evaluation (Table 2) also reflects results similar to the chrF metric (suggesting that BLEU is somewhat inaccurate for this language pair), training with the pivot language of Kiswahili with both 50k and 100k improves human evaluation.

Conclusion and future work

We have experimented with *pivot-prefinetuning* with the language of Kikamba, a method to enable more data-efficient adaptation to low-resource languages, combining rapid NMT adaptation techniques and pre-trained sequence-to-sequence models. However, the issue of data scarcity in low-resource languages persists. To address this challenge, in our future work, we explore alternative approaches, such as prompting and pivot prompting. In previous studies for high-resource languages [3], these approaches require significantly less data and allow us to leverage the power of large pre-trained language models (PLMs) for machine translation, without the need to train separate models.

References

[1] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions Of The Association For Computational Linguistics*, 2020. doi:10.1162/tacl_a_00343.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi:10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040`.

[3] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*, 2023.

[4] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi:10.18653/v1/W15-3049. URL `https://aclanthology.org/W15-3049`.

[5] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W18-6319`.

[6] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

[7] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *NAACL*, 2021.

skiilu@aimsammi.org