Interactions in Systematic Pre-Trained Neural Machine Translation

Ashleigh Richardson, Janet Wiles
The University of Queensland, Australia

ABSTRACT

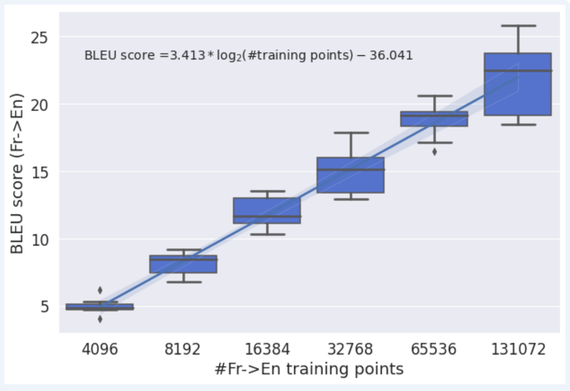
Transfer learning is a promising direction for low-resource NMT, but introduces many new variables. The effects of these variables, such as auxiliary task choice and dataset sizes, are typically analysed independently due to computational cost. We hypothesise that these factors are not independent and demonstrate initial evidence via a 3-way (7x6x2) systematic study which reveals statistically significant (p-value < 0.0018) non-trivial interactions between main and auxiliary dataset sizes and task relatedness.

EXPERIMENTAL SETUP

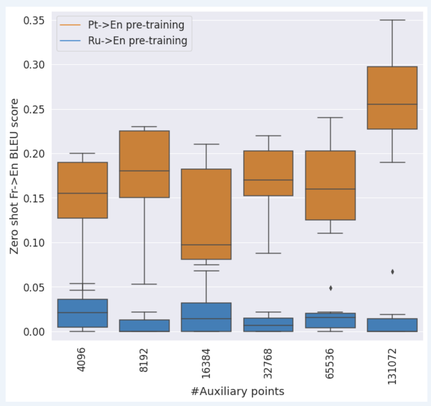
- **Main task:** French -> English (Fr -> En) translation
- **Aux. tasks:** Portuguese -> English (Pt -> En) & Russian to English (Ru -> En).
- **Main task dataset sizes:** [4096, 8192, 16384, 32768, 65536, 131072]
- **Aux. task dataset sizes:** [0, 4096, 8192, 16384, 32768, 65536, 131072]
- **Data:** screened language learning sentence pairs from the Tatoeba project [1].
- **Networks:** Transformer models mostly following details from the original transformer paper [2].
- **Metric:** BLEU scores [3] on auxiliary and main task test sets after pre-training and again after fine-tuning.

We conduct 8-fold cross-validation (CV) over all combinations of the 2 aux. tasks, 6 main dataset sizes, and 7 aux. dataset sizes. This results in a total of 672 networks over 84 conditions.

BASELINE AND ZERO-SHOT RESULTS

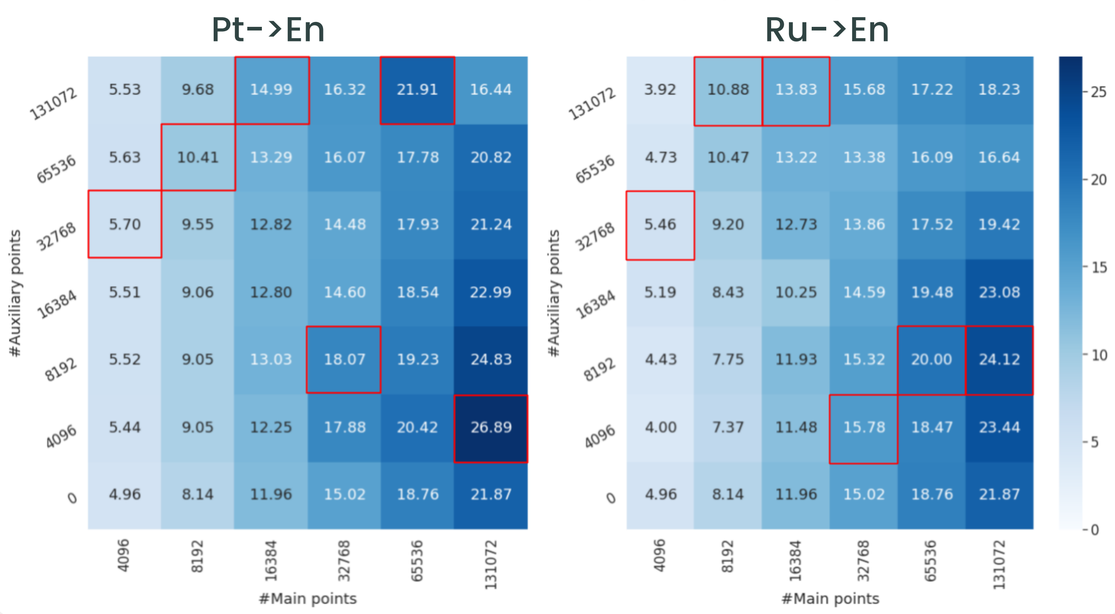


More (similar quality) data is always better for **baseline (bilingual) models.** However, returns are diminishing

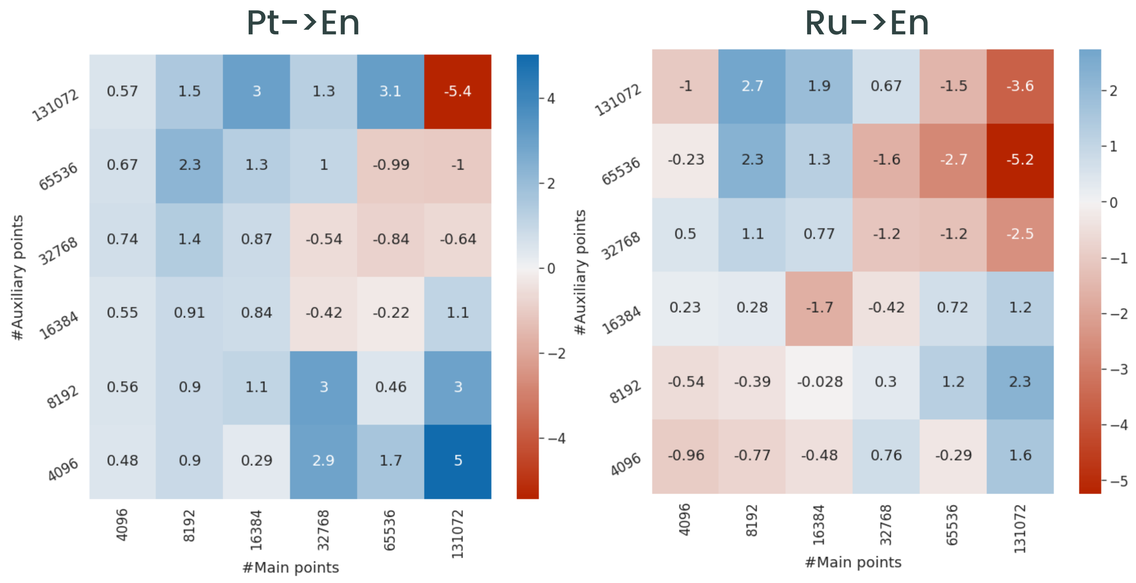


Closely-related data is always better for **zero-shot performance**.

MEAN BLEU SCORES



EFFECTS OF PRE-TRAINING



Raw mean BLEU shows initial trends, but #Main points significantly outweighs #Auxiliary points in effect on performance. Subtracting baseline bilingual results from each column, trends become clearer:

- For small main datasets, more aux. data is better up to ~8x#main points, at which point it is detrimental.
- Threshold between 16k-32k main points after which limited pre-training is beneficial while too much decreases performance below the baseline.
- More (similar quality) data is not always better.

EFFECTS OF OPTIMAL PRE-TRAINING

- Closely related data is typically better than distantly related data when the dataset sizes are well-chosen ('well-chosen' ≠ 'large').
- Optimal dataset sizes can matter more than similarity.

FUTURE WORK

- Extend studies to more main and auxiliary tasks (particularly to true low-resource languages) to investigate how results generalise.
- Replicate study to confirm results.
- Balance datasets before tokenization and limit effects of negative transfer for larger datasets.

RELATED WORK

Other work by the authors to improve low-resource NMT is aimed at facilitating easier, faster, and less error-prone research, and facilitating independent NMT research by language researchers:

- graphicalNMT: A system for code-free design, creation, and evaluation of NMT systems
- NLPDataTools: A system for code-free evaluation, analysis, and manipulation of NLP datasets.

REFERENCES
[1] http://www.manythings.org/anki/
[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems, pages 5998– 6008.
[3] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE
THE UNIVERSITY OF QUEENSLAND AUSTRALIA