Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1437–1443
Marseille, 20-25 June 2022

A Systematic Study Reveals Unexpected Interactions in Pre-Trained Neural Machine Translation

Ashleigh Richardson, Janet Wiles
The University of Queensland
Queensland, Australia
{a.richardson, j.wiles}@uq.edu.au

Abstract

A significant challenge in developing translation systems for the world's ∼7,000 languages is that very few have sufficient data for state-of-the-art techniques. Transfer learning is a promising direction for low-resource neural machine translation (NMT), but introduces many new variables which are often selected through ablation studies, costly trial-and-error, or niche expertise. When pre-training an NMT system for low-resource translation, the pre-training task is often chosen based on data abundance and similarity to the main task. Factors such as dataset sizes and similarity have typically been analysed independently in previous studies, due to the computational cost associated with systematic studies. However, these factors are not independent. We conducted a three-factor experiment to examine how language similarity, pre-training dataset size and main dataset size interacted in their effect on performance in pre-trained transformer-based low-resource NMT. We replicated the common finding that more data was beneficial in bilingual systems, but also found a statistically significant interaction between the three factors, which reduced the effectiveness of large pre-training datasets for some main task dataset sizes (p-value < 0.0018). The surprising trends identified in these interactions indicate that systematic studies of interactions may be a promising long-term direction for guiding research in low-resource neural methods.

Keywords: Machine Translation, Evaluation Methodologies, Less-Resourced Languages, Transfer Learning, Pre-training

1. Introduction

State-of-the-art techniques for neural machine translation (NMT), such as transformers (Vaswani et al., 2017), have raised societal expectations of language technologies. However, the significant data requirements for deep learning make it difficult to meet these expectations for many low-resource languages. This difference in access to technology exacerbates existing systemic inequality. The nature of low-resource languages makes it difficult or impossible to collect enough data for blackbox application of these techniques; for instance, many languages have very few (or no) remaining speakers. In some cases, the hardware required to train or pre-train on hundreds of millions of sentence pairs is not available to the communities who require this technology. Nonetheless, the development of machine learning systems for these languages is interesting for reasons of cultural promotion, accessibility, and as a means for testing the limits of technology.
Transfer learning can be used to boost performance on a low-resource task, by leveraging information from one or more high-resource tasks. However, transfer learning introduces many new variables that are not trivial to decide. For instance, consider the simplest case of pre-training on an auxiliary language pair before fine-tuning on the intended (main) language pair. In choosing an auxiliary task, we should consider task similarity, and amount of available data for both main and auxiliary tasks. However, the amount of computationally expensive trial-and-error involved in choosing an appropriate auxiliary task is a significant limiting factor in low-resource NMT. Previous work on guiding transfer learning in practical NMT mostly used older models, or examined variables independently. However, we show that by systematically studying the interactions between auxiliary dataset size, main dataset size, and task similarity in a low-resource setting, we can identify non-trivial, statistically significant interactions which affect performance in unexpected ways. This result demonstrates that systematic studies should be used for probing the effects of interactions between auxiliary task variables and guiding decision making in low-resource neural methods.

2. Background

2.1. Neural Methods for Machine Translation

Neural methods for machine translation often use an encoder-decoder structure. Traditionally, the encoder and decoder would both use some sort of recurrent neural network (RNN) architecture, such as a long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997) or a gated recurrent unit (GRU) (Cho et al., 2014). However, because these models maintain a hidden state across time, they become very deep and suffer from the vanishing/exploding gradient problem (LeCun et al., 2015).
Transformers (Vaswani et al., 2017) are the current state-of-the-art for machine translation. These networks are autoregressive, but not recurrent, so they overcome the issues of RNN-based models while retaining an ability to process sequences. However, these networks are still deep, and so effective learning still requires large datasets and extensive computational

power.

2.2. Pre-Training

Transfer learning techniques can be used to leverage unlabelled (Raina et al., 2007), related (Radford et al., 2019; McCann et al., 2018; Firat et al., 2017; Dong et al., 2015) or even unrelated (Romera-Paredes et al., 2012) data to reduce labelled data requirements. Pre-training is one approach to transfer learning, in which a system is first trained (pre-trained) on one or more auxiliary tasks before being fine-tuned on the actual (main) task. The process of pre-training in theory initializes the model well, meaning that less main task data is required. However, in practice, care is required, as pre-training can also lead to negative transfer (Wang et al., 2019); a phenomenon where prior out-of-domain knowledge interferes with the learning of new knowledge, reducing performance below the single task baseline.

2.3. Subword Segmentation and Tokenization

Before using sentences as input to a neural network, a subword segmentation algorithm is typically trained on the training data to split the sentence into 'subwords'. This process helps the network handle unseen or rare words, and allows the network to learn the meanings of useful subwords (e.g. 'un-', 'dis-', '-ly', '-s'). The SentencePiece library (Kudo and Richardson, 2018) implements many common subword segmentation strategies, such as byte-pair encoding (BPE) (Sennrich et al., 2016), unigram language models (Kudo, 2018), word piece, or character encodings.

When pre-training on a different language pair, tokens can change between tasks. Joint tokenization (Xiao et al., 2010) can allow for knowledge of shared tokens to be transferred. However, if datasets are unbalanced, joint tokenization can cause too few tokens to be allocated to the lower-resource task, particularly if there is little to no lexical overlap.

2.4. Task Similarity

The similarity between the auxiliary and main task is an important factor in transfer learning, however, there is no principled system for determining task similarity in translation (Ruder, 2017). While typology is generally a good metric for similarity, lexical overlap and auxiliary dataset size are (individually) more important for machine translation specifically (Lin et al., 2019). For the purpose of this study, since the target languages remain the same between auxiliary and main tasks, we define task similarity to be the lexical similarity between the auxiliary task source language and the main task source language.

3. Related Work

Prior research (Rodriguez et al., 2009; Morishita et al., 2017; Qi et al., 2018; Neishi et al., 2017) has provided empirical evidence for simple decision making in NMT, such as selection of batch size and number of folds for cross-validation, initialization strategy for embedding layers, and guidance on the appropriate situations to use pre-trained embeddings. However, little work exists to guide use of pre-training to aid low-resource NMT, and existing work examined variables independently, rather than as an interacting system. (Qi et al., 2018) report a 'sweet spot' (low but not too low-resource) in the efficacy of pre-trained word embeddings to improve low-resource NMT, but only scaled the main task dataset size, and used an RNN-based model. While the authors investigated the effects of relatedness between source and target languages within individual tasks, they did not consider the relatedness between main and auxiliary translation tasks.

Recently it was shown, via the first systematic study of negative interference, that low-resource languages can also be affected by negative interference when training multilingual models for fine-tuning on other NLP tasks (Wang et al., 2020). This is contrary to the popular belief that, in multilingual systems, negative interference affects the high-resource languages, while low-resource languages benefit from the additional auxiliary data. While (Wang et al., 2020) did not evaluate performance on NMT and did not study interactions, it did motivate the need for further study of how networks behave in controlled, systematic environments.

4. Experiments

The aim of this study was to investigate the interacting influence of language relatedness, and auxiliary and main task data set sizes on performance when pre-training transformer models for NMT with limited data. We performed a three-factor experiment ($2 \times 6 \times 7$) over the following variables:

- Auxiliary tasks (2): {Portuguese $\rightarrow$ English, Russian $\rightarrow$ English}
- Main task (French $\rightarrow$ English) dataset size (6): {4096, 8192, 16384, 32768, 65536, 131072}
- Auxiliary task dataset size (7): {0, 4096, 8192, 16384, 32768, 65536, 131072}

We keep both main and auxiliary task dataset sizes very small to moderately small to demonstrate how the factors interact in limited data, and limited hardware settings. These dataset sizes are sufficient to demonstrate interesting insights, and expansion to higher resource auxiliary task sizes is proposed as future work.

4.1. Datasets

While this research is part of a larger goal to provide guidance for selecting parameters for transfer learning in low-resource language tasks, those languages do not have sufficient data for factorial studies. Hence, large open source datasets for French (Fr), English (En), Portuguese (Pt) and Russian (Ru) were used to enable *systematic* exploration of the three factors. All networks

were trained to perform Fr → En translation as a main task. The dataset sizes were systematically scaled to reproduce the effects of variable levels of low-resource tasks, from very small to moderately large, enabling us to examine the effects of combinations of dataset sizes. Note that such systematic studies are not possible with actual low-resource languages, where thresholds between low and moderate dataset sizes cannot be examined. Pt → En and Ru → En translation were taken as closely and distantly related (lexically dissimilar) auxiliary tasks respectively. Datasets come from ManyThings.org [1], and are screened bilingual sentence pairs from the Tatoeba project ranging from short (e.g. two or three words) to very long (over 60 words). These language-learning datasets were selected to bias the network towards practical low-resource translation. Dataset sizes divide evenly by 512 for 8-fold cross-validation (CV) with a batch size of 64. Datasets were selected only from sentences that were 25 words or less. Because all sentences must be padded to the same length, and complexity of self-attention is $O(n^2)$ with respect to sequence length n (Vaswani et al., 2017), enforcing a maximum sentence length can vastly improve efficiency. Test sets of 1000 sentence pairs were drawn at random, and then from the remaining sentence pairs, datasets of the required sizes were drawn at random and remained consistent throughout all experiments. Subword segmentation was performed using the SentencePiece implementation of word piece, as we found through pilot studies that this strategy consistently performed the best on the datasets used.

4.2. Data Collected

We recorded training, validation, and test categorical cross-entropy loss, training time, and number of epochs to convergence. Main and auxiliary test BLEU scores (Papineni et al., 2002) were recorded after pre-training and again after fine-tuning.

4.3. Procedure and Network Architecture

All experiments used transformer models following the architecture and training details in (Vaswani et al., 2017), except as follows:

- Batch size: 64
- K-fold CV: K=8
- Maximum sentence length: 25 words
- Patience: 10 epochs
- Embedding dimensions: 512

Joint sentence piece tokenization over source languages was used (Kudo and Richardson, 2018), with max. vocab. size $= 35,000$; though most vocabularies were smaller than this limit. Separate sentence piece tokenization was used for the target language. When decoding at test time, greedy search was used for efficiency due to the number of models being trained. All networks were implemented in Tensorflow 2 and trained on a GPU cluster. Fig. 1 describes the experimental procedure.

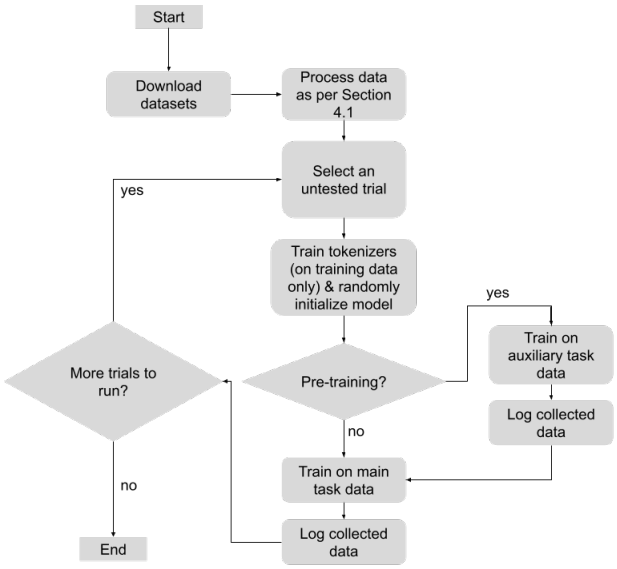

Figure 1: Experimental procedure.

5. Results and Discussion

5.1. Baseline Results

Bilingual (no pre-training) results revealed a logarithmic relationship between amount of training data and BLEU score; see Fig. 2 As expected, more data of a similar quality is beneficial in a bilingual context; though returns diminish as the size of the main dataset increases.

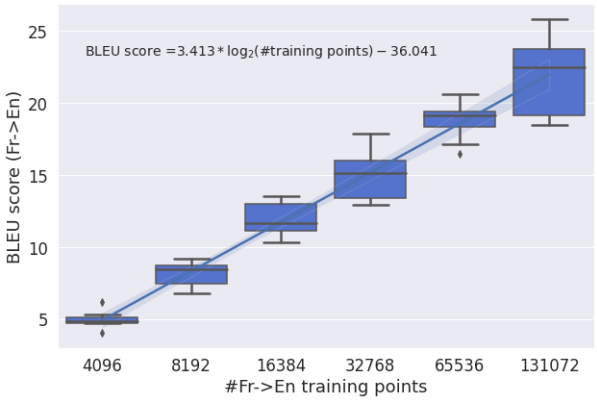

Figure 2: BLEU score vs. dataset size in baseline (bilingual) models. At each number of training points BLEU scores from 8-fold cross-validation on the Fr → En test set are used. Note the logarithmic scale on the x-axis.

[1] https://www.manythings.org/anki/

5.2. Effect of Pre-Training on Main Task Performance

Average Fr → En BLEU scores were calculated over 8 folds on all combinations of the three factors; see Fig. 3 and Fig. 4. To highlight the role of pre-training, baseline average BLEU for each #main points is subtracted from the BLEU scores achieved by models fine-tuned on that #main points; see Fig. 5 and Fig. 6. There is a general trend that more auxiliary data is beneficial when main task data is insufficient (4k to 16k points), with best results at #aux. points $\approx 8\times$ #main points. Beyond this #aux. points, performance decreases. We conjecture that unbalanced tokenization could be the main cause (more research is required, see Section 6). This effect is emphasized in lexically dissimilar (Ru → En) pre-training. As #main points passes a threshold between ∼16k and ∼32k, a small amount of pre-training is beneficial while too much decreases performance below the baseline. This result is likely due to a threshold at which negative interference outweighs the benefits of transfer learning, giving clues as to how much auxiliary data to use for maximum benefit. Overall, the 3-way interaction between the factors was statistically significant (p-value ≈ 0.00178).

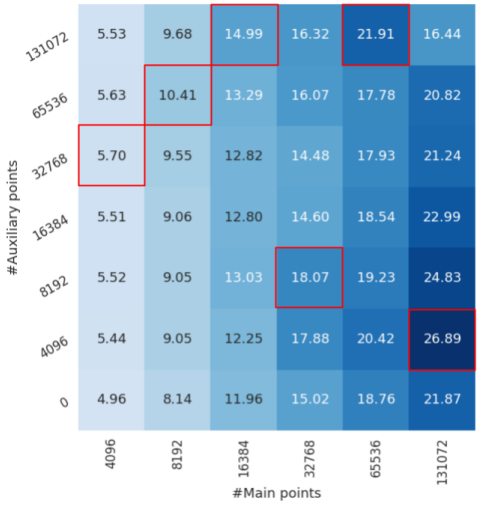

Figure 3: Interactions between auxiliary and main task dataset sizes with Pt → En pre-training. Values shown are BLEU scores (averaged over 8 folds) achieved on the Fr → En test set by transformers pre-trained on the specified amount of auxiliary data, and then fine-tuned on the specified amount of main task data.

5.3. Zero-Shot Performance

As expected, zero-shot test Fr → En BLEU scores were poor for both auxiliary tasks, but were consistently higher after pre-training on the closely related language pair, compared to the distantly related pair; see Fig. 7. Note that we have excluded two outliers for Pt → En pre-training to make the graph easier to read.

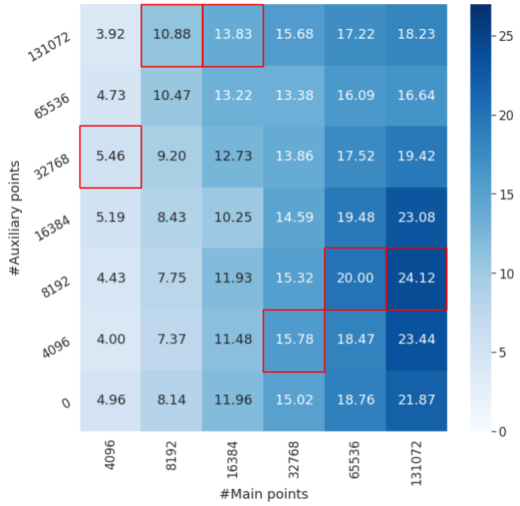

Figure 4: As per Fig. 3, but using Ru → En pre-training, rather than Pt → En.

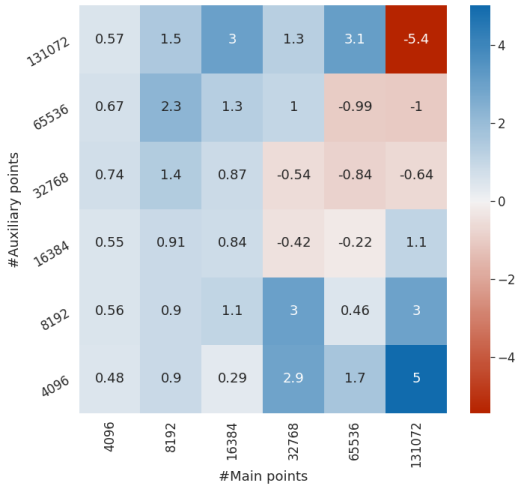

Figure 5: Impact of dataset sizes on effectiveness of transfer in Fr → En NMT with Pt → En pre-training. We transform Fig. 3 by subtracting baseline results for each #main points from results of pre-trained models that were fine-tuned on the same #main points. Initial trends indicate a 'threshold' between ∼16k and ∼32k data points, below which more auxiliary data ($\approx 8\times$ #main points) is better, and above which fewer auxiliary data points (but > 0) is better. Note the better performance in the top left and bottom right quadrants.

5.4. Overall Effect of Optimal Pre-Training

With fewer data points, optimal pre-training (highest average BLEU for each #main points achieved with any #aux. points) was equally beneficial with either task. As #main points increased, language relatedness helped boost performance; see Fig. 8.

5.5. Impact of Language Relatedness

While Pt → En provided more effective pre-training, Ru → En was surprisingly competitive (best score

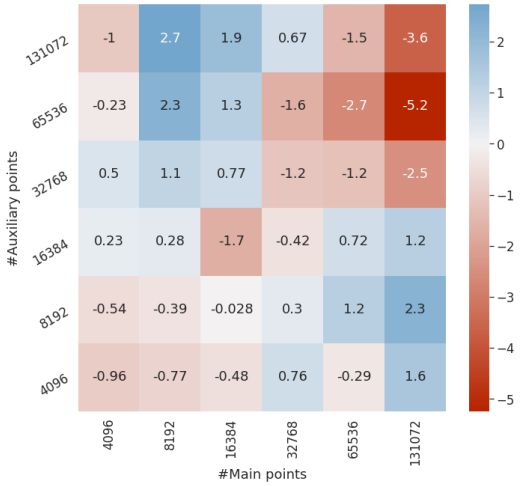

Figure 6: As per Fig. 5, except with Ru → En as the auxiliary task. Note that the same trends remain even with a different pre-training task.

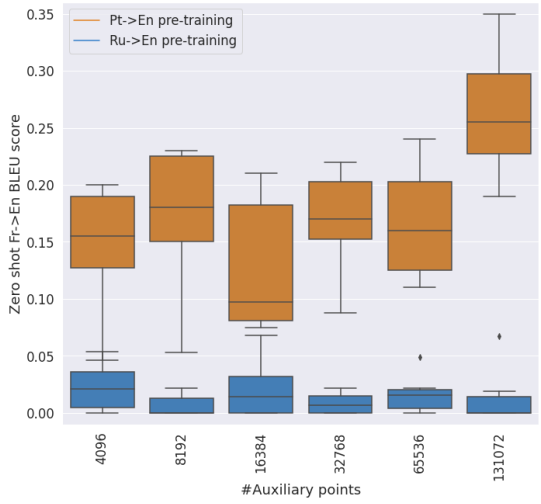

Figure 7: Zero-shot BLEU scores on the Fr → En test set after pre-training on different #auxiliary main points. Box and whisker plots shows average (over folds) zero-shot BLEU scores achieved on each #main points at the given #auxiliary points. Zero-shot performance is consistently higher with the closely-related auxiliary task.

26.89 for Pt compared to 24.12 for Ru) (see Fig. 3 and Fig. 4). This result may come from expanding the target-side monolingual corpus, or from the morphological similarity of Russian and French allowing for morphosyntactic information transfer via attention mechanisms.

5.6. Is Language Relatedness More Important than Auxiliary Dataset Size?

Optimal amounts of closely related data was almost always better than optimal amounts of distantly related data, and where it was not, the change in performance boost was small (see Fig. 8). An interesting practical

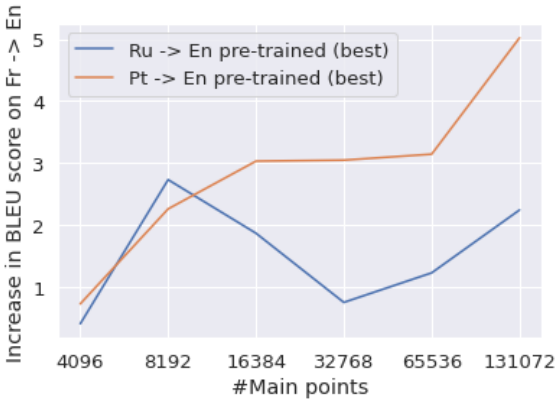

Figure 8: Best average improvement in BLEU score over the bilingual baseline achieved at each #main points for pre-trained models. At each #main points, the improvement from the most beneficial (not necessarily the largest) #auxiliary points was chosen for each auxiliary task. For small #main points (4k-8k) performance is similar for both auxiliary tasks, whereas for medium to large #main points, there is a significant benefit to using highly-related well-chosen auxiliary tasks.

question is whether large amounts of distantly related data is better than small amounts of highly related data. For each #main points, the best Ru → En (a less related dataset) pre-trained models outperformed the majority of Pt → En pre-trained models fine-tuned on the same #main points (22 of the 36 non-baseline Pt → En pre-trained trials); see results in red squares in Fig. 4. However, the systems pre-trained on the most (∼131k) Ru → En data before fine-tuning only outperformed Pt → En pre-trained systems fine-tuned on the same #main points in 14 of the 36 trials. Almost all of these 14 cases occurred when #main points was between ∼8k and ∼16k, implying that if the amount of main task data is small, but not too small, large auxiliary datasets are more important than highly related auxiliary data. Once the main task dataset is sufficiently sized, however, it is better to use highly related data, even if the quantity is limited. Overall, auxiliary dataset size can matter more than language relatedness but only when well-chosen, with a distinction between 'well-chosen' and 'large'.

6. Limitations and Future Work

This study provides initial insight into the non-trivial nature of the *interactions* between pre-training decisions. However, there are many avenues to extend these results. Computational cost is a limiting factor on any systematic analysis of transformer models, and in our study common settings were used for hyperparameters and network architecture, and greedy search was used in decoding. Extensions that could improve performance include beam search in decoding (Freitag and Al-Onaizan, 2017), balancing datasets before to-

kenization, and mitigating negative transfer effects as dataset sizes grow. Such investigations could also provide insight into the extent to which unbalanced tokenization and negative transfer explain results, and the role of statistical anomaly can be investigated through further replication of the study. Further work is required on a range of languages including evaluating the extent to which the current results generalize to truly low-resource tasks. Additionally, future work should investigate how systematic studies of interactions benefit other domains in natural language processing, transfer learning, and the broader field of machine learning.

7. Conclusion

The current results point to a threshold in transfer learning benefits which in the languages tested appears between $\sim$ 16k and $\sim$32k main task data points. Below this threshold, more auxiliary data is beneficial, provided the auxiliary and main task datasets are sufficiently balanced. Above this threshold, less auxiliary data ($\leq\sim$16k) is better. It is better to use a closely related task if the same amount of data is available. However, in practical NMT, it is not always the case that a highly related auxiliary task will have an abundance of data in the same way that distantly related tasks might. Our results demonstrate that it can be preferable to use well-chosen amounts of distantly related data, rather than suboptimal amounts of closely related data. Overall, our work highlights that decisions in low-resource pre-trained NMT interact in non-trivial ways, and provides initial evidence into how the use of systematic studies of interactions can benefit the development of low-resource machine learning technologies.

8. Bibliographical References

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

Firat, O., Cho, K., Sankaran, B., Vural, F. T. Y., and Bengio, Y. (2017). Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Freitag, M. and Al-Onaizan, Y. (2017). Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., et al. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Morishita, M., Oda, Y., Neubig, G., Yoshino, K., Sudoh, K., and Nakamura, S. (2017). An empirical study of mini-batch creation strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 61–68, Vancouver, August. Association for Computational Linguistics.

Neishi, M., Sakuma, J., Tohda, S., Ishiwatari, S., Yoshinaga, N., and Toyoda, M. (2017). A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 99–109.

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.

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana, June. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *URL https://openai.com/blog/better-language-models*.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM.

Rodriguez, J. D., Perez, A., and Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):569–575.

Romera-Paredes, B., Argyriou, A., Berthouze, N., and Pontil, M. (2012). Exploiting unrelated tasks in multi-task learning. In *International conference on artificial intelligence and statistics*, pages 951–959.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

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.

Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. (2019). Characterizing and avoiding negative transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11293–11302.

Wang, Z., Lipton, Z. C., and Tsvetkov, Y. (2020). On negative interference in multilingual language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Xiao, X., Liu, Y., Hwang, Y.-S., Liu, Q., and Lin, S. (2010). Joint tokenization and translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1200–1208.