Data and Parameter Scaling Laws for Neural Machine Translation

Anonymous ACL submission

Abstract

We observe that the development cross-entropy loss of supervised neural machine translation models scales like a power law with the amount of training data and the number of non-embedding parameters in the model. We discuss some practical implications of these results, such as predicting BLEU achieved by large scale models and predicting the ROI of labeling data in low-resource language pairs.

1 Introduction

As training neural networks becomes an organizational and multi-million dollar venture (Brown et al., 2020), it is imperative to quantifiably predict the benefits of scaling up neural networks. In this paradigm, machine learning is an engineering effort, in which money can buy resources (data, compute) and the main concern is to predict return-on-investment (ROI) while avoiding bottlenecks.

Recent work has observed that the cross entropy loss of neural language models and other autoregressive generative models scales like a power law in the amount of training data, compute, and number of model parameters over several orders of magnitude (Hestness et al., 2019; Kaplan et al., 2020; Henighan et al., 2020). Similar intuitions exist in the realm of supervised MT: doubling the amount of parallel training data leads to roughly a fixed improvement in BLEU in both phrase-based statistical MT (Irvine and Callison-Burch, 2013; Turchi et al., 2008) and neural MT (Koehn and Knowles, 2017; Sennrich and Zhang, 2019).

In Section 2, we show that these MT intuitions can similarly be quantified as power law scaling; using a handful of experiments on small subsets of MT datasets, we precisely predict the performance (both cross-entropy loss and BLEU) of large systems trained on orders of magnitude more data. In Section 3, we demonstrate how these trends might be utilized to make ROI predictions when annotating more data for low-resource language pairs.

2 Machine Translation Scaling Laws

To investigate the predictability of MT system performance as parameters/data increase, we train many Transformers of various sizes (Table 2) on randomly selected subsets of data ($\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$) for several standard MT datasets. The smallest data subsets contain $\sim$0.1% of the total data available.

We use three language pairs in our experiments:

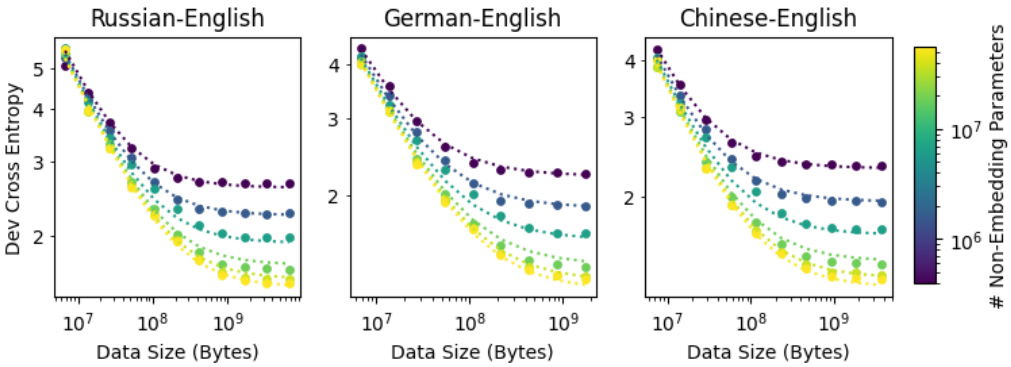

Figure 1: The development cross-entropy vs. the amount of data used to train each model from Section 2. Each point is colored by the number of non-embedding parameters in the model. The best fit of Equation 1 via least-squares regression for each language pair is shown as the dotted line.

Lang Pair	# Train Examples	α_N	$\log N_C$	α_D	$\log D_C$
Ru-En	51.1 M	0.11	21.62	0.38	13.81
De-En	28.3 M	0.13	18.81	0.35	13.43
Zh-En	35.9 M	0.12	19.90	0.43	12.73

Table 1: The number of training examples for each language pair, along with our estimates of the parameters of Equation 1 for those language pairs.

Layers	Hidden Dim	Non-Embed Params
2	128	393k
2	256	1.5M
2	512	6.3M
6	512	19M
12	512	38M
12	624	56M

Table 2: Number of non-embedding parameters of each model trained. The feed-forward size is $4\times$ the hidden size. Layers are allocated evenly between the encoder and decoder. Non-embedding parameters is roughly $12Ld^2$, where L is the number of layers and d is the hidden size.

German-English (de-en), Russian-English (ru-en), and Chinese-English (zh-en). The data for each language pair is a concatenation of WMT 2017 data (which includes news commentary, parliamentary proceedings, and web-crawled data) and OpenSubtitles2018 (Lison and Tiedemann, 2016; Tiedemann, 2016). Datasets are tokenized using the Moses[1] tokenizer, after which a 30k BPE vocabulary is constructed using the full dataset. For evaluation, we use newstest2016 concatenated with the last 2500 lines of OpenSubtitles2018.

Transformers are trained with early stopping and a learning rate of 0.0002[2] with a plateau-reduce schedule for a maximum of 350k updates. Other training details can be found in the code supplement.[3] The resulting losses are plotted in Figure 1, with model sizes ranging from 393k-56M parameters and data sizes from 40k-50M lines of text.

2.1 Cross-entropy vs. Data/Parameters

Kaplan et al. (2020) provide an ansatz that predicts cross-entropy loss given the amount of training data and the size of the neural model:

$$L(N, D) = \left[\left(\frac{N_C}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_C}{D} \right]^{\alpha_D} \quad (1)$$

[1] statmt.org/moses

[2] Determined via a brief grid search at medium model sizes.

[3] Anonymized code is attached to the submission.

Lang	Max Data %	$\lvert\alpha_N - \alpha'_N\rvert$	$\lvert\alpha_D - \alpha'_D\rvert$
ruen	6.25	0.014	0.013
	3.125	0.024	0.026
deen	12.5	0.003	0.004
	6.25	0.009	0.016
zhen	6.25	0.003	0.003
	3.125	0.006	0.009

Table 3: Difference between scaling exponents when using the full dataset (α_N,α_D) vs. estimating the scaling exponents using only models trained on smaller subsets of the data (α'_N, α'_D). We see that even when using 3-6% of the data, the best-fit scaling exponents of Equation 1 stay very similar.

where L is the per-token development cross-entropy loss (in nats), N is the number of non-embedding parameters, D is the amount of training data (in bytes), and α_N, α_D, N_C, and D_C are constants determined by the particulars of the data distribution and training setup.[4]

Figure 1 shows that this equation is highly predictive of our results.[5] The predictions are also fairly stable; Table 3 shows that the best-fit parameters of this equation stay similar even when restricting ourselves to using only 3-6% of the data. In Appendix A we perform a retrospective analysis of the results from Zhang and Duh (2020) to give some insight into how different hyper-parameter settings may influence scaling coefficients.

As either N or D approaches infinity, $L(N, D)$ simplifies to a "pure power law" in the other variable, which looks like a straight line on a log-log graph. For example, if we assume all models are large enough that data becomes the main performance bottleneck, then:

$$L(D) = \left(\frac{D_C}{D} \right)^{\alpha_D} \quad (2)$$

We will use this assumption later when dealing with very low-resource language pairs.

2.2 BLEU vs. Cross-Entropy Loss

Predicting cross-entropy loss by itself does not tell us much about the quality of the translation system; we would really like to predict the achieved

[4] It is unclear why these empirical trends hold so widely for auto-regressive modeling. Power law scaling can arise in complex systems for a variety of reasons (Hanel et al., 2018); Sharma and Kaplan (2020) suggest that scaling exponents may be related to the intrinsic dimension of the data manifold.

[5] The best-fit parameters are shown in Table 1; they are remarkably similar for each language pair, which may be attributed to the similarity of the domains each dataset was sourced from.

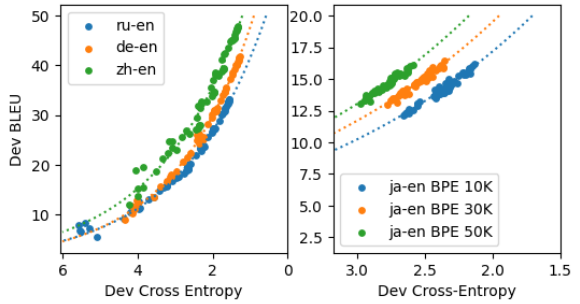

Figure 2: (Left) BLEU exponentially decays as cross-entropy increases. Different language pairs may have different exponents and constants. (Right) A retrospective analysis of Zhang and Duh (2020) (Appendix A). Even the same development dataset with different BPE applied may have a different constant multiplier.

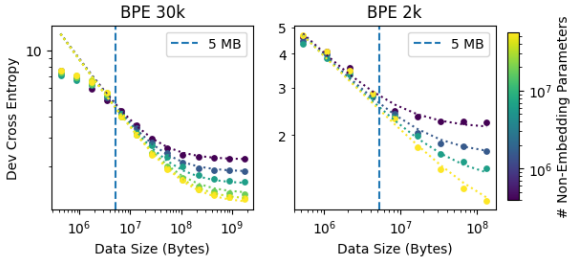

Figure 3: (Left) Models trained on <5 MB of data (around 40k lines) fall off-trend when using a BPE vocabulary of 30k, plateauing to an apparent maximum cross-entropy. (Right) When encoding the same dataset with a BPE of 2k, the plateau is rectified and returns to power-law scaling. Similar plots are shown for ru-en and zh-en in Appendix Figure 7.

BLEU score, which is more interpretable to humans as a measure of adequacy, fidelity, and fluency (Papineni et al., 2002). Figure 2 shows that the relationship between BLEU and cross-entropy can vary between different language pairs and BPE settings. However, when these factors are fixed, BLEU seems to exponentially increase as cross-entropy decreases:

$$\text{BLEU}(L) \approx Ce^{-kL} \tag{3}$$

This relationship is fairly predictable for high BLEU values, but becomes noisier as BLEU drops below 15. Notably, changing the BPE encoding does not seem to affect k, but does change the multiplying constant C.[6]

Why should this relationship be exponential? We might gain some insight by re-writing Equation 3 in terms of the per-token perplexity (P):

$$\text{BLEU}(P) \approx C\left(\frac{1}{P}\right)^k \tag{4}$$

where $(1/P)$ can intuitively be interpreted as the expected unigram precision of an autoregressively sampled translation with the same length as the reference sentence (Manning and Schutze, 1999). This is only intuition, however: in practice, we do not sample translations but decode using beam search, and BLEU combines multiple *modified* n-gram precisions besides unigram precision.[7]

[6]We evaluate BLEU using multi-bleu.perl from the Moses toolkit. De-bpe-ing, de-tokenizing, and using Sacrebleu (Post, 2018) adds a small amount of noise but does not qualitatively change our results. See Appendix Figure 5.

[7]The relationship between precision and perplexity for higher values of n is not clear. In general, expected bigram precision $\neq (1/P)^2$.

2.3 Preventing Breakdown At Smaller Dataset Sizes

Some extremely low-resource MT datasets (which we examine in Section 3) can have less than 5 MB of data (∼40k sentence pairs). Figure 3 shows that when we extend our previous experiments to datasets smaller than this size, using 0.05% - 0.0125% of the data, the data scaling power law seems to break down, casting doubt on our ability to extrapolate extremely low-resource results to medium and high-resource data regimes.

However, the results are not simply noisy but predictably plateau to an apparent ceiling of 7.8 nats. For reference, a unigram language model trained on only the English part of the training data (with a 30k BPE vocab) achieves a per-token cross-entropy of ∼7 nats. This leads us to suspect that models in this data regime are learning to rely on simple unigram statistics that do not change much as we decrease the data size.

Using a much smaller BPE vocabulary of 2k tokens rectifies this plateau and returns to power law scaling, even with datasets <5 MB. We believe this is because the smaller vocabulary makes it difficult to exploit unigram statistics for rare words. While this is not conclusive evidence, we recommend that cross-entropies near or above unigram LM performance should not be relied upon to extrapolate performance. Dataset subsets which contain less than half of the BPE vocabulary should similarly be avoided.[8]

[8]In ∼5 MB datasets, around half of the 30k BPE vocabulary is never seen during training. In contrast, ∼10 MB datasets contain almost every vocab token. See Appendix Figure 6.

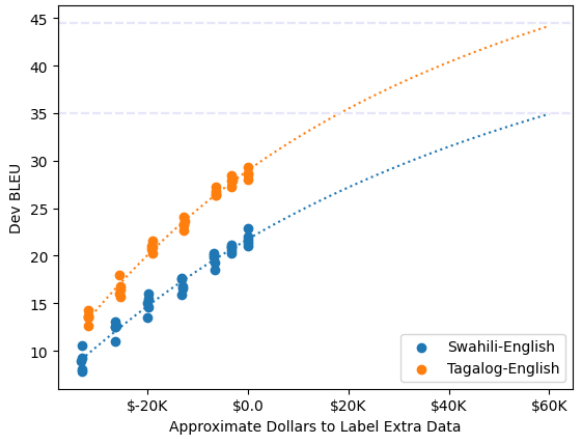

Figure 4: USD-to-bleu projections for low-resource language pairs, with a training setup similar to Section 2.[10] We assume each byte of data costs about 0.01 USD to acquire.[9] Negative dollars represent using less data than is currently available, whereas positive dollars represents our projections if we were to spend that much USD on acquiring more data.

3 Predicting ROI of Annotating Low-Resource Language Pairs

If we assume that data is the main performance bottleneck (as it is in many low-resource language pairs), we can plug Equation 2 into Equation 3 to directly model the relationship between BLEU and data size:

$$\text{BLEU}(D) = C \exp \frac{K}{D^{\alpha_D}} \tag{5}$$

where $K = k(D_C)^{\alpha_D}$. This can be further combined with the hourly cost of fluent human translators to give us an approximate USD-to-BLEU trade-off when annotating more data for low-resource language pairs.

Figure 4 shows some example projections for Tagalog-English (tl-en) and Swahili-English (sw-en), with each dataset containing less than 50k sentence pairs (Zavorin et al., 2020). Under some assumptions about the costs of human translation[9], we predict that spending ∼$60k USD to acquire more tl-en/sw-en data (which would roughly double the size of the either dataset) would lead to an improvement of around 10-15 BLEU.

[9]We assume translation costs around 0.10 USD per word, each word is composed of 5 characters on average, and each character requires around a byte of space.

[10]We train a 12 layer model using a 2k BPE dataset subsets (100%, 90%, ..., 50%) with five different data shuffling seeds. We also increase the checkpoint frequency for earlier stopping.

3.1 Limitations

There is a reasonable amount of noise in the cross-entropy/BLEU relationship at this scale (shown in Appendix Figure 8) which limits the precision and reliability of these predictions. In practice, we expect small amounts of data can be acquired in batches and predictions can be re-evaluated before deciding to continue. However, these predictions give a general sense of the cost of progress in low-resource machine translation. When engineering a real-world system, the simple option of acquiring more data and predictably improving performance should always be carefully weighed against more complicated and less predictable options.

That being said, predictably achieving a high BLEU score on a test dataset is not equivalent to "solving translation" for that language pair. Under-specification (D'Amour et al., 2020) still poses a challenge for effectively evaluating machine translation systems in real-world scenarios, especially in low-resource language pairs where evaluation data is usually from a narrow domain. More robust evaluation methods are needed, and it is not clear whether the output of these methods will be as predictable as cross-entropy loss or BLEU.

And finally, while our work demonstrates empirical power law scaling of NMT systems, it does not attempt to provide any causal explanation for these results. We also do not investigate the specific training factors that lead to a particular scaling exponent, but we expect this to be a fruitful research direction for future exploration.[11]

4 Conclusion

We have shown that supervised neural machine translation performance with Transformers scales like a power law in non-embedding parameters and training data, aligning with similar observations in unsupervised auto-regressive modeling. We've also seen that as development cross-entropy decreases, BLEU exponentially increases. These two relationships can be combined to predict an effective USD-to-BLEU trade-off when annotating more data, even in low-resource regimes.

[11]For example, Kasai et al. (2020) have found that a deep encoder and a shallower decoder can be more efficient, which may lead to better parameter scaling.

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are Few-Shot learners.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D Sculley. 2020. Underspecification presents challenges for credibility in modern machine learning.

R. Hanel, S. Thurner, and P. Klimek. 2018. Introduction to the theory of complex systems.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. 2020. Scaling laws for autoregressive generative modeling.

Joel Hestness, Newsha Ardalani, and Greg Diamos. 2019. Beyond Human-Level accuracy: Computational challenges in deep learning.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270, Sofia, Bulgaria. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. *arXiv preprint arXiv:1906.11943*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting Low-Resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Utkarsh Sharma and Jared Kaplan. 2020. A neural scaling law from the dimension of the data manifold.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522.

Marco Turchi, Tijl De Bie, and Nello Cristianini. 2008. Learning performance of a machine translation system: a statistical and computational analysis. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 35–43, Columbus, Ohio. Association for Computational Linguistics.

Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, and Richard Tong. 2020. Corpora for cross-language information retrieval in six less-resourced languages. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 7–13, Marseille, France. European Language Resources Association.

Xuan Zhang and Kevin Duh. 2020. Reproducible and efficient benchmarks for hyperparameter optimization of neural machine translation systems. *Transactions of the Association for Computational Linguistics*, 8:393–408.

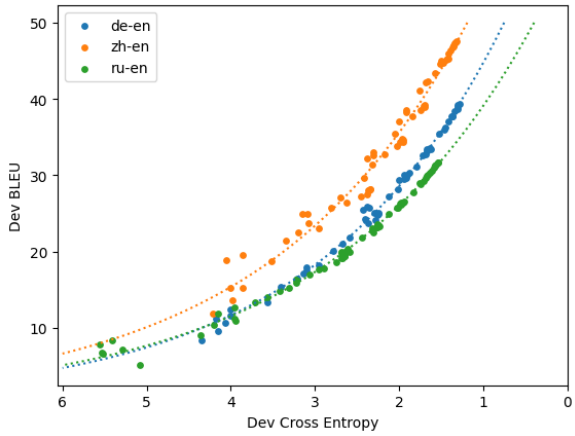

Figure 5: Same results as Figure 2 (Left), but translations are de-bpe'd, de-tokenized, and BLEU is computed using Sacrebleu (Post, 2018). This introduces some noise but does not qualitatively change the exponential relationship between cross-entropy and BLEU.

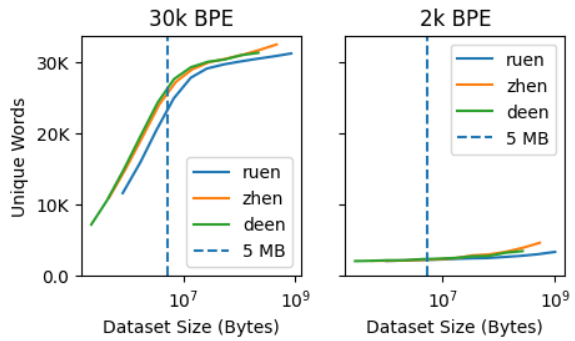

Figure 6: The number of unique words seen during training drops precipitously around 5 MB of data when using a BPE of size 30k, but remains constant when using a BPE of size 2k.

A Parameter Scaling in Japanese-English Translation

In this section, we provide a brief retrostpective analysis of the results of Zhang and Duh (2020), in which many MT systems were trained to evaluate the efficacy of hyper-parameter optimization techniques. Specifically, we examine their results on the Japanese-English WMT 2019 Robustness task (Li et al., 2019). Figure 9 shows power-law scaling of the development cross-entropy loss with the number of non-embedding parameters.[12] We see that changing the BPE encoding vocabulary size and the number of layers can affect the constant multiplier N_C, but does not seem to affect the exponent α_N. Furthermore, multiple attention

[12]In these experiments, only a single dataset size was used so we were unable to verify power-law data scaling.

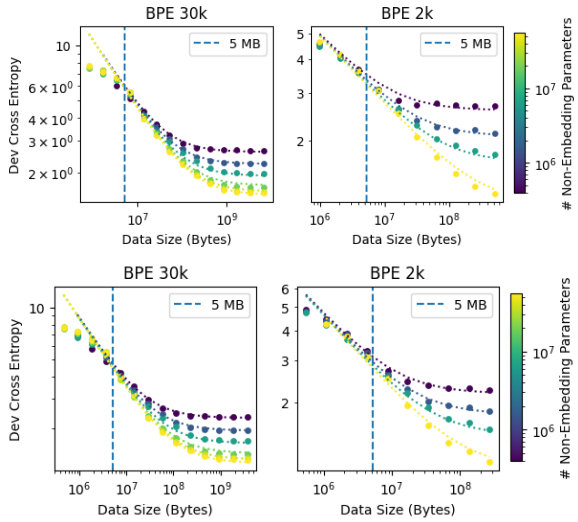

Figure 7: In both ru-en (Top) and zh-en (Bottom), models trained on <5 MB of data (around 40k lines) fall off-trend when using a BPE vocabulary of 30k. When encoding the same dataset with a BPE of 2k, the plateau is rectified and returns to power-law scaling.

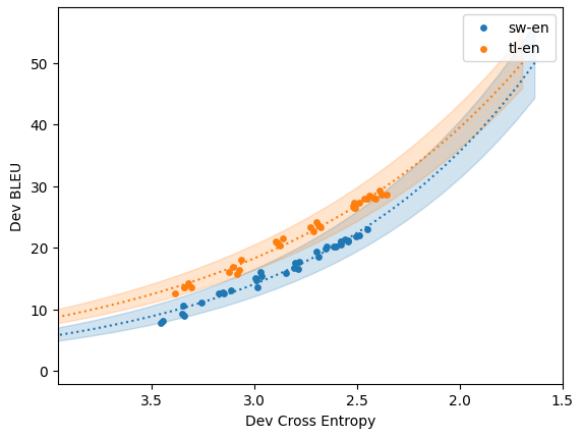

Figure 8: BLEU vs. cross-entropy development loss for the models trained in Section 3. Standard error is shown in the shaded region.

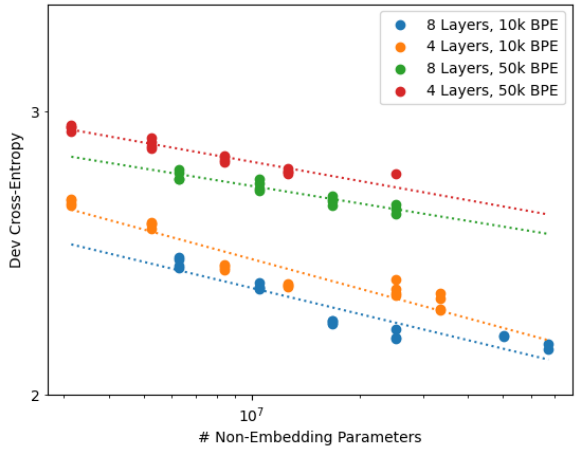

Figure 9: Non-embedding parameters vs. development cross-entropy for the Japanese-English models described in Section A. Changing the number of layers or the BPE vocab size or the number of Transformer layers seems to impact the multiplying constant N_C, but does not seem to change α_N much.

head settings (8, 16) were trained for each model size but they do not seem to impact scaling trends.

We exclude some outliers with unexpectedly large losses at for larger model sizes. This only occurs for specific learning rates, so we believe those models failed to converge due to improper learning rate tuning.