Data Cartography for Low-Resource Neural Machine Translation

Aquia Richburg
AMSC
University of Maryland
arichbu1@umd.edu

Marine Carpuat
Computer Science & UMIACS
University of Maryland
marine@umd.edu

Abstract

While collecting or generating more parallel data is necessary to improve machine translation (MT) in low-resource settings, we lack an understanding of how the limited amounts of existing data are actually used to help guide the collection of further resources. In this paper, we apply data cartography techniques (Swayamdipta et al., 2020) to characterize the contribution of training samples in two low-resource MT tasks (Swahili-English and Turkish-English) throughout the training of standard neural MT models. Our empirical study shows that, unlike in prior work for classification tasks, most samples contribute to model training in low-resource MT, albeit not uniformly throughout the training process. Furthermore, uni-dimensional characterizations of samples – e.g., based on dual cross-entropy or word frequency – do not suffice to characterize to what degree they are hard or easy to learn. Taken together, our results suggest that data augmentation strategies for low-resource MT would benefit from model-in-the-loop strategies to maximize improvements.

1 Introduction

While neural sequence-to-sequence models have led to dramatic increases in translation quality for many Machine Translation (MT) tasks, the large amounts of data needed to train high quality systems is only available for a small number of the 7000 languages spoken in the world (Haddow et al., 2022). Efforts to improve this state of affairs have focused on data cleaning, recognizing that parallel samples are often noisy and of limited domain coverage in these settings (Kreutzer et al., 2022; Caswell et al., 2020), and on data collection and augmentation, with techniques ranging from large scale crawling (Bañón et al., 2020; Schwenk et al., 2021) to automatic data augmentation via back-translation (Sennrich et al., 2016; Fadaee and Monz, 2018) to generating novel data through lexical or phrase replacement (Liu et al., 2021; Fadaee et al., 2017). However, to make the most of limited data, we argue that a better understanding of the role that data samples play in training low-resource MT models is needed.

This paper presents an empirical study of the role that training samples play in low-resource MT training so that future work on improving data quality and quantity can prioritize properties of data that matter most. Our study focuses on MT in a very low-resource setting (Swahili-English, 63k training samples) and a low resource setting (Turkish-English, 358K samples). As Joshi et al. (2020) note, the term "low resource" encompasses a wide disparity of situations. We work with languages from two distinct classes in their taxonomy of language resources. Turkish falls into the "underdogs" category, which comprises languages with large amounts of unlabeled data, comparable to those available for high-resource languages, and that are primarily challenged by lesser amount of labeled data (parallel text in the case of MT). Swahili falls into the "hopefuls" category, which refers to languages for which only small labeled datasets have been collected, often by a community which strives to keep them alive in the digital world.

In this work we use Data Maps (Swayamdipta et al., 2020), a technique that uses the model's training dynamics, to identify and characterize regions in realistic low-resource translation data and their impact on translation quality. Data Maps places training examples on a coordinate plane that describes a model's confidence and variability of these examples across the training process. We show that the notion of difficulty of an example changes at different stages of training, that samples are not uniformly useful, and that existing heuristics for sample difficulty do not characterize the samples that are hard to learn based on training dynamics. We hope that these findings can help inform data augmentation strategies in future work.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5594–5607
December 7-11, 2022

2 Methods

We describe the main methods used to analyze the role of different samples in low-resource MT training through training dynamics.

2.1 Applying Data Maps to MT

Swayamdipta et al. (2020) introduced Data Maps, a model-based tool to characterize and diagnose datasets by placing training samples on a two-dimensional map based on (1) the variability ($\hat{\sigma}_i$) of model scores for each sample throughout training on the x axis, and (2) the model confidence ($\hat{\mu}_i$) for each sample on the y axis. Their approach was developed for classification tasks such as Natural Language Inference. We use it here for sequence-to-sequence models by replacing the probability of predicting the gold class for a given input with the probability of the gold output sequence given the input sequence:

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} P_{\theta(e)}\left(y_i^{(m)}|x_i^{(n)}\right) \quad (1)$$

and

$$\hat{\sigma}_i = \frac{1}{\sqrt{E}}\sqrt{\sum_{e=1}^{E}\left(P_{\theta(e)}\left(y_i^{(m)}|x_i^{(n)}\right) - \hat{\mu}_i\right)^2} \quad (2)$$

On classification tasks, training on various subsets of the data revealed that samples in three of the map regions play a distinct role:

1. samples with high confidence and low variability (upper left corner) are **easy to learn**, as the model consistently predicts their class correctly with high confidence, and their presence in the training data encourages convergence without over fitting.

2. samples with low confidence and low variability (lower left corner) are **hard to learn**, possibly because they are noisy and mislabeled.

3. samples with high variance (right) are **ambiguous**, and are found to contributing most to improving the models ability to generalize to out-of-domain data.

We divide the MT Data Maps into these three regions by using the midpoint between the max and min values of confidence and variability scores. Figure 1 previews Data Maps for our two low-resource MT systems (Swahili-English on the left, Turkish-English on the right). We use BLEU, an automatic translation quality metric, as an analog to the accuracy of the converged model's translation output compared against the reference. We will discuss the specific settings used to generate it and the findings in Sections 3 and 4 respectively.

2.2 Distinguishing MT Training Phases

MT requires generating a complete sequence that is well formed in the target language and adequately conveys the meaning of the source. We expect that the multi-faceted nature of these predictions might not be captured in a single data map.

We build on insights from Voita et al. (2021), who showed that MT training can be split into three distinct phases, to decompose Data Maps across phases of training. Throughout training, MT models learn to model

1. the target language (**phase 1**)

2. the lexical translation between source and target words (**phase 2**),

3. the word reorderings needed to generate well-formed output sequences (**phase 3**).

We follow their approach to determine boundaries between the three training phases. Phase 1 ends when the scores of a language model on MT outputs produced by model checkpoints plateaus. Phase 2 ends when development BLEU decreases and non-monotonic alignments between input and model outputs increase.

3 Empirical Study Settings

We focus on translation from Swahili and Turkish into English, since these are two low-resource tasks involving under-studied languages that raise interesting reordering and morphology challenges for MT. Swahili is a Bantu language spoken by an estimated 200 million people. It has an agglutinative morphology and uses a subject-verb-object order. Turkish is a Turkic languages, with 70 to 80 million speakers. It uses extensive agglutination and follows a subject-object-verb order.

Data The Swahili-English data is a combination of data taken from the IARPA MATERIAL program[1], GlobalVoices and CommonCrawl. The training data totals 63k sentence pairs and we have

[1]https://www.iarpa.gov/index.php/research-programs/material

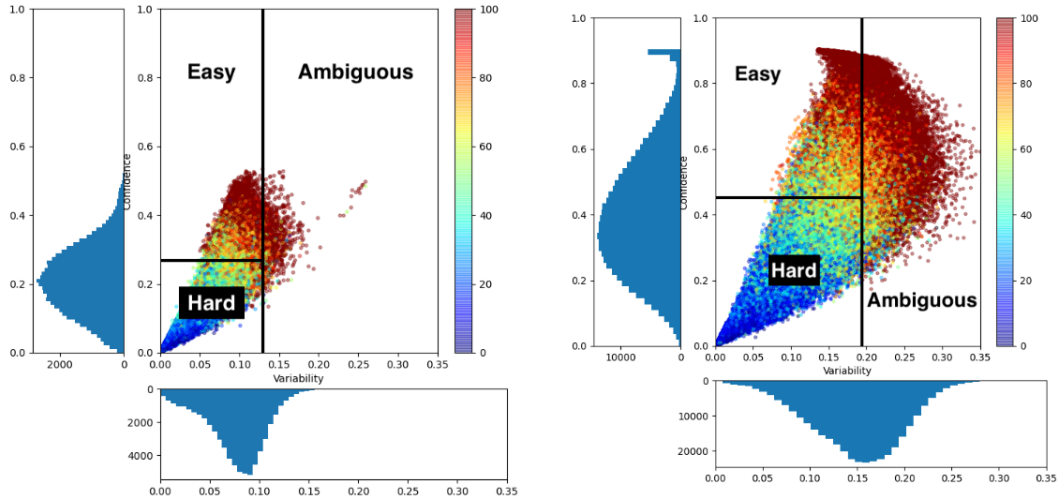

Figure 1: Data Maps generated for the Sw-En (left) and Tr-En (right) training corpora. On the y-axis we plot confidence scores and on the x-axis variability scores. The vertical and horizontal bars separate the easy (top left), ambiguous (right) and hard regions (bottom left). The color gradation represents the sentence-level BLEU for the final model's predictions, from blue (lower quality) to red (higher quality).

a validation set and two test sets from MATERIAL of sizes 2000 and 3367. The training corpus for Turkish-English is a combination of publicly available data from WIT[3] (Cettolo et al., 2012) and SETimes (Alperen et al., 2010) at around 350k sentence pairs. The validation and test sets come from the 2016 WMT news translation task (Bojar et al., 2017); we use newsdev2016 as validation and newstest2016 and newstest2017 as test sets; with sizes 3000 and 3007, respectively.

All data is tokenized and true-cased using the Moses toolkit. We use SentencePiece to generate subword tokens with a joint vocabulary and decide on the sizes of 4000 and 16000 for Sw-En and Tr-En, respectively through hyper-parameter tuning.

MT Systems We utilize the Transformer model with self-attention (Vaswani et al., 2017) implemented within the Sockeye toolkit (Hieber et al., 2020). Inspired by Araabi and Monz (2020) we alter some of the hyper-parameters for the low-resource data setting. For Sw-En, we set the number of layers for the encoder and decoder to 5, the number of heads for self-attention to 2 and label smoothing to 0.6. Both Transformer attention and activation layer dropout is 0.3 while target word dropout is 0.1. For Turkish-English, there are 6 layers for the encoder and decoder, and 8 self-attention heads. Label smoothing is 0.1, Transformer attention and activation layer dropout is 0.1. We do not use target word dropout. For both systems the number of units in the feed-forward layer is 2048, the source and target embedding dimension size is 512, the number of hidden units is 512. We maintain the same training configurations for the models trained on the entire data-set and all filtering experiments. Models are considered converged when perplexity on the validation set output has not improved in 20 checkpoints. We measure translation quality with BLEU (Papineni et al., 2002). All translation results are averaged over three random seeds.

Data Maps For generating Data Maps, we compute model predictions at each checkpoint, 1000 parameter updates, and plot all training examples. The Data Maps visualized are the result of a single training run with a single random seed, as varying the random seed leads to very similar looking plots.

Phase Boundary Detection To detect the transition between phase 1 and 2, we use a 5-gram KenLM (Heafield, 2011) language models (LM) on the tokenized target side of the training bi-text. At each training checkpoint, we score the model translation output of the development set. To detect the transition between phase 2 and phase 3, we use fastAlign (Dyer et al., 2013) to obtain automatic word alignments between the source and model translation of the development set at each checkpoint, and compute the average Kendall tau distance between the word orderings induced from the alignments.

	SW-EN		TR-EN	
	ANALYSIS1	ANALYSIS2	Newstest16	Newstest17
Baseline (100% data)	$\mathbf{32.43 \pm 0.15}$	$\mathbf{32.52 \pm 0.15}$	$\mathbf{20.61 \pm 0.09}$	$\mathbf{20.10 \pm 0.06}$
Subset 33%				
Most Ambiguous	27.59 ± 0.08	27.94 ± 0.14	16.39 ± 0.29	16.02 ± 0.27
Easiest	27.63 ± 0.24	28.12 ± 0.31	15.62 ± 0.12	15.12 ± 0.19
Random	25.01 ± 0.25	25.17 ± 0.06	15.63 ± 0.06	15.29 ± 0.12
Hardest	11.43 ± 0.58	10.82 ± 0.52	12.38 ± 0.17	12.19 ± 0.20

Table 1: Experiments on training on subsets from the different regions for Sw-En and Tr-En.

4 Overall Data Maps

Maps Overview Figure 1 shows the Data Maps for our Swahili-English model (Sw-En; on the left) and for our Turkish-English model (Tr-En; on the right). The distribution of samples across regions differs from the classification maps in prior work (Figures 1-2 in Swayamdipta et al. (2020)), reflecting the complexity of the MT task and the low-resource training regime. The spread of confidence and variability scores are smaller for MT, leading to samples being more concentrated in the lower left part of the map. MT maps have more hard samples than in classification maps, and the distinction between easy and ambiguous is not as clear cut. Unlike in classification maps, even the easiest samples do not reach upper left corner of the map. Furthermore, for the very low resource Swahili-English, training samples have lower confidence and variability than for Turkish-English, and are concentrated in the easy and hard regions. The translation quality of the final model increases with the distance from the origin in the map: decoding outputs for hard samples have the worst BLEU scores, while high confidence and high variability samples are translated with higher BLEU. This confirms that neither confidence nor variability alone can characterize sample difficulty. Note that low BLEU scores might arise from bad MT or from good MT that do not match noisy references.

Role of regions Following Swayamdipta et al. (2020), we conduct filtering experiments to characterize the role that samples in different regions play in the training process. We train MT models until convergence on a third of the original training set, selecting samples with 1) high confidence (easiest), 2) high variability (most ambiguous) and 3) low confidence (hardest). We also include a random sample of the same size as a basis for comparison. The MT configuration is kept constant throughout these experiments as described in Section 3.

As can be seen in Table 1, training on subsets of the data degrades performance substantially compared to the full data baseline. Swayamdipta et al. (2020) found that when using the most ambiguous samples, training on one third of the data maintained the baseline performance in-domain and slightly improved out-of-domain performance. Unsurprisingly, this does not hold for our low-resource MT tasks.

The relative drops based on different selection methods are more revealing. Compared to the baseline trained on the complete dataset, the lowest degradation in BLEU is obtained by training on either the easiest or ambiguous samples (tied) for Swahili-English, and by ambiguous samples for Turkish-English. Specifically, easiest samples improve translation quality over random samples by 2.6-3 BLEU points on Swahili-English, but reach the same quality for Turkish-English. The quality obtained with ambiguous samples is the same as for the easiest samples for Swahili-English and slightly better (and statistically significant) for Turkish-English. While overall these results corroborate the findings of Swayamdipta et al. (2020) that easy and ambiguous samples are the most useful for training, the two regions do not play as distinct a role as in classification settings: on the SNLI and MultiNLI tasks, ambiguous samples were found to lead to improved out-of-domain performance, with minimal degradation of in-domain performance, while easy samples hurt performance in large amounts but help convergence in small amounts. In our settings, the distinction between easy and ambiguous samples is not as clear cut in the map, which is confirmed by these filtering experiments.

Overall, these results call for further analysis to better understand map regions throughout training.

5 Data Maps Across Training Phases

Maps Overview The Data Maps for each of the three training phases (Figure 2) show that samples do not play a constant role throughout the training process. For both language pairs, samples are initially distributed across the diagonal during phase 1 – target language modeling. Samples variability decreases in later stages focused on cross-lingual learning, leading to more easy samples and fewer to none ambiguous samples. The range of confidence scores increases throughout training for Turkish-English, with a small peak in distribution in the easiest upper left right corner in the later phases, while the confidence remains under 0.5 for most samples in the very low Swahili-English setting. For both language pairs, a large fraction of samples remain in the hard region throughout training.

Sample Movement We characterize movement across regions more directly by reporting the proportion of samples that move across the different regions over the three training phases in Figure 3. The main type of movement consists of hard samples becoming easy from one phase to the next: 30% of Swahili-English samples and 25% of Turkish-English samples that were hard in phase 1 become easy in phase 2, while 23% of Swahili-English samples and 24% of Turkish-English samples that were hard in phase 2 become easy in phase 3, reflecting that model confidence increases while variability remains in the same lower range for these samples as training progresses. Easy samples in one phase remain easy in the next. Ambiguous samples also remain ambiguous, except for Turkish-English where 15% of samples move from the ambiguous to the easy regions between phases 1 and 2. Notably, by the end of training, 42% and 35% of the Swahili and Turkish data respectively are still classified as hard examples.

Overall, this analysis suggests that the role of samples changes during training, but in a way that reflects overall model improvement rather than the modeling focus of each phase: for instance we do not observe a trend where samples that are easy for language modeling in phase 1 become hard in phase 2 while learning lexical translations. We conduct further filtering experiments to assess how sample movement impacts training more directly.

Filtering Experiments We conduct experiments where we filter out 20% of the training samples in various ways (Appendix Table 4), thus controlling the amount of training data throughout. Filtering out only the samples that are hard in phase 3 leads to no loss in translation quality compared to using all the data, while filtering out samples that are hard during each phase of training leads to no difference for Swahili-English and a small but significant decrease in BLEU for Turkish-English, where significance is defined by performance outside the confidence interval from averaging scores over three random seeds. By contrast, filtering a random sample of the same size hurts BLEU significantly.

We also filter out all the samples that are hard in phase 3, which means that they remained hard throughout the three phases of training. This represents a substantial portion of the training set – 42% of the data for Swahili and 35% for Turkish – which leads to very small training sets in our low-resource settings. As expected, this hurts translation quality significantly, but not as much as removing the same amount of randomly selected data.

Overall, these results suggest that while the hard samples do not contribute as much as other samples to improving MT models in these low-resource settings, they are not just noise that should be filtered out. Samples that are hard to learn only in early phases can have a small positive impact on BLEU. This leads us to examine the properties of samples in each of the regions.

6 Characterizing Region Properties

Table 2 shows randomly selected samples in each of the Data Maps regions. As can be seen from these examples, it is not immediately obvious what makes a sample hard vs. easy. We note that sentence length does not correlate with region placement. Minor lexical changes between the reference and Google translation can place samples in either the easy or ambiguous regions.

6.1 Correlation with Data Difficulty Heuristics

Metrics We compare the notion of sample difficulty from Data Maps against heuristic data characterizations from the literature. At each phase and for the overall Data Maps, we divide the data by regions and compute Pearson's correlation coefficient between the distance of a training example from the origin (using the variability and confidence values as a coordinate pair) and various measures of difficulty. There is no clear way to

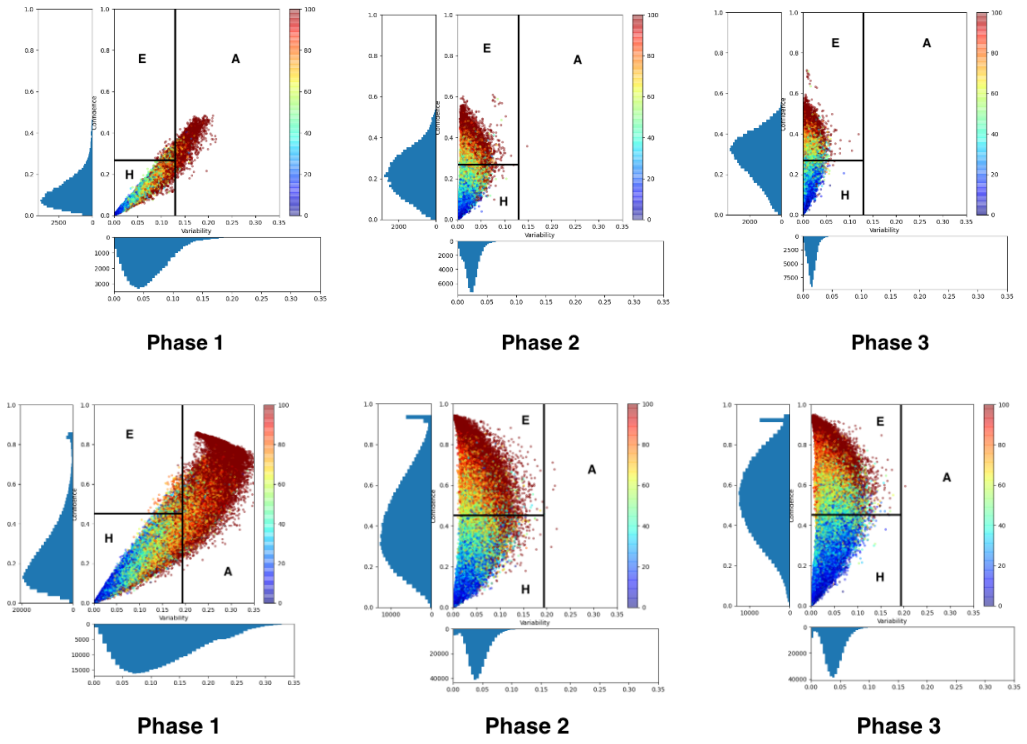

Figure 2: Data Maps for the Sw-En (top) and Tr-En (bottom) data broken down according to phases 1 to 3 (from left to right). Region boundaries and BLEU color gradations are defined as in Figure 1.

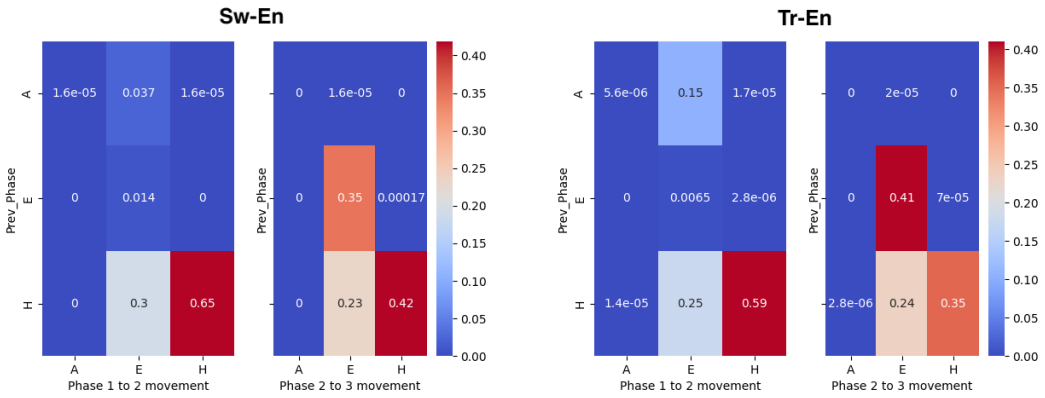

Figure 3: Heat maps tracking the movement of samples between regions across phases for Sw-En and Tr-En. The vertical axis represents the location from the previous phase and the horizontal axis the location in the current phase.

Region	Example source, reference English translation, Google translation of source
Easy	SW SRC: Mtandao wa haki za Binadamu illitoa ripoti yake ya 2013 kuhusu 'hali mbaya' ya haki za binadamu nchini Vietnam: GT SRC: *The human rights network released its 2013 report on the 'dire situation' of human rights in Vietnam:* EN REF: The Vietnam Human Rights Network released its 2013 report about the 'worsening' human rights situation in Vietnam:
Ambiguous	SW SRC: Sata: umetokea mji gani? GT SRC: *Sata: What city are you from?* EN REF: Sata: which town are you from?
Hard	SW SRC: Miongoni mwa wasichana hawa, 41 walifariki kufuatia tukio hili la mauaji ya kijinsia na 15 wakiachwa na majeraha mabaya sana. GT SRC: *Of these girls, 41 died as a result of the genocide and 15 were left with serious injuries.* EN REF: Of these girls, 41 died as a result of this femicide and 15 are badly hurt.
Easy	TR SRC: Hareket etmenin beynimizin en önemli fonksiyonu olduğuna inanıyorum – kimsenin size bunun yanlış olduğunu söylemesine izin vermeyin. GT SRC: *I believe movement is the most important function of our brain – don't let anyone tell you that it's wrong.* EN REF: I believe movement is the most important function of the brain – don't let anyone tell you that it's not true.
Ambiguous	TR SRC: Bugün, iki oğlum David ve Daniel, annem ve babamla konuşabiliyor, onları tanıyabiliyorlar. GT SRC: *Today, my two sons, David and Daniel, are able to talk to and get to know my parents.* EN REF: Today, my two sons David and Daniel can talk to my parents and get to know them.
Hard	TR SRC: Bu süreç 20-30 yıl civarı sürer. GT SRC: *This process takes around 20-30 years.* EN REF: The process takes two to three decades.

Table 2: Sentence samples from the three Data Map regions for Swahili and Turkish.

Classifier	SW-EN	TR-EN
Ambiguous, Easy or Hard?	0.65	0.64
Hard or not?	0.78	0.82

Table 3: Accuracy of classifiers for placing examples in different map regions.

define what makes a MT training sample difficult, however hypotheses about properties of easy vs. difficult samples emerge from prior work on data cleaning and curriculum learning. We include dual cross entropy (Junczys-Dowmunt, 2018) as a measurement of data noise, and sentence length, token frequency and word norm embeddings as sentence difficulty criteria from the curriculum learning literature (Zhang et al., 2019; Platanios et al., 2019; Liu et al., 2020). Sentence length and token frequency are computed at the word level; although we saw similar trends at the subword level.

Findings We find that most metrics considered have a weak correlation with the distance from the origin according to Data Maps (Appendix Figures 4, 5). Only dual cross-entropy achieves a moderate correlation, primarily in the hard regions, which is not surprising given that it is based on model scores just like Data Maps. We use a rule of thumb for cutoffs at 0.3 and 0.5 for weak and moderate correlation, as outlined by Hinkle et al. (2003). This might explain why no clear data difficulty metrics has emerged from the curriculum learning literature, where data ordering strategies that improve translation quality do not always align with meaningful intuitions about sample difficulty (Zhang et al., 2019), and raise the question of whether the difficulty captured by Data Maps is purely related to model uncertainty rather than intrinsic data uncertainty.

6.2 Map Regions Can Be Automatically Predicted

While uni-dimensional heuristics do not help explain sample positions in Data Maps, we ask whether data properties suffice to automatically predict in which region a sample lies.

We train supervised classifiers that predict the region of a sample by fine-tuning an XLM-RoBERTa model, and evaluate on held-out data. The training data consists of a balanced set drawn from the easiest, most ambiguous and hardest Data Maps samples. On the binary task that consists of distinguishing hard samples from others, the classifier achieves a high accuracy, in comparison to a random baseline. On the three-way classification task, we find that the classifier confuses easy and ambiguous samples, confirming that the boundary between these two regions is not as clear.

Overall these results indicate that data intrinsic patterns captured by the classifier partially explain map regions, even though these patterns are not as interpretable as the uni-dimensional characterizations from Section 6.1. However, the classifier does not entirely separate data uncertainty from model uncertainty since the underlying multilingual language model shares basic properties with the Transformer model used for translation.

7 Related Work

It is well established that data quality can have a large impact on the performance of MT systems. Khayrallah and Koehn (2018) shows how different types of training set noise impact the quality of neural MT models. Dual cross-entropy (Junczys-Dowmunt, 2018) and other techniques are routinely used to clean crawled datasets that are known to contain many types of noise (Caswell et al., 2020).

Training sample difficulty is often considered in the context of curriculum learning, which aims to present training examples to the model in an order that benefits model performance or convergence speed. Some of the earlier works by Bengio et al. (2009); Cirik et al. (2016) show that ordering training samples by sentence length can decrease training speed or improve model performance. Zhang et al. (2018); Platanios et al. (2019) go on to develop curriculum learning strategies building on the use of linguistic features such as sentence length or word frequency; but their results are mixed. Later works have looked at more model-focused techniques for deciding the difficulty of training examples such as the norm of source words (Liu et al., 2020), language model scores on monolingual text (Zhou et al., 2020) or the change in the decrease of MT model loss (Xu et al., 2020). These techniques have shown more consistent improvements in translation quality but still do not query information from the MT model itself; except in the case of the latter which still only looks at local changes in the model loss. In contrast Data Maps are based on a holistic view of the training process.

Prior work on low-resource settings has primarily focused on modeling strategies or supplement-

ing the training data. Araabi and Monz (2020) determined appropriate hyper-parameters for Transformer models that were initially developed in high resource settings. Sánchez-Cartagena et al. (2021) paired data augmentation with a multi-task modeling approach to strengthen the power of the model's encoder and decoder. Others considered restricted sampling from MT decoding (Li et al., 2020), leveraging a high resource pivot language (Xia et al., 2019), generating new contexts for infrequent words (Fadaee et al., 2017) and learning to edit noisy training samples (Briakou et al., 2021).

8 Conclusion

This work used training dynamics in the form of Data Maps to understand the role of samples in two low-resource machine translation tasks (Swahili-English and Turkish-English). The Data Maps show that the role of samples changes across the training of neural MT models. Filtering experiments show that all samples contribute to training and that hard samples do not hurt translation quality. Further analysis shows that the role of samples cannot be explained by simple uni-dimensional heuristics, although classifiers can be trained to detect which region a sample belongs to with a potentially useful level of accuracy. This suggests future work on using Data Maps to guide data augmentation, possibly using automatic classifiers to target the easy or ambiguous samples that are more likely to benefit further training.

Limitations

Languages Our empirical study is limited to two language pairs, and only considers translation into English. While Swahili and Turkish are both morphologically rich languages that are under-studied in MT and NLP research, and fall in different categories in the taxononmy of Joshi et al. (2020), this work alone does not indicate how our findings generalize to other language pairs, translation directions, and data conditions.

Models We also experiment with a single Transformer model architecture. However this is the dominant architecture in MT research, which makes the findings more broadly applicable.

Evaluation Finally, our evaluation of MT quality is done entirely automatically by comparing system outputs to reference translations using BLEU, which is an imperfect yet useful measure of quality.

References

Murat Serdar Alperen, Oleg Kapanadze, A. Ceausu, Carlos Ramisch, and A. Fotopoulou. 2010. Southeast european times : A parallel corpus of balkan languages , francis tyers and.

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation.

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Y. Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. volume 60, page 6.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Eleftheria Briakou, Sida I. Wang, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Bitextedit: Automatic bitext editing for improved low-resource machine translation. *CoRR*, abs/2111.06787.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

M. Cettolo, C. Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *EAMT*.

Volkan Cirik, Eduard H. Hovy, and Louis-Philippe Morency. 2016. Visualizing and understanding curriculum learning for long short-term memory networks. *ArXiv*, abs/1611.06204.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, Lisboa, Portugal. European Association for Machine Translation.

Dennis E Hinkle, William Wiersma, and Stephen G Jurs. 2003. *Applied statistics for the behavioral sciences*, volume 663. Houghton Mifflin College Division.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data Rejuvenation: Exploiting Inactive Training Examples for Neural Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2255–2266, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*.

Yu Li, Xiao Li, Yating Yang, and Rui Dong. 2020. A diverse data augmentation strategy for low-resource neural machine translation. *Information*, 11(5).

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. Mitchell. 2019. Competence-based curriculum learning for neural machine translation. *CoRR*, abs/1903.09848.

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, Belgium, Brussels. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8502–8516, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *EMNLP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1126–1140. Association for Computational Linguistics.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1282–1292.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *WMT*.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Chen Xu, Bojie Hu, Yufan Jiang, Kai Feng, Zeyang Wang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2020. Dynamic curriculum learning for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3977–3989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Jeremy Gwinnup, Marianna J. Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *CoRR*, abs/1811.00739.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915, Minneapolis, Minnesota. Association for Computational Linguistics.

Lei Zhou, Liang Ding, Kevin Duh, Shinji Watanabe, Ryohei Sasano, and Koichi Takeda. 2021. Self-guided curriculum learning for neural machine translation. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 206–214, Bangkok, Thailand (online). Association for Computational Linguistics.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6934–6944, Online. Association for Computational Linguistics.

A Appendix

	SW-EN		TR-EN	
	ANALYSIS1	ANALYSIS2	Newstest16	Newstest17
Baseline (100% data)	32.43 ± 0.15	32.52 ± 0.15	20.61 ± 0.09	20.10 ± 0.06
Filter 20%				
All/All/Phase 3	$\mathbf{32.65 \pm 0.06}$	$\mathbf{32.58 \pm 0.24}$	$\mathbf{20.68 \pm 0.10}$	$\mathbf{20.20 \pm 0.23}$
All/Phase 2/Phase 3	32.52 ± 0.24	32.49 ± 0.24	20.45 ± 0.04	20.20 ± 0.14
Phase 1/Phase 2/Phase 3	32.32 ± 0.16	32.30 ± 0.09	20.06 ± 0.08	19.74 ± 0.06
Filter 42% (SW), 35% (TR)				
All/All/Random	29.38 ± 0.19	29.46 ± 0.15	19.10 ± 0.19	18.68 ± 0.02
All/All/Phase 3	31.33 ± 0.10	31.37 ± 0.17	19.55 ± 0.07	18.87 ± 0.24

Table 4: Filtering experiments for Sw-En and Tr-En. All indicates 100% of the data is used and Phase k indicates filtering a percentage of the hardest samples as scored by Phase k.

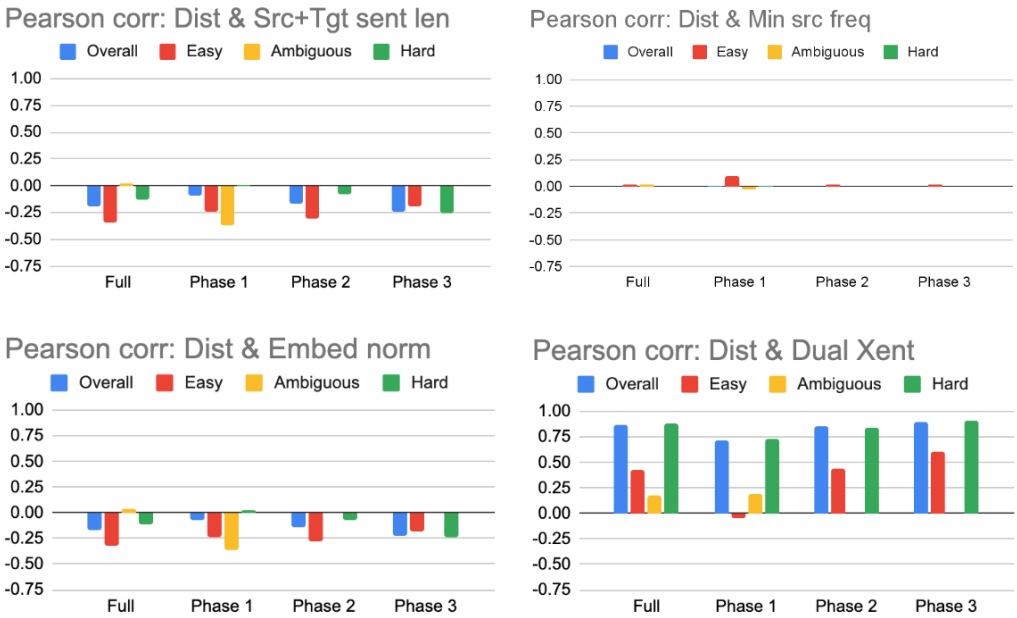

Figure 4: Pearson correlations of rankings according to distance from the origin on the Data Maps and other metrics for Sw-En. We include correlations for the overall data (in blue) and a breakdown by sub-region. We also show the change in correlations when looking at the Data Map as a whole vs. according to each phase.

Figure 5: Pearson correlations of rankings according to distance from the origin on the Data Maps and other metrics for Tr-En. We include correlations for the overall data (in blue) and a breakdown by sub-region. We also show the change in correlations when looking at the Data Map as a whole vs. according to each phase.