Robust Domain Adaptation for Low-Resource Biomedical Translation

Stan Fris
University of Amsterdam
s.c.j.fris@uva.nl

Yanxu Chen
University of Amsterdam
yanxu.chen@student.uva.nl

Abstract

This study explores domain adaptation for low-resource English to Russian biomedical translation. We compare fine-tuning a pre-trained NMT model with backtranslation (BT) of monolingual biomedical texts, finding that combining BT and fine-tuning yields the best results on a biomedical test set, while BT alone excels on related-domain data. We introduce temperature-based sampling for BT, which outperforms standard BT and noise-based augmentation. Additionally, we evaluate data selection methods: 5-gram ranking, BM25, and a gradient-based "LESS" algorithm. We find that BM25 and LESS offer modest gains over random sampling, particularly in noisy settings. Our work highlights the efficacy of temperature-based BT and robust data selection for domain-specific, low-resource NMT. Code is made available through google drive, as this project is not fully public at this moment https://drive.google.com/file/d/1WOlrH-9AbiWNAbvl2fchrl64bxptDDGz/view?usp=sharing.

1 Introduction

In Natural Language Processing, the vast majority of research is focused on 20 out of the 7000 languages that exist (Magueresse et al., 2020), leaving most of the languages understudied. The understudied languages are referred to as "low-resource languages", for which fewer resources like human-translated sentences exist. Because of this, a core challenge of training models for Neural Machine Translation (NMT) in low-resource languages, is to effectively train models without large, high-quality datasets. What also necessitates these techniques is the phenomenon that models are often needed for a single specific application, such as assistance in translating medical documents, in which the available data is also limited. For this type of task, a technique called domain adaptation is used, where a model is fine-tuned to perform well on a different but related domain (Chu and Wang, 2018). However, this does still require a significant amount of relevant data for the training objective. Therefore, there is a need for identifying ways to increase the amount or quality of existing datasets.

Backtranslation (BT) (Sennrich et al., 2016), which generates synthetic parallel data by translating monolingual target-language text into the source language, has been shown to be an effective augmentation technique for low-resource NMT without modifying the core training objective (Fadaee and Monz, 2018; Caswell et al., 2019). This is useful, as parallel data is scarce, but monolingual data exists significantly more for many languages. Research by Edunov et al. (2018) has shown that adding noise to backtranslation when using beam search can significantly improve performance. The way noise is added in this paper is by adding random permutations and deletions to sentences, increasing variance. However, other methods, such as increasing temperature, are currently not investigated.

A technique to increase the effectiveness of pre-existing parallel data is data selection. Here, selection methods are used for finding relevant data for the target domain in general datasets, or for ordering data to improve training effectiveness (Koehn et al., 2018; Zhang et al., 2019b). Research by Zhang et al. (2019b) has shown that using data selection to augment training, by training extensively on a subset of data, rather than the full dataset, can improve model performance. However, this paper only evaluates 5-gram data selection, which is a method with known limitations (inability to handle synonyms and long-range dependencies). In the field of instruction-tuning, data selection using Gradient Similarity Search with the LESS algorithm has shown effective results (Xia et al., 2024).

This study examines the effectiveness of BT and data selection for domain adaptation in a low-

resource setting, biomedical note translation from English to Russian. Specifically, the following topics are investigated: 1. To what extent does BT complement or replace fine-tuning for domain adaptation? 2. In what ways can BT be modified to increase the variance of sentences, and what is the effect on domain-adaptation? 3. How do varying ranking methods affect curriculum learning for NMT? 4. Can using the LESS algorithm for data selection improve model performance in Domain Adaptation for Neural Machine Translation?

We investigate these research questions through several extensions of existing methods. Investigation of fine-tuning and backtranslation finds strong improvements, where extending backtranslation with noise shows further improved scores. Extensive testing is done of data selection methods, including 5-gram, BM25, and an adaptation of the LESS algorithm. Models are evaluated across related-domain and general-domain datasets, including partially corrupted and combined datasets. Here, it is found that using selection methods can, in certain situations, give improved results over randomly selected subsets, and therefore allow for more efficient and effective training of NMT models.

2 Related Work

A common approach in NMT for domain adaptation is fine-tuning (Luong et al., 2015): a model is first trained on large general data for general tasks, and then fine-tuned to adapt to the task. Research by Kumari et al. (2021) shows that backtranslation can also effectively be used for domain adaptation. This motivates the question of whether backtranslation can improve upon fine-tuning. Investigating combinations offers insight into how these can complement one another, as the methods do not necessarily rely on the same data (one-sided data and parallel data, respectively) and could therefore be combined.

Backtranslation (BT) has shown effective improvement of learning in NMT models for low-resource languages (Fadaee and Monz, 2018; Edunov et al., 2018; Caswell et al., 2019). The method for generating synthetic sentences is a core modelling choice in backtranslation. Methods such as greedy search and beam search attempt to maximize the MAP with respect to the output sentence (Edunov et al., 2018). However, research by Edunov et al. (2018) has shown that adding noise to backtranslation when using beam search can significantly improve performance. Similarly, Lample et al. (2017) shows that by jointly optimizing denoising auto-encoding, cross-domain reconstruction, and adversarial alignment of latent representations between source and target languages, improved performance can be reached. The way noise is added in Edunov et al. (2018) is by adding random permutations and deletions to sentences, which is theorized to add to a richer training signal and an increased difficulty to predict target translations. Increasing the temperature is another way to raise generation variance in language models. Investigating this gap in existing research could offer insight into the effect of increasing variance without altering the input signal, and increase insight into the effects of backtranslation.

For domain-adaptation as well as backtranslation, it can be difficult to find a sufficient amount of data that specifically matches the task, leading to worse performance in NMT (Koehn and Knowles, 2017). For fine-tuning, a positive effect has been shown when adding information from generally related domains, such as adding information from general medical data when fine-tuning for a biomedical task (Koehn et al., 2018). Research by Zhang et al. (2019b) has shown that adapting curriculum learning for domain adaptation can lead to improved results in NMT. In their adaptation, similarity scores from 5-gram language models are used to rearrange the order of training samples to improve learning order and increase the relative frequency of examples that are more relevant to the task. The paper makes use of a 5-gram language model and does not test other ranking methods. In this work, we address this gap and investigate whether advanced models for data ranking and selection could lead to improved performance.

A novel method for data selection is through estimation of gradient influence (Pruthi et al., 2020). Here, the gradient of datapoints is compared to the gradient of test datapoints in order to gain an improved intuition of which datapoints are relevant for training. This approach has shown promising results in several related fields, such as instruction tuning and in-context learning (Xia et al., 2024; Han et al., 2023). Research by Xia et al. (2024) proposes the LESS algorithm, which is designed to select relevant data for instruction tuning, and shows promising results for data selection. The algorithm builds a gradient datastore from various datasets and selects data with similar low-dimensional gra-

dient features to validation examples. This allows the algorithm to effectively find relevant data for a target task. This technique has proven very effective in instruction tuning. Models tuned with LESS-selected data consisting of only 5% of the full data can outperform training on the full dataset (Xia et al., 2024). The previous success of data selection and ranking methods in NMT, as well as the very good performance of LESS on the closely related task of instruction tuning, suggests that adapting LESS for NMT could lead to improved results.

3 Methodology

3.1 Neural Machine Translation

The core NMT loop is formulated as a sequence-to-sequence (seq2seq) generation task based on the Transformer architecture (Vaswani et al., 2017).

3.2 Fine-tuning

Fine-tuning is a commonly used technique to adapt a pre-trained NMT model to specific domains or datasets. It involves continued training on some natural or synthetic data to improve the model's performance on some specific domains or downstream tasks. Specifically for domain adaptation, fine-tuning is necessary to increase performance on the downstream task.

3.3 Backtranslation

Backtranslation is extended by implementing the possible improvements discussed in Section 2. BT experiments are extended by attempting to increase the amount of variance in backtranslated sentences. A method to add noise in BT, as proposed by Edunov et al. (2018) works by altering input text by deleting words, replacing words with a filler token, and swapping words. We propose increasing the temperature to investigate the effect of increasing the variance of sampling without altering input tokens. Furthermore, an experiment is performed in which we vary the level of beam search in combination with sampling.

For backtranslation and fine-tuning, we expect to see improved results, especially when combining the two methods. With respect to general domain performance, it is possible that Fine-tuning might lose performance compared to the baseline, as it optimizes the model for task-specific performance, whereas backtranslation is performed on more general domain data, and could therefore improve these results.

3.4 Data Selection

Experiments in data selection will first be performed by applying the 5-gram methods for data selection and ranking in curriculum learning as discussed in Zhang et al. (2019b). Furthermore, this method will be extended by investigating different methods for relevance ranking. This will be done by evaluating performance when using BM25 as a ranking method.

We expect to see that ranking methods can provide an improvement with respect to random selection, as samples more similar to the target set are expected to allow for more effective learning. Specifically, we expect this to scale with the effectiveness of the ranking method, where BM25 is expected to perform better than 5-gram selection.

3.5 Advanced Data Selection Methods

The pipeline of LESS (Xia et al., 2024) is adapted to be used for NMT data selection. The task for the LESS algorithm is defined as follows: from the previously defined dataset $\mathcal{D}$, select $\mathcal{D}_{\text{train}}$ such that training on $\mathcal{D}_{\text{train}}$ achieves the lowest loss on validation data. This is done by computing the influence of datapoints using the SGD or Adam (Kingma and Ba, 2014) optimizer, as defined in Xia et al. (2024).

Our data selection is simplified from the original LESS, as in this case, only a single subtask exists, and we only use one checkpoint. We compute gradient features for model checkpoint $\boldsymbol{\theta}$:

$$\overline{\nabla}\ell(\mathcal{D}_{\text{val}};\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_{\text{val}}|}\sum_{\boldsymbol{z}'\in\mathcal{D}_{\text{val}}}\tilde{\nabla}\ell(\boldsymbol{z}';\boldsymbol{\theta})$$

The influence Inf of each datapoint z on the task when using the SGD or Adam optimizer with respect to the validation dataset can then be written as:

$$\text{Inf}(\boldsymbol{z};\mathcal{D}_{\text{val}}) = \frac{\left\langle \bar{\nabla}\ell\big(\mathcal{D}_{\text{val}};\,\boldsymbol{\theta}\big),\ \tilde{\Gamma}(\boldsymbol{z},\boldsymbol{\theta})\right\rangle}{\left\|\bar{\nabla}\ell\big(\mathcal{D}_{\text{val}};\,\boldsymbol{\theta}\big)\right\|\ \left\|\tilde{\Gamma}(\boldsymbol{z},\boldsymbol{\theta})\right\|}$$

where $\Gamma(\boldsymbol{z},\boldsymbol{\theta}) = \frac{\boldsymbol{m}}{\sqrt{\boldsymbol{v}+\epsilon}}$ is the normalized update for the Adam optimizer, and $\Gamma(\boldsymbol{z},\boldsymbol{\theta}) = \nabla\ell(\boldsymbol{z},\boldsymbol{\theta})$ is the gradient for the SGD optimizer. $\tilde{\nabla}\ell$ and $\tilde{\Gamma}$ represent $\nabla\ell$ and Γ randomly projected into a lower dimension respectively.

This then gives us an influence score for each datapoint, which can be sorted to get a list of data points ordered by expected relevance. The dataset for training the NMT model can be obtained by taking a small split from the top of this dataset.

We expect that the advanced selection using LESS will allow for even better performance using selected samples, as using the gradient similarity can allow for even more relevant sampling compared to BM25 and 5-gram methods.

4 Experiments

4.1 Tasks and Datasets

The chosen task is translating English to Russian in the biomedical field. The performance in completing this task will be represented by performance on the WMT Biomedical Translation Task (Neves et al., 2022).

As training datasets, several related domain datasets, such as TICO-19 (Anastasopoulos et al., 2020) (3.1k samples) and Medline (Bawden et al., 2020) (8.5k samples). Furthermore, datasets from more general domains have been used, including a Wikimedia-based dataset (300k samples), and SciPar (10k samples) (Tiedemann and Thottingal, 2020; Roussis et al., 2022).

Evaluation will be done on the official WMT test set as described in Neves et al. (2022). In addition, the FLORES dataset has been used to evaluate retained general performance (NLLB Team, 2022).

4.2 Models

The baseline model is the FairSeqMachineTranslation (FSMT) English–Russian translation model developed by Facebook AI as part of their submission to the WMT 2019 News Translation Task (Ng et al., 2019).

Backtranslation is performed using the corresponding FSMT Russian–English model, also developed by Facebook AI for the same task.

4.3 Metrics

BLEU (Papineni et al., 2002) measures the n-gram precision between the output and the references, and uses a brevity penalty to penalize candidate strings that are too short. It is a simple and widely used metric in machine translation, but struggles with synonyms and rephrasing due to its simplicity (Wieting et al., 2019).

COMET (Rei et al., 2022) is a learned evaluation metric that correlates strongly with human judgments. It uses pretrained multilingual encoders and is fine-tuned on human-rated translation data to align with human preferences.

BERTScore (Zhang et al., 2019a) uses contextual embeddings from pretrained language models like BERT to evaluate the similarity between the output and references.

4.4 Fine-tuning

We fine-tune the model on the Medline parallel dataset, with an AdamW optimizer, with a batch size of 32, a learning rate of 2×10^{-5}. More details in the hyperparameters can be found in the appendix A.1.

4.5 Backtranslation

We backtranslate the Medline monolingual dataset with the FSMT Russian-English translation model. Unless specified, the temperature is set to 0 during backtranslation, and the maximum generation length is 512 tokens.

For noise augmentation methods, the data augmentation will be performed by altering inputs using the same method as discussed in (Edunov et al., 2018). Augmentation is applied to the full Medline monolingual dataset.

4.6 Data Selection

Data Selection methods require choosing a training dataset as well as a reference dataset. As reference datasets, TICO-19 and Medline are used. Base selection experiments are performed by selecting a 20% subset of full training sets, and evaluating whether the selected subset improves performance. For wikimedia, a smaller subset of 1% is used, resulting in 2000 training samples for Scipar, and $\pm$4000 samples for wikimedia.

4.7 Further Investigation in Data Selection

To test the selection of relevant source datasets in a controlled setting, custom datasets are also constructed, which allows for evaluating whether selection methods effectively identify relevant data and filter out incorrect data. For this, 2 different datasets are constructed through concatenation. To evaluate the ability to select domain-specific data from mixed datasets, TICO_SP is constructed as a combination of the TICO-19 and SciPar datasets, resulting in a 10,000/2,000 split of TICO/SciPar Data. To evaluate the ability to find relevant or high-quality data, a combination of SciPar and a portion of the Wikimedia, Wiki_SP is constructed, where 90,000 samples were randomly selected from the Wikimedia dataset to create a dataset with 100,000 samples along with SciPar samples.

To evaluate the ability of selection methods to filter out incorrect data in a setting where data

is partially corrupted, a custom corrupted dataset 'Corrupted TICO_SP' is created using the same TICO/SciPar subset, where all SciPar samples are corrupted by randomly replacing 50% of letters. This results in a dataset with $10,000$ corrupted data and $2,000$ useful data.

LESS is adapted for use with all datasets. The implementation is taken directly from Xia et al. (2024), and used with default hyperparameters. Two versions of LESS can be used: a version where stochastic gradient descent is used, and one where the Adam optimizer is adapted for use. Due to resource constraints, the Adam variant is not used in all experiments.

5 Results

5.1 Baseline, Fine-tuning and Backtranslation

In Table 1, evaluation metrics for using backtranslation and fine-tuning on the FSMT model are shown. It can be observed that fine-tuning on parallel data along with backtranslated data gives the best performance across all metrics. Backtranslated data gives a better BLEU score than parallel data, while parallel data gives better neural-based scores than backtranslated data. This may indicate that fine-tuning on backtranslated data gives a better word-level precision, while fine-tuning on natural parallel data gives a better generation quality. Table 2 shows results when evaluating on the TICO Dataset, where backtranslation shows the most improvement over all scores, and fine-tuning shows decreased performance when compared to the baseline. This suggests that fine-tuning might hurt performance for different tasks, and that backtranslation can transfer effectiveness to some extent.

	BLEU	COMET	BERTScore
Baseline	27.75	0.8234	0.8121
FT	31.18	0.8722	0.8928
BT	33.14	0.8473	0.8292
FT+BT	**33.68**	**0.8800**	**0.8970**

Table 1: Metrics on WMT Biomedical dataset comparing performance of Baseline, fine-tuning (FT), backtranslation (BT), and a combination of BT and FT (FT+BT), overall, FT+BT shows the best scores across all metrics.

5.2 Extension of Backtranslation

In Table 3, the results are shown for augmenting backtranslation to improve variance. Experiments have been conducted using the original method of Edunov et al. (2018), as well as the proposed increase in temperature. A hybrid approach has also been tested.

	BLEU	COMET	BERTScore
Baseline	28.67	0.8387	0.8615
FT	26.90	0.8362	0.8578
BT	**29.98**	**0.8531**	**0.8685**
FT+BT	29.44	0.8526	0.8671

Table 2: Metrics on TICO-19 comparing the usage of fine-tuning, backtranslation, or both methods.

Looking at the base augmentation methods, where words are replaced or altered, we observe a decrease in performance across all metrics. A possible explanation is that Edunov et al. (2018) used a significantly larger dataset, meaning that augmentation had a relatively smaller effect on the overall vocabulary and word frequency distribution. This method was also evaluated with a lower augmentation rate, which yielded slightly better performance than the original method, though still lower than standard backtranslation across all datasets.

From the experiments evaluating the effect of increased temperature, it can be observed that a non-greedy decoding approach leads to improved performance. A temperature of 1.0 improves scores across all metrics compared to regular backtranslation, while a temperature of 2.0 improves COMET and BERTScore but results in lower BLEU scores. One possible explanation is that BLEU relies on n-gram precision and is less tolerant of synonyms. Since higher temperature increases variance, this may introduce more paraphrasing or synonym usage, reducing BLEU scores despite potential improvements in semantic similarity. COMET, being less sensitive to surface-form variation, reflects these semantic gains more effectively. A higher temperature of 5.0 leads to a substantial degradation in performance across all metrics, indicating a clear trade-off between variance and backtranslation quality.

The hybrid approach, which combines preprocessing-based augmentation and increased temperature, shows improved results over standard backtranslation. However, it does not outperform the use of increased temperature alone, possibly due to the limited effectiveness of augmentation.

	BLEU	COMET	BERTScore
Baseline	27.753	0.8234	0.8121
BT	33.141	0.8473	0.8292
BT_aug_0.05	32.913	0.8442	0.8284
BT_aug_0.1	32.993	0.8357	0.8268
BT_tmp_1.0	**33.836**	0.8490	0.8310
BT_tmp_1.5	33.438	0.8470	0.8316
BT_tmp_2.0	32.165	**0.8643**	**0.8536**
BT_tmp_3.0	23.815	0.8406	0.8501
BT_tmp_5.0	12.125	0.7843	0.8109
BT_tmp_2.0_aug_0.05	31.536	0.8539	**0.8536**

Table 3: Comparison of usage of Fine-tuning, Backtranslation or Both methods, using only Backtranslation gives the best results when looking at BLEU score.

5.3 Data Selection

In order to evaluate the effect of using data selection methods, as described in 4.6, several experiments have been conducted to evaluate the performance of selection methods on custom datasets. First, the base effects of using data selection for improving performance on a general domain dataset $\mathcal{D}$ when compared to a baseline of random selection. The results of these experiments can be observed in Table 4. Looking at data selection when using TICO as the reference dataset $\mathcal{D}_{\text{ref}}$, there is an increase in performance for LESS and BM25 across all metrics, suggesting an improved quality of data after selection. For 5-gram, performance is decreased over all metrics across both datasets compared to random selection. When Medline is used as reference dataset, both BM25, and both versions of LESS show no improvement over random selection, suggesting that Medline is a less effective reference dataset to use for reference than TICO for improvement in the WMT task; the only notable point is that the Adam version of LESS is much more effective in terms of BLEU score for the Medline reference set, improving even over the results of the selection using TICO. As BLEU is a direct measure of translation similarity, this could indicate that LESS (Adam) is an effective method for selecting similar samples in this scenario.

The results of evaluating the performance of different selection methods on the general domain wikimedia dataset are shown in Table 5. Only 5-gram shows improved performance across all metrics when compared to random selection, while both BM25 and LESS do not improve the performance consistently. When compared to Table 4, it can be concluded that this training data was less effective in improving WMT performance, given that final performance is lower, while more data was used for training after selection (2000 samples for Scipar, compared to ± 4000 for wikimedia).

$\mathcal{D}$	$\mathcal{D}_{\text{ref}}$	Method	BLEU	COMET	BERTScore
	Baseline		27.753	0.8234	0.8121
SciPar	TICO	Random	28.932	0.8240	0.8140
SciPar	TICO	5-gram	27.083	0.8208	0.8127
SciPar	TICO	BM25	**29.276**	**0.8244**	**0.8152**
SciPar	TICO	LESS (SGD)	29.270	0.8238	0.8147
SciPar	TICO	LESS (Adam)	29.233	0.8228	0.8143
SciPar	Medline	Random	28.075	**0.8284**	**0.8156**
SciPar	Medline	5-gram	27.949	0.8275	0.8154
SciPar	Medline	BM25	28.389	0.8260	0.8145
SciPar	Medline	LESS (SGD)	28.750	0.8237	0.8137
SciPar	Medline	LESS (Adam)	**29.475**	0.8222	0.8140

Table 4: Effect of using different selection methods and reference datasets for the SciPar datasets on WMT Performance, results show slightly improved performance when using BM25 and LESS with TICO as reference, and slightly worse performance when using Medline as Reference.

We observe that even with random selection, performance is minimally improved over baseline performance, specifically with random selection from the Scipar Dataset, which matches the conclusions made by Koehn et al. (2018).

$\mathcal{D}$	$\mathcal{D}_{\text{ref}}$	Method	BLEU	COMET	BERTScore
	Baseline		27.753	0.8234	0.8121
Wikimedia	TICO	Random	27.903	0.8250	0.8238
Wikimedia	TICO	5-gram	**28.262**	**0.8322**	**0.8277**
Wikimedia	TICO	BM25	27.911	0.8263	0.8151
Wikimedia	TICO	LESS (SGD)	27.477	0.8296	0.8207

Table 5: Effect of using different selection methods for the medline dataset on WMT Performance, results show improved performance when using the 5-gram method.

5.4 Further Investigation in Data Selection

In this section, several components of selection are evaluated. Initially, the effectiveness of models in selecting data from mixed datasets for different models. For this, the importance of the size of the reference dataset and the amount of target data retrieved by selection methods is evaluated when varying the total selection threshold. Furthermore, we investigate the effectiveness of models in retrieving relevant data from a partially corrupted dataset.

5.4.1 Effect of Reference Dataset Size

In Figure 1, the effects of varying the amount of reference data given to selection methods on the amount of target dataset data can be observed. Results show that across selection models, there is no observable increase in performance as the sam-

pling percentage increases over 20% (200 samples). Evaluating on other datasets and with different save percentages also showed no observable difference above 200 samples. Especially, 5-gram methods can be observed to receive very little effect from the sampling percentage of the reference set. The effect when using Medline as the reference dataset is shown and explained in A.2, which is similar.

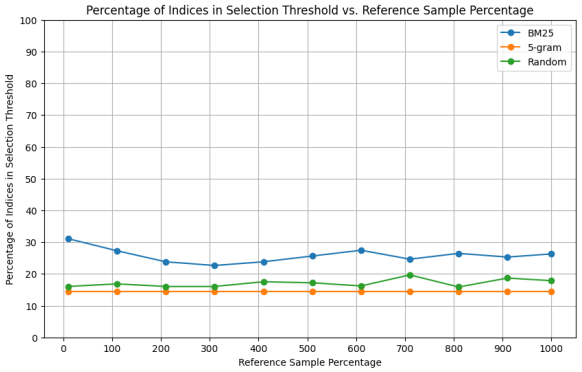

Figure 1: The effect of varying the amount of relevant information supplied to retrieval methods in retrieving correct samples, evaluated on the TICO-Scipar mixed dataset with a save percentage of 5%. Results show that after around 200 samples, there is no observable further increase in performance across retrieval models.

5.4.2 Effect of Target Dataset Size

Furthermore, Figure 2 shows the effect of varying the percentage of samples retrieved from the target dataset, and what the effect is on the percentage of samples taken from the target group when looking at the TICO-Scipar mixed dataset, where the TICO test set is used as reference. Results show a strong selection bias for the TICO dataset in up to 5% of 5-gram selected data, and 20% for BM25. Above this sample size, samples from the other dataset are selected more. Above these percentages, there is a strong decrease in sampling from this dataset. The observation of a decrease in both methods suggests that the SciPar dataset contains samples that are considered more relevant than a portion of the samples from TICO. It should be noted that this decrease is much less sharp for BM25 than it is for 5-gram. LESS consistently selects an above-average number of samples from the related dataset.

Overall, it should be noted that a high or low selection amount for any dataset can not allow us to conclude a model is not selecting well, as although generally, we expect samples from inside the dataset to be more similar to other samples, than to samples from another dataset, but this is not always true. In Appendix A.2, the selection curves are shown, where it is found that selection for 5-gram methods is quite erratic, and does not show consistent selection.

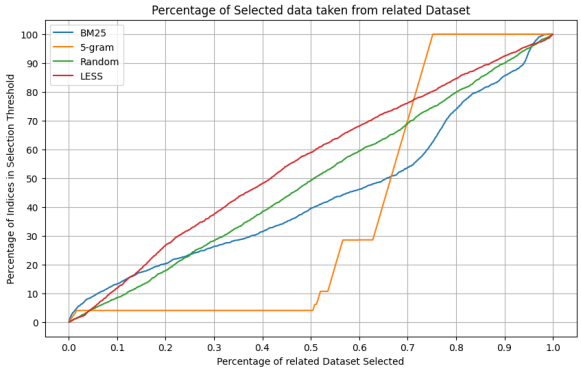

Figure 2: The effect of increasing the number of samples taken from the total available mixed dataset on the amount of data selected data taken from the related dataset

Given the similar or worse performance of 5-gram methods, they are omitted from further experiments.

5.4.3 Performance on Mixed Dataset

In Table 6, the results are shown for evaluating models across combined datasets. The results do not show a clear advantage for selection methods over randomly selecting samples. In both dataset tests, models show different improvements, and we see that non-gradient selection methods perform better in terms of COMET and BERTScore for the TICO_SP dataset, but perform worse in this aspect on the Wiki_SP dataset. LESS (SGD) shows an advantage over random selection in both datasets; due to resource constraints, no experiment was conducted for LESS (Adam) for the Wiki_SP dataset. However, the results for TICO_SP indicate no significant difference in performance of the models, except a leaning towards the BLEU score for the Adam version. This leaning towards BLEU, combined with the previous result where LESS (Adam) outperformed other methods in BLEU, could indicate that LESS (Adam) is specifically helpful in terms of the BLEU score. We also observe that all models perform over the baseline scores again. A possible explanation for these results could be that randomly overlapping between datasets is also an effective method for finetuning, or that, due to the limited sizes of datasets, it is difficult to observe general trends.

The results of investigating selection methods in

$\mathcal{D}$	Method	BLEU	COMET	BERTScore
Baseline		27.75	0.823	0.812
TICO_SP	Random	27.73	0.828	0.813
TICO_SP	BM25	27.70	0.827	**0.816**
TICO_SP	LESS (SGD)	27.69	**0.831**	0.815
TICO_SP	LESS (Adam)	**27.90**	0.821	0.813
Wiki_SP	Random	27.21	0.829	0.816
Wiki_SP	BM25	**28.45**	0.825	0.814
Wiki_SP	LESS (SGD)	27.52	**0.832**	**0.827**

Table 6: Results for evaluating selection methods when selecting from combined datasets of Tico and Scipar (15% selected), and Wikimedia and Scipar on the WMT dataset (5% selected). Results show nearly identical performance for all evaluated models.

a more one-sided setting, where part of the data is corrupted and does not benefit training, are shown in Table 7. Here, we see that although random selection loses all performance, BM25 and LESS show very similar results to the baseline, meaning that performance is retained and improved in some metrics. This indicates that selection methods could be an effective approach for partially corrupted data or highly different data.

$\mathcal{D}$	Method	BLEU	COMET	BERTScore
Baseline		27.753	0.8234	**0.8121**
Corrupted Tico_SP	Random	0.0000	0.3219	0.0000
Corrupted Tico_SP	BM25	26.833	**0.8278**	0.8111
Corrupted Tico_SP	LESS (SGD)	**27.8390**	0.8269	0.8113

Table 7: Comparison of results when using data selection on the partially corrupted dataset

6 Discussion

Several aspects of domain adaptation for low-resource languages in NMT have been investigated.

6.1 Backtranslation

In examining fine-tuning and BT, it was found that combining methods offers improved performance on target tasks. However, fine-tuning has the drawback of significantly decreasing general-domain performance, confirming our hypothesis. Regarding BT, methods proposed by Edunov et al. (2018) did not transfer effectively to the target task. A novel approach of increasing sampling temperature to boost variance was introduced, leading to improvements across target metrics.

6.2 Data Selection

Data selection methods were also explored. When selecting a subset of data, the 5-gram base method performs erratically, probably due to the possibility of 5-gram overlapping being low or even zero on the experimental dataset. Methods such as BM25 and LESS were shown to slightly outperform random selection in some cases. However, these improvements were relatively small and inconsistent, varying across target and reference datasets. This suggests that no single selection method universally outperforms others in all settings, which contradicts our hypothesis that the general performance of selection methods is most important in the effectiveness of training on selected data.

Further investigation of selection methods on mixed datasets showed that the amount of reference data had a limited effect on selection quality. Even with a small reference set, similar samples were chosen as if using a larger set. It was also found that selection methods do not consistently prioritize the dataset most similar to the reference. When evaluating performance on mixed datasets, selection methods did not significantly outperform random selection, indicating limited utility for identifying relevant data. While LESS demonstrated effectiveness in related tasks such as instruction tuning, no significant improvement was found in these experiments over random selection. However, in scenarios with corrupted data, selection methods maintained performance across all metrics, whereas random selection suffered a substantial decline, suggesting robustness in such settings.

7 Conclusion

In conclusion, several aspects of fine-tuning and backtranslation for domain adaptation have been investigated, including several novel approaches listed below:

- Increasing backtranslation variance using the temperature parameter, which shows improved results over existing methods and baseline performance.
- Investigation of several aspects of data selection, including both in-domain and general datasets, and analysis of important factors in selection quality.
- Adaptation of the LESS algorithm for data selection, allowing for gradient-based data selection.

Together, these contributions provide a clear analysis for methods for improving low-resource NMT, identifying several directions for future work on more adaptive and flexible translation systems.

8 Distribution of work

8.1 Experiments

Stan implemented the experimental framework consisting of the Neural Machine Translation (NMT), Fine-tuning, Backtranslation, Data Selection and the generation of mixed and corrupted dataset. Yanxu contributed to the development and debugging of backtranslation and developed the code for the LESS method. Stan conducted the experiments for fine-tuning, extended backtranslation methods and data selection (5-gram and BM25). Stan and Yanxu jointly conducted the experiments for basic backtranslation and LESS.

8.2 Writing

Stan carried out the framework of the entire report and the initial draft, engaging in the writing of all sections. Yanxu contributed to parts of the methodology, experiments, discussion, limitations, and wrote the appendix, and did overall revisions during the review phase.

9 Limitations

Future work could explore additional use cases or modifications to improve performance under current settings. In all experiments, reference and test datasets were kept strictly separate to prevent data leakage. However, due to the limited overlap between TICO-19/Medline and the WMT test set, the quality of reference data for selection may have been constrained. Using a reference dataset more closely aligned with the WMT test set could improve results, especially for methods like LESS, which depend heavily on gradient similarity.

Moreover, in the mixed dataset experiments, it remains unclear whether datasets like Scipar or TICO are more beneficial for training. While TICO is closely related to the target domain and would intuitively enhance performance, further experiments could quantify each dataset's contribution to overall performance. This would offer clearer insight into the effectiveness of selection methods.

Finally, data selection was only applied during training. However, Zhang et al. (2019b) shows that selection methods can also be used during evaluation by reusing selected data, effectively increasing the amount of relevant training data seen by the model. Evaluating models with this approach could provide additional insights into the performance of selection methods. LESS, in particular, may benefit due to its gradient-based relevance mechanism.

All experiments in this work focused on the English-Russian translation task in the biomedical domain using WMT data. Further experiments could assess whether these findings generalize to other domains, datasets, and language pairs, potentially revealing broader applicability of the investigated methods.

All experiments are only conducted once for each case due to computational budget and time constraints. Further experiments of running cases with interesting findings multiple times with different seeds and disturbance could help us show the robustness of the methods and further determine the statistical significance.

According to the plot in A.2, keeping 20% of data may not be the most beneficial threshold to represent the difference between selection methods like BM25 and random selection. It could be beneficial if we test multiple saving thresholds to see if similar results hold as the chosen threshold.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *CoRR*, abs/1906.06442.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *CoRR*, abs/1808.09381.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. *CoRR*, abs/1808.09006.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. *Preprint*, arXiv:2306.15091.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Kevin Duh, and Brian Thompson. 2018. The JHU machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 438–444, Belgium, Brussels. Association for Computational Linguistics.

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, Vancouver. Association for Computational Linguistics.

Surabhi Kumari, Nikhil Jaiswal, Mayur Patidar, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, and Lovekesh Vig. 2021. Domain adaptation for NMT via filtered iterative back-translation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 263–271, Kyiv, Ukraine. Association for Computational Linguistics.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-Lopez, Eulalia Farre-Maduel, Martin Krallinger, Cristian Grozea, and Aurelie Neveol. 2022. Findings of the wmt 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation*, pages 694–723, Abu Dhabi. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. 2020. Estimating training data influence by tracking gradient descent. *CoRR*, abs/2002.08484.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. 2022. SciPar: A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2652–2657, Marseille, France. European Language Resources Association.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. pages 86–96.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: selecting influential data for targeted instruction tuning.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019b. Curriculum learning for domain adaptation in neural machine translation. pages 1903–1915.

A Appendix

A.1 Hyperparameters and compute

Most of the hyperparameters for training are adapted from the setup used by LESS. The detailed numbers used are listed in Table 8. Due to resource constraints and the large number of distinct tasks, no hyperparameter tuning was run. Experiments were carried out on a single NVIDIA A100 40GB GPU along with 9 Intel Xeon Platinum 8360Y CPUs.

Hyperparameter	Value
Epoch	5
Learning rate	2e-5
Batch size	32
Weight decay	0
β_1 for Adam	0.9
β_2 for Adam	0.999
Warmup steps	100

Table 8: Default hyperparameters for training. Unless specified, training was conducted with these hyperparameters.

A.2 Proportions of selected data

Figure 3 shows the percentage of data from the relevant sample group (SciPar) in the selected dataset under varying sizes of the reference dataset (a random subset of Medline), when selecting from the Wiki_SP dataset, similar to section 5.4. The improvement from increasing the size of the reference dataset is minimal. However, the 5-gram method performs better than random selection here, possibly due to the 5-gram distribution of the Wikimedia dataset being much different than SciPar.

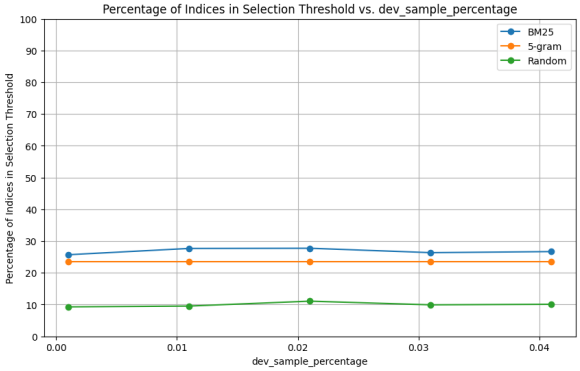

Figure 3: Percentage of samples from the relevant sample group (TICO) in the selected dataset, under varying relative size of reference dataset, when selecting from the Tico_SP dataset.

Figure 4 shows the percentage of data from the relevant sample group in the selected dataset under varying saving thresholds. With the random method, the percentage is consistently between 15% to 20%, reflecting the ratio of the number of samples in the relevant sample group and the entire sample pool. The BM25 method steadily selects more data than the random selection method when the saving threshold is below 20%.

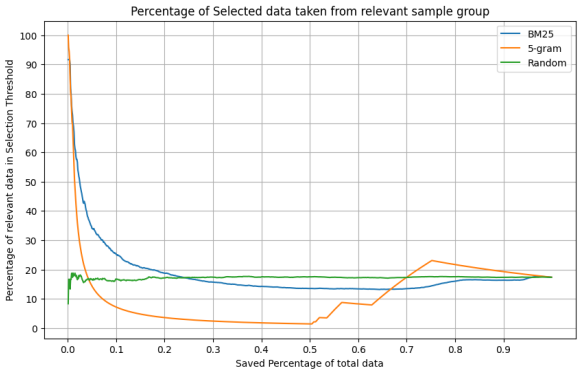

Figure 4: Percentage of samples from the relevant sample group (TICO) in the selected dataset, under varying saving threshold, when selecting from the Tico_SP dataset.

Figure 5 shows the percentage of data from the useful sample group in the selected dataset under varying saving thresholds. he random method also shows a consistent percentage between 15% and 20%, while BM25 and 5-gram based methods can rank useful samples highly to select them early. LESS performs better than random selection, but cannot consistently select useful samples at the beginning, probably due to the stochastic nature of its projection step and the normalization of gradients.

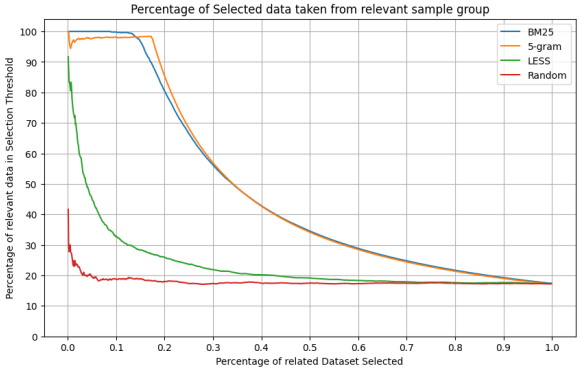

Figure 5: Percentage of samples from the useful sample group in the selected dataset, under varying saving threshold, when selecting from the corrupted Tico_SP dataset.

In figure 6, we can observe a similar trend to 5.4. The 5-gram method still performs erratically, while the BM25 method consistently selects more data than random selection, which is better than the case of the Tico_SP dataset, where BM25 only performs better in the first 20% data.

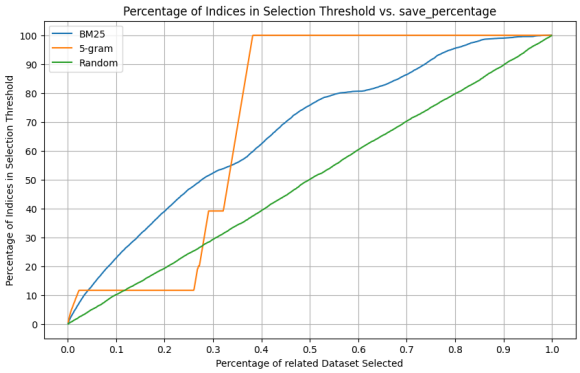

Figure 6: Percentage of selected samples from the relevant group over the size of the relevant group, under varying saving threshold, when selecting from the Wiki_SP dataset; this was not run for LESS due to resource constraints.