Cost-Effective Training in Low-Resource Neural Machine Translation

Anonymous ACL submission

Abstract

While Active Learning (AL) techniques are explored in Neural Machine Translation (NMT), only a few works focus on tackling low annotation budgets where a limited number of sentences can get translated. Such situations are especially challenging and can occur for endangered languages with few human annotators or having cost constraints to label large amounts of data. Although AL is shown to be helpful with large budgets, it is not enough to build high-quality translation systems in these low-resource conditions. In this work, we propose a cost-effective training procedure to increase the performance of NMT models utilizing a small number of annotated sentences and dictionary entries. Our method leverages monolingual data with self-supervised objectives and a small-scale, inexpensive dictionary for additional supervision to initialize the NMT model before applying AL. We show that improving the model using a combination of these knowledge sources is essential to exploit AL strategies and increase gains in low-resource conditions. We also present a novel AL strategy inspired by domain adaptation for NMT and show that it is effective for low budgets. We propose a new hybrid data-driven approach, which samples sentences that are diverse from the labelled data and also most similar to unlabelled data. Finally, we show that initializing the NMT model and further using our AL strategy can achieve gains of up to 13 BLEU compared to conventional AL methods.

1 Introduction

There are several thousand languages in today's world, with millions of people knowing only their native language. This creates a language barrier and is a hindrance to communication in this globalized world. Translation technologies are essential to overcome the language barries and enable communication between monolingual speakers. Neural Machine Translation (NMT) systems (Bahdanau et al., 2015; Vaswani et al., 2017) have significantly advanced translation quality to alleviate this problem. Supervised NMT models rely on vast amounts of parallel sentences to translate between languages with high quality. But, the labelled data is not available for many language pairs.

Unsupervised NMT (UNMT) (Lample et al., 2018; Artetxe et al., 2018) and UNMT with multilingual transfer (Fraser, 2020; Garcia et al., 2021; Li et al., 2020) are promising research directions to tackle this problem. The former learns to translate, relying on monolingual corpora but fails in practical conditions when dealing with distant low-resource language pairs (Kim et al., 2020; Marchisio et al., 2020). The latter approach uses parallel data between similar high-resource language pairs and generates decent quality. However, it is not enough to produce high-quality translations for several language pairs in both directions (source $\leftrightarrow$ target). Labelled data between the language pair in focus is necessary to attain SOTA performance.

However, human annotation of sentences poses several challenges: 1) Costly and time-taking; 2) Bilingual translators for several language pairs are hard to find. Hence, annotating large amounts of parallel sentences for low-resource languages is impractical and expensive. We need to design a training procedure that is cost-effective but also enables the model to translate with adequate quality.

One way to save costs is by employing Active Learning (AL) strategies with NMT (Zeng et al., 2019; Ambati, 2012; Haffari et al., 2009). The goal of AL is to maximise translation quality for an annotation budget of labelling B sentences. We label only the most informative B sentences in the whole unlabelled dataset using *selection strategies*. Previous works on AL (Zeng et al., 2019; Peris and Casacuberta, 2018) consider annotation budgets between hundred thousand to million sentences. But, it is not always possible to afford the annotation such amounts of data for low-resource languages. Also, current AL frameworks do not utilize the monolingual data which does not require any labelling. Analysis on AL for low-annotation budgets[1] with exploiting monolingual data is necessary to build good quality NMT systems in realistic scenarios.

Another way to improve the model without spending significant money is by integrating small, inexpensive bilingual dictionaries. Word translations are compact, can cover different domains and are a cheaper knowledge source to annotate. Exploiting this additional information with monolingual data and combining it with AL can further improve the performance of the model. However, our methods should be robust and be able to utilize smaller dictionaries.

In this work, we address the challenges above by the following contributions:

- We show that improving the model's quality by pretraining is necessary before applying AL strategies with low annotation budgets. (Table 4)
- We present a novel "*Cross-entropy difference*" selection strategy for AL that is effective in low-resource scenarios. (§ 3.3)
- We propose a inexpensive pretraining procedure by incorporating a small dictionary (1146 entries) and show that combining this with AL can increase the translation quality up to 13 BLEU. (Table 4)

2 Background: Active Learning in NMT

There are several language pairs for which parallel data is hardly available. To build NMT systems for these languages, we need to create bi-texts by annotating the unlabelled data. Given an annotation budget, we can only afford to label a certain amount of sentences in unlabelled data. However, choosing data points randomly might include annotating uninformative data and incur a waste of resources.

AL is an effective solution to reduce the amount of labelling. It uses selection strategies[2] (ψ) to mitigate this problem. $\psi(\cdot)$ is simply a scoring function to estimate the "importance" of each sentence of the unlabelled data. Choosing the top-scoring sentences can help in maximising the translation quality for an annotation budget. It can use any of the following as input: 1) Labelled data[3] ($\mathcal{L}$) 2) Unlabelled source data ($\mathcal{U}_S$) 3) Batch size ($\mathcal{B}$) 4) Model ($\mathcal{M}$) trained on the available data.

One paradigm is to use the model $\mathcal{M}$ to score each sentence in the unlabelled data. They are grouped as *model-driven* strategies. The key idea is to determine sentences in $\mathcal{U}_S$ for which the model is relatively weaker. Round-trip-translation-likelihood (RTTL) (Zeng et al., 2019; Haffari et al., 2009) is the current SOTA approach for model-driven strategies. It gives higher score to sentences for which, the model is unsure during back-translation. We generate a intermediate translation $\hat{t}$ for a sentence s. Then, we take the average of the log-probability at token level giving $\hat{t}$ as input and asking to reconstruct s at the output. Higher value indicates that the model is more confident and hence s obtains a lower score.

Another paradigm is to compare each sentence s in unlabelled data to the labelled data $\mathcal{L}$ or the whole unlabelled source data $\mathcal{U}_S$ itself. These methods can be called as *data-driven* strategies. They rely on the following heuristics:

- **Diversity**: Sampling sentences that are diverse from the existing labelled data $\mathcal{L}$ is important.
- **Density**: The test set follows the same distribution as the unlabelled data. Hence, sampling from dense regions of unlabelled data $\mathcal{U}_S$ is beneficial.
- **Hybrid**: Accounting to both of the above metrics with a trade-off.

N-gram overlap (Eck et al., 2005) is simple yet an effective data-driven strategy. It only accounts for the diversity metric. Sentences in the unlabelled data $\mathcal{U}_S$ are given a higher score, if they have more number of n-grams that are not present in the labelled data $\mathcal{L}$.

3 Cost-Effective Training in NMT

We design a sequence of training steps to exploit additional inexpensive data sources with AL to increase translation quality. The overview of the process is illustrated in Figure 1. We utilize the dictionary and monolingual data by training a UNMT

[1]We consider budgets that can annotate between 0 to $50k$ sentence pairs as low-annotation budgets

[2]We follow the terminology in Zeng et al. (2019)

[3]We generate the initial labelled data by annotating random batch of sentences.

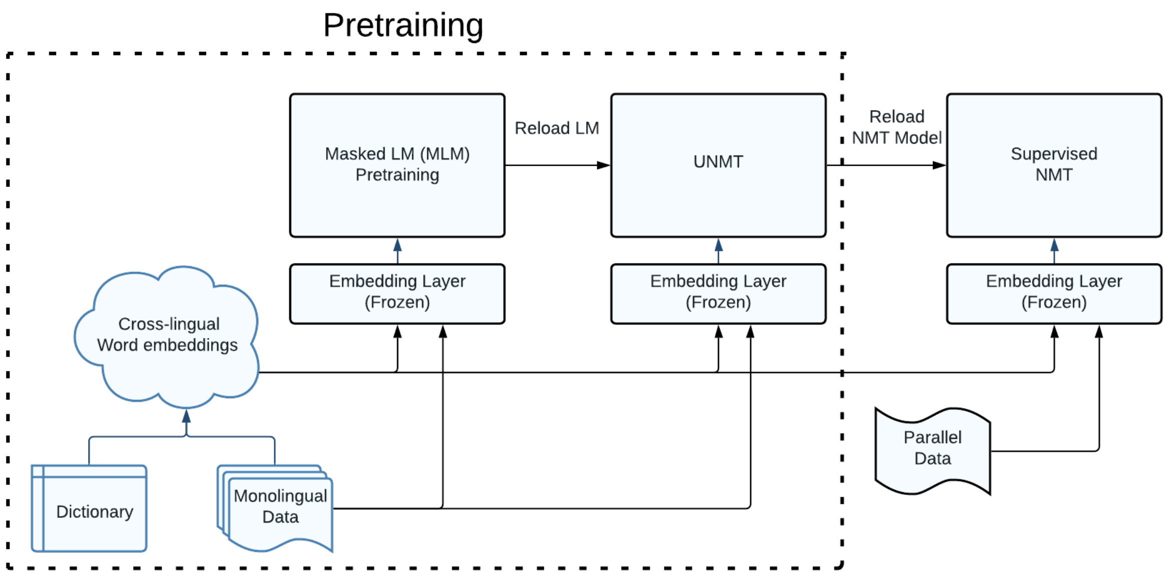

Figure 1: Proposed pretraining procedure to integrate CLWE. The black dotted box denotes the pretraining stage that only uses monolingual data. upervised NMT denotes the fine-tuning phase where we use the parallel data

system to improve the model. Then, we apply AL to sample informative data and maximise gains.

As a first step, we use the bilingual dictionary to provide supplementary supervision signal by constructing cross-lingual word embeddings (CLWE) (§ 3.1). We extract embeddings from monolingual data (Bojanowski et al., 2017) and map them into common space using a small dictionary (Artetxe et al., 2017). We hypothesize this is useful for supervised NMT in low-resource conditions.

For the second step, we use the monolingual data with CLWE to provide a strong initialization for the NMT model (§ 3.2). We leverage Masked Language modelling (MLM) (Devlin et al., 2019) and UNMT (Lample et al., 2018; Artetxe et al., 2018) objectives (self-supervised) on monolingual data to provide a better initialization for the NMT model without the need of annotation. While training on these objectives, we reload the embedding layer with CLWE created in the first step and freeze them for the entire process to always provide cross-lingual signal (Banerjee et al., 2021).

The last step is to employ AL for labelling and prioritize the annotation of the most informative sentences. We present a novel AL strategy "Cross-entropy difference" that is effective in these low-resource conditions (§ 3.3). We reload the model trained using self-supervised objectives above as initialization before fine-tuning on the sampled parallel data using AL to achieve higher performance.

3.1 Integrating Dictionaries

Incorporating word-to-word translations can increase the potential of NMT models to be handle a wider range of words, especially in low-resource conditions. We propose to take advantage of a small dictionary by learning CLWE and utilizing them for low-resource NMT. These embeddings can help in building generalised and cross-lingual NMT models which might be particularly useful in our setup. We can learn the mapping between the monolingual embeddings using a dictionary to create CLWE. Then, we can integrate them with the embedding layer of our NMT model. The only constraint is that the dictionary should contain single token-token entries. But, the current NMT models operate on sub-words using Byte-pair encoding (BPE) (Sennrich et al., 2016b). This is a problem when learning CLWE from the dictionary. Entries consisting of translating rare words would split into multiple tokens. Discarding these (particularly informative) entries would lead to losing information about the mapping between infrequent words.

We can include the infrequent words by simply operating on the word level data. However, this leads to losing all the advantages of operating with sub-words. Chronopoulou et al. (2021) has shown that CLWE is beneficial for UNMT even on sub-word level data. Therefore, we propose a modification to standard BPE in order to retain advantages operating on both word and sub-word

Dictionary Words	[**tomorrow**, **training**, center]
Source sentence	Academic Skills center will focus on **training tomorrow**
BPE	Academic S@@ kills center will focus on train@@ ing tom@@ orrow
DP-BPE	Academic S@@ kills center will focus on **training tomorrow**
	Academic S@@ kills center will focus on train@@ ing tom@@ orrow

Table 1: Example for DP-BPE. Words highlighted in bold indicate rare words present in dictionary that get split into multiple words after BPE. The two new encoded sentences after applying DP-BPE are added to the training data.

tokens. We explain this technique below and denote it as *Dictionary-preserving BPE* (DP-BPE)

First, we create a list of all the words that are present in the dictionary. We consider all the words in the list that will get split into multiple tokens as rare words. Next, we apply standard BPE for sentences that do not contain these rare words. For the remaining text that consists of rare words, we perform the following operations on each sentence:

1. Apply BPE on tokens that are not rare words. So, the rare words remain as single tokens.

2. Apply BPE on all the tokens including the rare words. In this case, these words get split into multiple tokens

3. Add the above two BPE processed sentences to the existing data.

We illustrate this process with an example in Table 1. The word "tomorrow" and "training" are rare words available in the dictionary which would split into different sub-words. We create two different sentences with selectively applying BPE. We ignore the rare words while applying BPE to form the first sentence. We create another sentence by applying BPE with including the rare words. Finally, we join these two sentences to our dataset.

There is no alignment between the texts for monolingual data. However, parallel data is aligned between the source and target sentences. The rare words might occur only in the source or only in target or in both sentences. Here, we simply apply standard BPE and DP-BPE at a time and create two new sentence pairs.

Training on the new dataset will result in both the rare word and corresponding sub-words to have similar representation. The word/sub-words will appear in the same context and eventually be treated similarly by the model. Therefore, applying DP-BPE allows us to integrate CLWE with retaining advantages from sub-word based NMT models.

After pre-processing the monolingual and parallel data using DP-BPE, we can start creating CLWE. First, we create the sub-word monolingual embeddings for both languages using a *fasttext* (Bojanowski et al., 2017) on the monolingual data.

Next, we align the monolingual embeddings using all the words in the dictionary to build CLWE. As we want to minimize the costs, we only assume having a small dictionary. Hence, we use a *semi-supervised* learning algorithm that is robust to small dictionaries and map the embeddings in a common space using *VecMap* (Artetxe et al., 2017). Therefore, we are able to build CLWE without spending large amounts on collecting dictionaries.

3.2 Exploiting Monolingual Data

Pretraining in low-resource conditions has been shown to improve the models quality significantly (Conneau and Lample, 2019; Liu et al., 2020). Therefore, we propose to use the monolingual data to improve the models performance in these challenging conditions. Moreover, having a better model increases its ability to exploit both model and data-driven AL strategies. It is easier for the model to learn from the data selected through various heuristics. Especially, the model-driven strategies need the model to be good enough to accurately identify and learn from data points where it is weak.

We extend the process in Chronopoulou et al. (2021) by integrating dictionaries and use that as a initialization before fine-tuning with AL. We begin by training the encoder using the Masked Language Model (MLM) (Devlin et al., 2019) objective on monolingual data for both languages. We build this cross-lingual language model to promote cross-lingual contextual representations. Then, we use this language model for initializing the encoder and decoder and train a UNMT (Lample et al., 2018; Artetxe et al., 2018) system. The UNMT training consists of Denoising auto-encoding (Vincent et al., 2008) and on-the-fly back translation (Sennrich et al., 2016a). Although this system often struggles to translate between distant languages adequately (Kim et al., 2020; Koneru et al., 2021), it provides a good initialization for the cross-attention and the decoder for fine-tuning.

After training the model as described above, we can start the AL process to select samples. Then, we can fine-tune the model developed using monolingual data on the chosen data points.

3.3 Effective Sampling for fine-tuning

Model-driven strategies depend on the model to estimate where it is weak. However, in low-resource conditions, the model is not strong enough to accurately select the data points where it is weak. Relying totally on diversity will lead to a challenging and small dataset, making it hard for the model to learn. Depending on density alone will lead to a small subset of similar sentences with uninformative samples causing unnecessary costs. We need hybrid approaches that account for both density and diversity to increase gain in low or very low-resource conditions.

Inspired from the strategy to select in-domain data by Moore and Lewis (2010), we present a new hybrid data-driven AL strategy called "Cross-entropy difference". The key idea is to use cross-entropy loss of causal language models (CLM) trained on the labeled and unlabeled data to estimate both diversity and density metrics.

Consider a CLM trained on the unlabelled source data. If a sentence would obtain a smaller cross-entropy loss, it indicates that this sentence is similar to the data distribution of the unlabelled source data. This allows us to measure the density metric and help in selecting sentences that are highly representative. Similarly, higher cross-entropy loss on a language model trained on the labelled source data indicates that the sentence is quite diverse. We use these heuristics and explain how we measure the density and diversity.

Let the labelled source data be denoted as $\mathcal{L}_S$. We train a CLM[4] on $\mathcal{L}_S$ and denote it as $\mathcal{M}_{LS}$. Further, we denote the cross-entropy loss of a sentence s on $\mathcal{M}_{LS}$ as $H(\mathcal{M}_{LS}, s)$. We can simply use $H(\mathcal{M}_{LS}, s)$ to measure diversity. If the cross-entropy loss is high, than the sentence would score greater in the diversity metric.

[4]Note that while training a CLM, we initialize with the MLM trained on the monolingual data for better contextualized representations.

Recall that the selection strategy scores each sentence in unlabelled source data to estimate its importance. To measure the density metric, we cannot train a language model and evaluate cross-entropy loss on sentences that the model has seen during training. This causes over-fitting and does not provide accurate scores. Therefore, we propose to split the unlabelled source data into two halves and train two separate language models. Then, the first half of the data can be scored using the model trained on the other half and vice-versa.

Let the unlabelled source data be denoted as $\mathcal{U}_S$. Due to reasons mentioned above, we split this into two halves $\mathcal{U}_{S1}$ and $\mathcal{U}_{S2}$. We denote the CLM trained on $\mathcal{U}_{S1}$ and $\mathcal{U}_{S2}$ as $\mathcal{M}_{US1}$ and $\mathcal{M}_{US2}$. Now for a sentence s present in $\mathcal{U}_{S1}$, we use $\mathcal{M}_{US2}$ (trained on the other half) to evaluate the cross-entropy loss. Similarly, we use $\mathcal{M}_{US1}$ if s is present in $\mathcal{U}_{S1}$ and estimate the density metric.

Finally, we combine the diversity and density metric using the above cross-entropy losses. A sentence s in $\mathcal{U}_S$ is scored with "Cross-entropy difference" strategy using the following formula:

$$\psi_{\text{ce-diff}}(s) = H(\mathcal{M}_{LS}, s) - \mathcal{I}(s \in \mathcal{U}_{S2}) \cdot H(\mathcal{M}_{US1}, s) - \mathcal{I}(s \in \mathcal{U}_{S1}) \cdot H(\mathcal{M}_{US2}, s) \quad (1)$$

where $\mathcal{I}(s \in D)$ is 1 if s is present in D and 0 otherwise. Higher scores on $H(\mathcal{M}_{LS}, s)$ and lower scores on $H(\mathcal{M}_{US}, s)$ indicate diversity and density. Therefore, we take the difference of the two to estimate the importance of a sentence.

4 Experiments and Results

In this section, we consider English (En) and Kannada (Kn) as our language pair of interest. We chose this as it is truly low-resource, have different writing systems and replicates the challenges faced where AL is needed. We analyze the importance of the proposed techniques to integrate CLWE and evaluate several AL strategies with various annotation budgets.

4.1 Datasets

We assume the availability of monolingual data for the two languages. We use Wikipedia dumps for English and AI4Bharat-IndicNLP corpus (Kunchukuttan et al., 2020) for Kannada. We chose not to use Wikipedia for Kannada to replicate practical use cases between diverse languages.

The parallel data between the languages is from PM-India dataset (Haddow and Kirefu, 2020). We train and evaluate according to the split provided by WAT 2021 MultiIndicMT (Nakazawa et al., 2021). We created our dictionary between English and Kannada using *Kaikki*[5]. We discared entries that are not single word-word translations. Statistics about the data are summarized in Table 2.

		Total Examples		
Dataset	Type	Train	Valid	Test
Wikipedia	Mono (En)	46M	5K	5K
AI4Bharat	Mono (Kn)	15M	5K	5K
PMIndia	Parallel (En $\leftrightarrow$ Kn)	29K	1.1K	2.4K
Kaikki	Dictionary (En $\leftrightarrow$ Kn)	1.1K	–	–

Table 2: Overview of the available data.

Word Embedding	BPE		DP-BPE	
	Kn $\rightarrow$ En	En $\rightarrow$ Kn	Kn $\rightarrow$ En	En $\rightarrow$ Kn
MWE	26.5	28.7	26.4	28.7
CLWE	25.1	27.8	**27.3**	**30.0**

Table 3: Performance of word embeddings v/s pre-processing approach. We report the BLEU scores. Best scores are highlighted in **bold** for each direction.

4.2 Results on integrating dictionary

What is the benefit of applying DP-BPE and integrating CLWE? We evaluate the proposed pretraining approach described to integrate dictionaries in § 3.2. First, we create monolingual word embeddings (MWE) by joining *fasttext embeddings* for En and Kn and CLWE by mapping the MWE into a common space. Then, we pretrain the models using MWE/CLWE with standard BPE/DP-BPE techniques. Finally, we fine-tune these models on all the parallel data available and report the scores in Table 3. Comparing these 4 approaches gives us insight into the role of CLWE and DP-BPE. In the case of "MWE + DP-BPE", we do not have access to dictionary words. However, we simply assume that there is a dictionary and use that for DP-BPE. This tells us if CLWE are necessary. For "CLWE + BPE", the rare words in the dictionary would split into multiple tokens. Therefore, we removed these entries and ended with 390 word pairs in the dictionary. We mapped the monolingual embeddings with *VecMap* using only these entries. We do this experiment to evaluate the importance of rare words.

We observe similar scores for monolingual embeddings with different type of representations. This shows that the gains from applying DP-BPE are not due to better generalization as in the case of applying dropout in BPE. For CLWE, we find decrease in the performance compared to monolingual embeddings when using standard BPE. We hypothesize this is because of discarding the infrequent words when building CLWE. However, we obtain the best scores by combining CLWE with DP-BPE and gain up to 0.8 and 1.3 BLEU in English and Kannada respectively. In this case, we included the rare words in the dictionary when creating our CLWE. This shows that retaining rare words when learning the mapping between embeddings is helpful in exploiting dictionaries for NMT.

Do CLWE improve the ability to predict words in the dictionary? Evaluation metrics like BLEU is not enough to understand the models ability to predict words in the dictionary. We have to also evaluate how many times we predict these words accurately. Therefore, we calculate precision, recall and F1 scores on the dictionary words in the test set. Note that this does not consider the positional information of these words. However, we can judge them together with BLEU. If the model is predicting these words at the wrong positions, then the BLEU scores will be lower.

We consider two pretraining model configurations: 1) CLWE and DP-BPE (**With Dict**) 2) MWE with standard BPE (**No Dict**). Then, we fine-tune these models on different parallel dataset sizes. Finally, we evaluate the models ability to predict English words in the dictionary and report scores in Table 5.

We observe that the model's with CLWE are consistently better at predicting these words with relative increase of F1 score up to 3.1%. By including rare words in dictionary with help of DP-BPE, we are able to obtain higher performance on these words. Also, the scores in Table 1 show that including dictionaries with DP-BPE obtain higher BLEU. This indicates the correctness of the predicted positions. However, as we do not explicitly teach the model to predict the dictionary translation (Niehues, 2021), we don't expect significant gains.

[5] `https://kaikki.org/dictionary/Kannada/words.html`

Annotation Budget	*Selection Strategy* Random			RTTL (Zeng et al., 2019)			n-gram Overlap (Eck et al., 2005)			Cross-entropy diff (**ours**)		
	No Init	UNMT (MWE) Init	UNMT (CLWE) Init	No Init	UNMT (MWE) Init	UNMT (CLWE) Init	No Init	UNMT (MWE) Init	UNMT (CLWE) Init	No Init	UNMT (MWE) Init	UNMT (CLWE) Init
						Kn → En						
5k	7.8	16.9	18.1	-	-	-	-	-	-	-	-	-
10k	10.4	20.3	21.7	9.9	20.2	21.3	9.2	17.8	19.2	**10.6**	20.2	**22.2***
15k	12.8	22.4	23.1	11.3	22.0	**24.1***	10.2	20.2	21.2	12.0	22.1	**23.8**
20k	13.3	24.3	25.2	**13.4**	23.9	25.2	12.9	21.9	23.2	**13.5**	**24.4**	**25.5***
						En → Kn						
5k	7.6	18.7	20.3	-	-	-	-	-	-	-	-	-
10k	11.6	22.5	24.0	11.2	22.4	23.7	10.3	20.3	21.8	**12.2**	22.5	**24.6***
15k	14.5	25.1	26.5	12.9	24.1	**26.8***	12.7	23.1	24.4	14.5	25.0	26.5
20k	15.1	26.8	27.8	**15.9**	26.5	**28.4***	15.5	24.9	26.5	**16.0**	**27.0**	**28.3**

Table 4: Evaluation of AL strategies with respect to different types of pretraining modes and annoation budgets. UNMT (MWE or CLWE) indicates a UNMT model trained using MWE or CLWE while pretraining. We report BLEU scores and append * for the best model given an annotation budget. We highlight in **bold** if the score is higher than random for that pretraining configuration and budget.

	Precision (%)		Recall (%)		F1 (%)	
Dataset Size	No Dict	With Dict	No Dict	With Dict	No Dict	With Dict
10k	44.6	**49.2**	48.5	**50.0**	46.5	**49.6**
15k	46.4	**48.1**	**51.6**	51.3	48.9	**49.6**
20k	48.9	**49.0**	51.7	**52.5**	50.3	**50.7**
Full (∼30k)	51.0	**53.2**	51.4	**55.1**	52.6	**54.1**

Table 5: Impact of CLWE on the test set for predicting English words in the dictionary. We report precision, recall and F1 scores for total 2091 occurrences. Best scores for each configuration are highlighted in **bold**.

4.3 Comparision of AL Strategies

We perform a set of experiments using several AL selection strategies with multiple pretraining configurations. This enables us to assess the role of dictionary in AL and advantages of selection strategies. We consider a batch size of $5k$ and report the scores in Table 4. For the first batch, there is no available labelled data. Therefore, we randomly select $5k$ sentences and initialize our model and labelled data.

Without any initialization, we mostly do not achieve better scores than random with using RTTL or n-gram overlap strategy. Our proposed approach Cross-entropy difference is able to beat random most of the time but only with slight gains. Also, the translation quality is not adequate. For pretraining using monolingual embeddings as initialization, we only obtain slight gains than random with our strategy for a budget of $20k$. But, the performance of these models has increased significantly with at least 10 BLEU.

For models using pretraining with our proposed approach as initialization, we are consistently able to exploit AL strategies by only spending small amounts on dictionary. Random sampling with a budget of $10k$ and pretraining with monolingual embeddings achieves 20.3 BLEU when translating to English. While, "Cross-entropy difference" sampling with the same budget but using a small dictionary increases the models performance by 1.9 BLEU. This shows that building CLWE can be highly beneficial. Furthermore, we observe that the impact of CLWE decreases from around 2 to 1 BLEU as we increase the parallel data. Therefore, building CLWE has a bigger impact on very-low resource conditions and might not be as impactful with large amounts of parallel data.

We can conclude that our proposed "Cross-entropy difference" strategy is highly competitive to RTTL in almost all scenarios while RTTL being better in Kannada. However, the "n-gram overlap" strategy fails throughout all cases and shows that diversity alone is not a sufficient metric. We need to estimate both density and diversity to gain from data-driven methods for low-annotation budgets.

4.4 Impact of freezing the embedding layer

We proposed to freeze the embedding layer during all stages of training. To understand its role, we evaluate our method with/without freezing at different phases using the full dataset. We report the scores in Table 6. We observe that freezing at all stages leads to the best performance. By always providing cross-lingual and forcing the model to learn from the CLWE enables the model to exploit

Freezing	Kn → En	En → Kn
None	25.9	27.9
MLM	25.0	26.9
↪ + UNMT	25.6	28.4
↪ + Supervised NMT	**27.3**	**30.0**

Table 6: Analysis on freezing the embedding layer. We report BLEU scores starting from not freezing the embedding layer at any stage and sequentially consider freezing until each phase. ↪ + UNMT indicates freezing the embeddings at both MLM and UNMT. Best scores are highlighted in **bold**.

them in these low-resource conditions. Also, freezing during only MLM is worse than not freezing at all. We force the model to use CLWE in the pretraining stage and later allow the freedom to alter the embedding layer. We believe this hinders the ability to transfer learning and therefore does not achieve the best results. Moreover, freezing prevents erasing the knowledge from the dictionary and does not allow to drastically change the embeddings weights based on limited parallel data.

5 Related Work

There are several works on AL in the context of MT (Eck et al., 2005; Haffari et al., 2009; Ambati, 2012). These methods operated and evaluated using phrase-based machine translation systems. Zeng et al. (2019) provides a comprehensive summary of AL strategies using the current SOTA transformer architecture. They propose a novel model-driven strategy RTTL and show it outperforms other data-driven methods. However, they consider large annotation budgets in their analysis. We focus on scenarios with small budgets and show that the model's quality is insufficient to exploit this strategy. We show pretraining is necessary to enable model-driven sampling methods like RTTL in low budgets. Moreover, we propose a data-driven strategy "Cross-entropy difference" adapted from (Moore and Lewis, 2010), that is competitive to RTTL in these challenging low-resource conditions.

Instead of relying on heuristics with selection strategies, Liu et al. (2018) uses Deep Imitation Learning to learn the best way to sample using a high resource language pair. They also consider a scenario of limited labelling budgets ($10k$) and show their approach's effectiveness. However, these methods are computationally expensive and rely on having auxiliary parallel data.

Our pretraining approach is similar to and largely inspired from (Chronopoulou et al., 2021). Their work operates only on sub-word level data using identical sub-words as a seed dictionary to build CLWE. They show that these lexically aligned embeddings are beneficial when training a UNMT system between distant languages. We use this approach to include a dictionary and provide better supervision for the pretrained model. We show how we can further include the rare words using DP-BPE, when learning the mapping between monolingual embeddings.

6 Conclusion

The main goal of the paper was to design a high-quality NMT system with limited annotation costs. To achieve this, we designed a cost-effective training procedure by proposing improvements in multiple avenues. First, we showed the necessity of pretraining with monolingual data. This is useful as the monolingual data does not require any labelling and improves the models significantly. Moreover, it enables us to gain from selection strategies. Second, we suggested a pretraining procedure by integrating a dictionary which can be created cheaply. We proposed DP-BPE to include the rare words in the dictionary while learning the alignment. Further, we showed the importance of including these rare words from our experiments. Using our approach, we were able to increase the models ability to predict these words. Finally, we presented a novel data-driven strategy "Cross-entropy difference" that is helpful in low-resource scenarios. We empirically showed that sampling using our strategy achieves better scores than random consistently and is competitive to the SOTA approach RTTL.

Pretraining with auxiliary data of similar high-resource languages can substantially increase the model's quality. Building such multilingual models can greatly increase the potential of model-driven strategies. Also, designing AL strategies for constructing a dictionary can even further decrease costs while increasing gains. We leave these directions as future work.

References

Vamshi Ambati. 2012. *Active learning and crowdsourcing for machine translation in low resource scenarios*. Ph.D. thesis, Carnegie Mellon University.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Tamali Banerjee, Rudra V Murthy, and Pushpak Bhattacharya. 2021. Crosslingual embeddings are essential in UNMT for distant languages: An English to IndoAryan case study. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 23–34, Virtual. Association for Machine Translation in the Americas.

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.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2021. Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 173–180, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low cost portability for statistical machine translation based on n-gram frequency and tf-idf. In *International Workshop on Spoken Language Translation (IWSLT) 2005*.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur Parikh. 2021. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1126–1137.

Barry Haddow and Faheem Kirefu. 2020. PMIndia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423, Boulder, Colorado. Association for Computational Linguistics.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Sai Koneru, Danni Liu, and Jan Niehues. 2021. Unsupervised machine translation on dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 55–64.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020. Reference language based unsupervised neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4151–4162, Online. Association for Computational Linguistics.

Ming Liu, Wray Buntine, and Gholamreza Haffari. 2018. Learning to actively learn neural machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 334–344, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Jan Niehues. 2021. Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, 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 *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Álvaro Peris and Francisco Casacuberta. 2018. Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 151–160, Brussels, Belgium. Association for Computational Linguistics.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Empirical evaluation of active learning techniques for neural mt. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 84–93.

A Appendix

A.1 AL Framework

Algorithm 1 General AL Algorithm

Require: Parallel Data $\mathcal{D}_P$,
Monolingual Data $\mathcal{D}_M$,
Unlabelled in-domain source data $\mathcal{U}_S$,
Batch size $\mathcal{B}$, Selection strategy $\psi()$
$\mathcal{M}_{PRE} \leftarrow PRETRAIN(\mathcal{D}_M, empty)$;
$\mathcal{M} \leftarrow SNMT(\mathcal{D}_P, \mathcal{M}_{PRE})$;
while $Budget \neq 0$ **do**
  for $x \in \mathcal{U}_S$ **do**
    $f(x) \mathrel{+}= \psi(x, \mathcal{U}_S, \mathcal{D}_P, \mathcal{M})$;
  end for
  $X_B = Topscoring(f(x), \mathcal{B})$;
  $Y_B = HumanTranslated(X_B)$;
  $\mathcal{U}_S = \mathcal{U}_S - X_B$;
  $\mathcal{D}_P = \mathcal{D}_P \cup (X_B, Y_B)$;
  $\mathcal{M} \leftarrow SNMT(\mathcal{D}_P, \mathcal{M}_{PRE})$;
end while
return $\mathcal{M}, \mathcal{D}_P$

A.2 Pre-processing and Hyperparameters

We tokenize the data with Moses (Koehn et al., 2007) for English and *Indic-NLP-Library*[6] for Kannada. We learn sub-words using BPE (Sennrich et al., 2016b) with $50k$ merge operations on concatenating subset of English and Kannada data. We report detokenized BLEU (Papineni et al., 2002) using SacreBLEU[7] (Post, 2018). We use the SOTA

[6] `https://github.com/anoopkunchukuttan/indic_nlp_library`

[7] BLEU+case.mixed+numrefs.1+smooth.exp+tok.spm+version.1.4.12

Transformer architecture (Vaswani et al., 2017) for building NMT models. For pretraining, we use a transformer with 6 layers and 8 heads and an embedding dimension of 1024. While fine-tuning on the parallel data, we use label-smoothing of 0.2, activation dropout of 0.2 and attention dropout of 0.2 as we have limited data. The language models for the "Cross-entropy difference" strategy use the pretrained MLM model as initialization before training on the CLM objective. For the models that do not use any initialization in Table 4, we use a smaller model with 5 layers and 2 heads and an embedding dimesion of 512. We use the same value for the regularization parameters as mentioned in the pretraining architecture. Furthermore, the CLM for the "Cross-entropy difference" strategy use a transformer with 3 layers and 2 heads as there is no pretrained model. We use the *XLM*[8] code base to perform our experiments and set the other parameters to default.

[8] https://github.com/facebookresearch/XLM