Active Learning Approaches to Enhancing Neural Machine Translation

Yuekai Zhao[1] **Haoran Zhang**[1] **Shuchang Zhou**[2] **Zhihua Zhang**[3]
[1] Academy for Advanced Interdisciplinary Studies, Peking University
[2] Megvii Inc.
[3] School of Mathematical Sciences, Peking University
{yuekaizhao, haoran_zhang}@pku.edu.cn
zsc@megvii.com
zhzhang@math.pku.edu.cn

Abstract

Active learning is an efficient approach for mitigating data dependency when training neural machine translation (NMT) models. In this paper we explore new training frameworks by incorporating active learning into various techniques such as transfer learning and iterative back-translation (IBT) under a limited human translation budget. We design a word frequency based acquisition function and combine it with a strong uncertainty based method. The combined method steadily outperforms all other acquisition functions in various scenarios. As far as we know, we are the first to do a large-scale study on actively training Transformer (Vaswani et al., 2017) for NMT. Specifically, with a human translation budget of only 20% of the original parallel corpus, we manage to surpass Transformer trained on the entire parallel corpus in three language pairs.

1 Introduction

Many impressive progresses have been made in neural machine translation (NMT) in the past few years (Luong et al., 2015; Gehring et al., 2017; Vaswani et al., 2017; Wu et al., 2019). However, the general training procedure requires tremendous amounts of high-quality parallel corpus to achieve a deep model's full potential. The scarcity of the training corpus is a common problem for many language pairs, which might lead to the NMT model's poor performance.

However, constructing a parallel corpus is a slow and laborious process. Professional human translators and well-trained proofreaders are needed. Although several dual learning (He et al., 2016; Bi et al., 2019) and unsupervised learning (Artetxe et al., 2018; Lample et al., 2017; Lample and Conneau, 2019) approaches have been successfully used, they are often inferior to the supervised models. In such cases, active learning might be a good choice. The goal of active learning in NMT is to train a well-performing model under a limited human translation budget. We achieve this goal by using some particularly designed acquisition functions to select informative sentences to construct a training corpus.

Acquisition functions can be categorized into two types: model related and model agnostic. For the former, the methods we use are all based on the idea of uncertainty. For the latter, we devise a word frequency based method which takes linguistic features into consideration. Both types of acquisition functions have been proven to be beneficial in active NMT training, especially when they are appropriately combined.

Data augmentation techniques that consume no human translation budget are worth exploring in active NMT training. If the parallel corpus of a related language pair is available, transfer learning (Zoph et al., 2016; Kim et al., 2019) might be a good choice. Otherwise, we propose a new training framework that integrates active learning with iterative back-translation (IBT) (Hoang et al., 2018). We achieve success in both the settings, especially when active learning bonds with IBT.

The main contributions of this work are listed as follows: 1) To the best of our knowledge, we are the first to give a comprehensive study of active learning in NMT under various settings. 2) We propose a word frequency based acquisition function which is model agnostic and effective. This acquisition function can further enhance existing uncertainty based methods, achieving even better results in all settings. 3) We design a new training framework for active iterative back-translation as well as a simple data augmentation technique. With a human translation budget of only 20% of the original parallel corpus, we can achieve better BLEU scores than the fully supervised Transformer does (Vaswani et al., 2017).

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1796–1806
November 16 - 20, 2020.

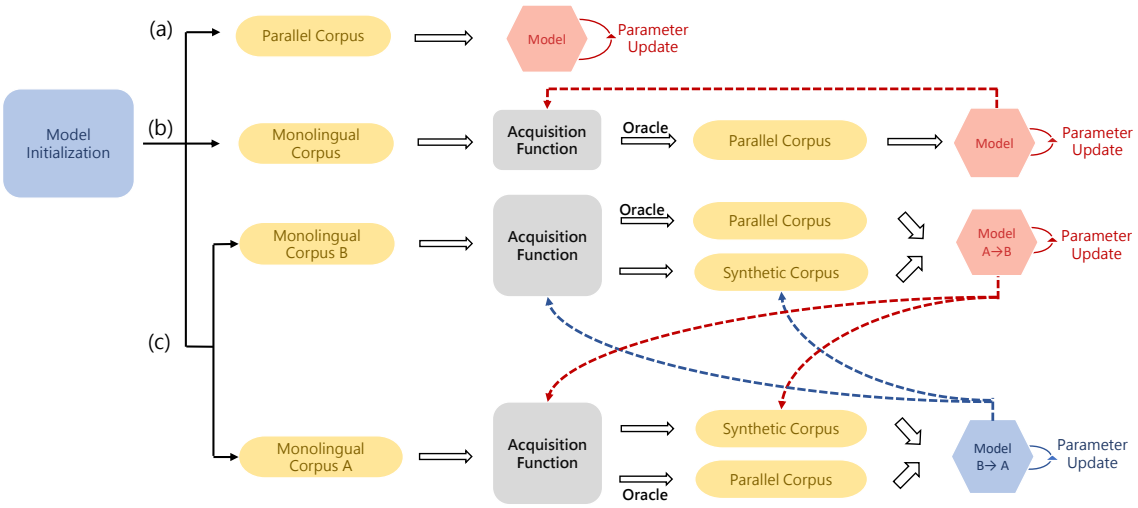


Figure 1: (a) shows the diagram of vanilla supervised NMT training. A parallel corpus is available and used to train the model. (b) shows active NMT training. An acquisition function can use the model to score each sentence in the source side monolingual corpus. A parallel corpus is gradually constructed by employing an oracle (human translator) to translate the sentences with high scores. (c) shows active iterative back-translation. An acquisition function can use $Model_{A\to B}$ to score the untranslated sentences in language A. One part of the high score sentences are translated by an oracle (new parallel corpus), another part are translated by $Model_{A\to B}$ (new synthetic corpus). New parallel corpus and new synthetic corpus are used for training $Model_{B\to A}$ and vice versa.

2 Related Work

Active learning As for natural language processing, active learning is well studied in text classification (Zhang et al., 2017; Ru et al., 2020) and named entity recognition (Shen et al., 2017; Siddhant and Lipton, 2018; Prabhu et al., 2019). Peris and Casacuberta (2018) applied attention based acquisition functions for NMT. Liu et al. (2018) introduced reinforcement learning to actively train an NMT model.

Data selection in NMT Although active learning has not been thoroughly studied in NMT, the related data selection problem attracts some attention. van der Wees et al. (2017); Wang et al. (2018a) deliberately designed weighted sampling methods, which accelerates training and improves performance. Wang et al. (2018b); Pham et al. (2018) focused on noisy data, coming up with algorithms to filter harmful sentence pairs. Wang et al. (2019) simultaneously dealt with domain data selection and clean data selection. Fadaee and Monz (2018); Poncelas et al. (2019); Dou et al. (2020) considered domain data selection in back-translation. Wang and Neubig (2019) proposed a method to select relevant sentences from other languages to bring performance gains in low resource NMT. Furthermore, Ruiter et al. (2019) tried to extract possible parallel data from bilingual Wikipedia.

Interactive NMT Interactive NMT exploits user feedback to help improve translation systems. Real-world (Kreutzer et al., 2018) or simulated user feedback includes highlighting accurate translation chunks (Petrushkov et al., 2018) or correct errors made by machine (Peris and Casacuberta, 2018; Domingo et al., 2019). Kreutzer and Riezler (2019) took the cost of different types of supervision (feedback) into account, which resembles the idea of active learning.

3 Methodology

We give a detailed description of active neural machine translation (NMT) in this section. Basic settings and some terminologies are introduced in Section 3.1. In Section 3.2 and Section 3.3, various acquisition functions are presented and explained. Section 3.4 deals with combining active learning with transfer learning and iterative back-translation. Figure 1 is an illustration of different training frameworks in NMT.

3.1 Active NMT

Several terminologies need to be clarified before introducing the active NMT circulation, namely, acquisition function, oracle and budget.

Acquisition Function An acquisition function gives a score to each untranslated sentence in the monolingual corpus. Sentences with higher scores are more likely to be selected as the training corpus. Acquisition functions fall into two types, model related and model agnostic. A model related acquisition function takes a sentence as the model input and gives a score depending on the model output. A model agnostic acquisition function often concerns about the informativeness of the sentence itself, which can score each sentence before training the model.

Oracle An oracle is a gold standard for a machine learning task. For NMT, an oracle can output the ground truth translation given a source sentence (specifically an expert human translator). A parallel corpus is gradually constructed by employing an oracle to translate the selected sentences.

Budget Budget means the total cost one can afford to employ an oracle. For NMT, we need to hire human experts to translate sentences. In order to simulate active NMT training, throughout all our experiments, the cost is the number of words been translated.

In the beginning, we have a large-scale monolingual corpus of the source language. We do several rounds of active training until the total budget is used up. In each round, five steps are taken:

- Use an acquisition function to score each untranslated sentence.
- Sort the untranslated sentences according to the scores in descending order.
- Select high score untranslated sentences until the token budget in this round is used up.
- Remove the selected sentences from the monolingual corpus and employ an oracle to translate them.
- Add these new sentence pairs to the parallel corpus and retrain the NMT model.

Transformer is what we use throughout our experiments. As this architecture is commonly used and our implementation has little difference with the original, we skip an exhaustive background description of the underlying model. One can refer to Vaswani et al. (2017) for some details. The active NMT training circulation is shown in part (b) of Figure 1.

3.2 Model Related Acquisition Functions

All model related acquisition functions we try are based on uncertainty. Settles and Craven (2008) tried these methods on sequence labeling tasks. For NMT, we use greedy decoding to generate a synthetic translation of each sentence $x = (x_1, \cdots, x_n)$ in the monolingual corpus U. We denote this synthetic translation as $\hat{y} = (\hat{y}_1, \cdots, \hat{y}_m)$. In the i^{th} decoding step, the model outputs a probability distribution over the entire vocabulary $P_\theta(\cdot|x, \hat{y}_{<i})$.

Least Confident (lc) A direct interpretation of model uncertainty is the average confidence level on the generated translation. We strengthen the model on its weaknesses and force it to learn more on intrinsically hard sentences.

$$\frac{1}{m}\sum_{i=1}^{m}\Big[1 - P_\theta(\hat{y}_i|x, \hat{y}_{<i})\Big] \quad (1)$$

Minimum Margin (margin) Margin means the average probability gap between the model's most confident word $y^*_{i,1}$ and second most confident word $y^*_{i,2}$ in each decoding step. With a small margin, the model is unable to distinguish the best translation from an inferior one.

$$-\frac{1}{m}\sum_{i=1}^{m}\Big[P_\theta(y^*_{i,1}|x, \hat{y}_{<i}) - P_\theta(y^*_{i,2}|x, \hat{y}_{<i})\Big] \quad (2)$$

Token Entropy (te) Concentrated distributions tend to have low entropy. Entropy is also an appropriate measurement of uncertainty. In NMT, we calculate the average entropy in each decoding step as given by the following equation.

$$\frac{1}{m}\sum_{i=1}^{m} entropy(P_\theta(\cdot|x, \hat{y}_{<i}))) \quad (3)$$

Total Token Entropy (tte) To avoid favoring long sentences, we choose to take average over sentence length in the above three methods. However, it remains a question whether querying long sentences should be discouraged. We design an acquisition function to figure out this issue by removing the $\frac{1}{m}$ term from Token Entropy.

3.3 Model Agnostic Acquisition Functions

Uncertainty based acquisition functions depend purely on probability. We propose a model agnostic acquisition function that focuses on linguistic features. In NMT, it is important to enable the model

Algorithm 1 Decay Logarithm Frequency Acquisition Function

Input:
Selected Corpus L, Untranslated Corpus U;
Token Budget b;
Positive Constants λ_1, λ_2;
Output: New Selected Sentences B

$B = \emptyset; \hat{U} = \emptyset$
for s in U **do**
  calculate $lf(s)$ by Equation (6)
end for
for s in sort(U) by lf score **do**
  calculate $delfy(s)$ by Equation (7)
  $\hat{U} = \hat{U} \cup \{\text{s}\}$
end for
for s in sort(U) by $delfy$ score **do**
  if Cost($B \cup \{\text{s}\}$) $> b$ **then**
    break
  end if
  $B = B \cup \{\text{s}\}$
end for

to translate unseen future sentences. In other words, we wish to choose those sentences that are representatives of all the untranslated sentences but less similar with what has previously been selected.

In each active training round, we have a set of untranslated sentences in the source language side, which is denoted as U. Also, those sentences that have been selected in previous active training rounds are denoted as L. We denote a sentence as $s = (s_1, \cdots, s_K)$ which is different from what it is in Section 3.2 because we are now working on word level instead of the subword level. First, we define the logarithm frequency of a word w in U, namely, $F(w|U)$.

$$G(w|U) = \log(C(w|U) + 1) \tag{4}$$

$$F(w|U) = \frac{G(w|U)}{\sum_{w' \in U} G(w'|U)} \tag{5}$$

Where $C(w|\cdot)$ means the occurrence number of a word w in a certain sentence set.

As shown in Equation (6), the representativeness of a sentence s is determined by its average logarithm word frequency in U. A decay factor $\lambda_1 \geq 0$ is introduced to assist the model to pay more attention to the uncommon words in the previously selected corpus L.

$$lf(s) = \frac{\sum_{i=1}^{K} F(s_i|U) \times e^{-\lambda_1 C(s_i|L)}}{K} \tag{6}$$

Directly using lf scores is problematic. The algorithm favors a small number of function words (like "a", "the") which account for a high proportion of the entire corpus. Also, redundancy breaks out since sentences of similar content share similar scores. These two drawbacks are disastrous for building a well-performing translation system.

A gradual reranking is used to ease these two problems. Equation (6) is employed for the first round of sorting. $\hat{U}(s)$ is the set of all sentences that have a higher lf score than s. If s has a high lf score, but each word s_i in s frequently appears in $\hat{U}(s)$, we use a decay term $e^{-\lambda_2 C(s_i|\hat{U}(s))}$ to cut down its score. In this way, we tend to discard repetitive sentences and filter out insignificant function words. Details can be found in Equations (7) and (8). λ_1 and λ_2 are non-negative constants.

$$delfy(s) = \frac{\sum_{i=1}^{K} F(s_i|U) \times Decay(s_i)}{K} \tag{7}$$

$$Decay(s_i) = e^{-\lambda_1 C(s_i|L)} \times e^{-\lambda_2 C(s_i|\hat{U}(s))} \tag{8}$$

We name this model agnostic acquisition function as **de**cay **l**ogarithm **f**requenc**y** (**delfy**) which is summarized in Algorithm 1.

3.4 Active NMT with Data Augmentation

Directly incorporating active learning into NMT can be beneficial. However, is there any technique that consumes no extra budget to further improve translation performance? The answer depends on the availability of some related parallel corpus. Transferring knowledge from a related language pair can be considered if an extra parallel corpus is available. Iterative back-translation is worth trying if not.

Transfer Learning We assume that there exists a rich parallel corpus in a related translation direction, *e.g.*, we try to build a German-English NMT system and we have access to French-English sentence pairs. The model is initialized by training on this related parallel corpus. Active NMT training is carried out as described in Section 3.1 after model initialization.

Iterative Back-Translation Iterative back-translation (IBT) (Sennrich et al., 2016a; Hoang et al., 2018) proves to be of help in boosting model performance. IBT offers a data augmentation technique that is budget free (no human translator needed) when considering active NMT training. However, simply using all monolingual corpus

Algorithm 2 The Framework for Active Iterative Back-Translation (IBT)

Input:
Active IBT Rounds R;
Parallel Corpus $L = \{L_A, L_B\}$;
Monolingual Corpus U_A,U_B;
Initialized NMT Model $M_{A\to B}$, $M_{B\to A}$;
Acquisition Function Φ;
Token Budget b, Oracle O;
Token Number in Synthetic Sentences α;
Output: $M_{A\to B}$, $M_{B\to A}$;

for i in 1 to R **do**
  $\overrightarrow{A_i} = \Phi(U_A, L_A, M_{A\to B}, b)$
  $\overrightarrow{B_i} = O(\overrightarrow{A_i})$; $U_A = U_A \setminus \overrightarrow{A_i}$
  $\overrightarrow{P_i} = \Phi(U_A, L_A, M_{A\to B}, \alpha)$
  $\overrightarrow{Q_i} = M_{A\to B}(\overrightarrow{P_i})$
  $L_A = L_A \cup \overrightarrow{A_i}$, $L_B = L_B \cup \overrightarrow{B_i}$
  Train $M_{B\to A}$ on $\{(L_B \cup \overrightarrow{Q_i}), (L_A \cup \overrightarrow{P_i})\}$
  $\overleftarrow{B_i} = \Phi(U_B, L_B, M_{B\to A}, b)$
  $\overleftarrow{A_i} = O(\overleftarrow{B_i})$; $U_B = U_B \setminus \overleftarrow{B_i}$
  $\overleftarrow{Q_i} = \Phi(U_B, L_B, M_{B\to A}, \alpha)$
  $\overleftarrow{P_i} = M_{B\to A}(\overleftarrow{Q_i})$
  $L_A = L_A \cup \overleftarrow{A_i}$, $L_B = L_B \cup \overleftarrow{B_i}$
  Train $M_{A\to B}$ on $\{(L_A \cup \overleftarrow{P_i}), (L_B \cup \overleftarrow{Q_i})\}$
end for

to generate a synthetic parallel corpus will hurt instead of improving the model performance. We designed some experiments to validate this argument. Detailed results can be seen in Appendix B.

Two reasons may cause these poor results. First, the quality of synthetic corpus varies. Some of the synthetic sentence pairs can be beneficial, while others only introduce chaos into the NMT model. Second, the percentage of the synthetic corpus in the entire training corpus is too high. To cope with these two problems, we propose a new Active IBT framework. Models of opposite translation directions are responsible for constructing training corpus for each other. Sentences with the highest acquisition function scores are divided into two parts. One part is translated by an oracle to enrich the parallel corpus. Another part is used to generate a new synthetic corpus. In this way, we manage to control the quality as well as the percentage of the synthetic corpus.

This framework is shown in part (c) of Figure 1, and some details can be found in Algorithm 2.

Algorithm 3 Active IBT++ (LAN A to LAN B)

Input:
Active IBT Rounds R; Merge Number k_1, k_2;
Final Parallel Corpus $L^{++} = \{L_A, L_B\}$;
$M_{A\to B,i}$, $M_{B\to A,i}$, $i \in \{1, 2, \cdots, R\}$;
Synthetic Corpus $\overleftarrow{P_i}$, $\overleftarrow{Q_i}$, $i \in \{1, 2, \cdots, R\}$;
Output: $M_{A\to B}$;

for j in 1 to k_1 **do**
  $\tilde{L}_{A,j} = M_{B\to A,R-j+1}(L_B)$;
  $\tilde{L}_{B,j} = M_{A\to B,R-j+1}(L_A)$;
  $L^{++} = L^{++} \cup \{\tilde{L}_{A,j}, L_B\} \cup \{L_A, \tilde{L}_{B,j}\}$
end for
for j in 1 to k_2 **do**
  $L^{++} = L^{++} \cup \{\overleftarrow{P}_{R-j+1}, \overleftarrow{Q}_{R-j+1}\}$
end for
$M_{A\to B} =$ Retrain $M_{A\to B,1}$ on L^{++}

Active IBT++ Active learning aims at choosing informative sentences to train the model. Is there any way that we can exploit more value from these selected sentences? Inspired by Nguyen et al. (2019), we propose some further data augmentation techniques after Active IBT is done. Models of the last k_1 rounds are used for translating the final parallel corpus, such that each selected sentence will have diversified translations. We merge the diversified parallel corpus with the synthetic corpus of a specific translation direction in the last k_2 rounds. Duplicate sentence pairs are filtered out. The NMT model is re-initialized and trained on this enlarged training corpus.

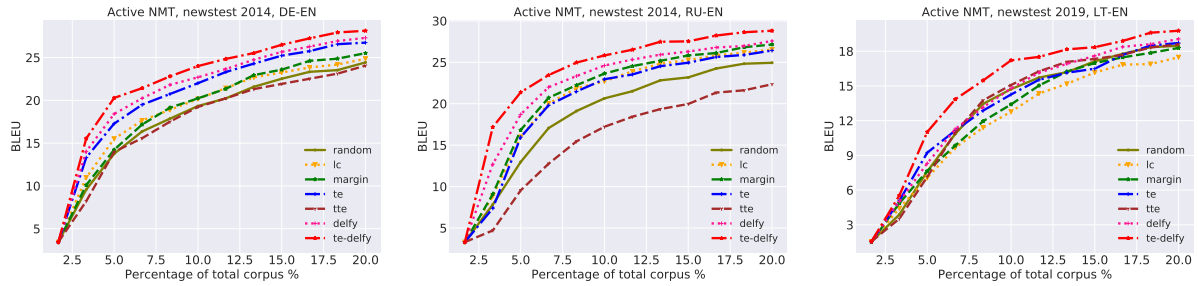
We name this technique Active IBT++ and summarize it in Algorithm 3. For simplicity, we only consider one translation direction in Algorithm 3. The same technique can be easily done in another translation direction.

4 Experiments

4.1 Dataset, Preprocessing and Implementation

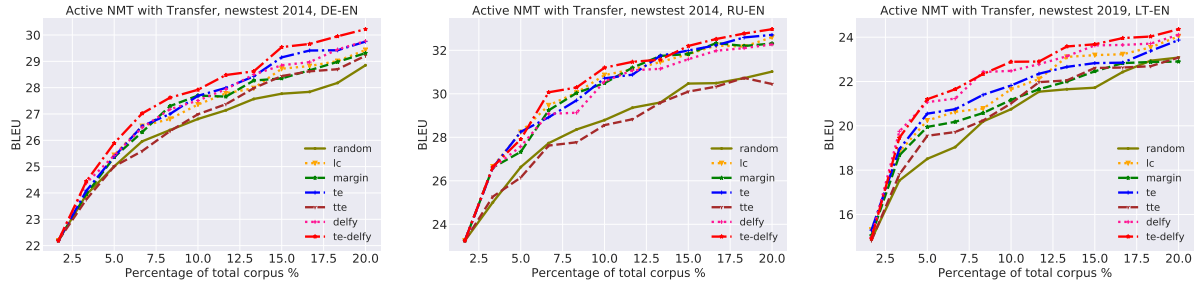
We experiment on three language pairs, namely, German-English (DE-EN), Russian-English (RU-EN) and Lithuanian-English (LT-EN). To simulate active NMT training, we use parallel corpus from the WMT 2014 shared task (DE-EN, RU-EN) and the WMT 2019 shared task (LT-EN). For Russian-English, we randomly choose extra 2M sentence pairs from the UN corpus[1]. The number of sen-

[1]https://conferences.unite.un.org/UNCorpus/



(a) news test 2014, DE-EN (b) news test 2014, RU-EN (c) news test 2019, LT-EN

Figure 2: Active NMT, BLEU scores on the test dataset.



(a) news test 2014, DE-EN (b) news test 2014, RU-EN (c) news test 2019, LT-EN

Figure 3: Active NMT with Transfer Learning, BLEU scores on the test dataset.

tence pairs in each language pair is 4M (DE-EN), 4M (RU-EN) and 0.8M (LT-EN). Tokenization is done by Moses[2]. We employ BPE (Sennrich et al., 2016b) to generate a shared vocabulary for each language pair. The BPE merge operation numbers are 20K (LT-EN), 32K (DE-EN,RU-EN). For active NMT with or without transfer learning, we only experiment on translating into English. Instead, for active iterative back-translation (IBT), evaluation is carried out on translating from English and into English. The evaluation metric is BLEU (Papineni et al., 2002).

Model hyper parameters are identical to Transformer base (Vaswani et al., 2017). Adam optimizer (Kingma and Ba, 2014) is used with a learning rate of 7×10^{-4}. We use the same learning rate scheduling strategy as Vaswani et al. (2017) does with a warmup step of 4000. During training, the label smoothing factor and the dropout probability are set to 0.1. λ_1, λ_2 in Algorithm 1 are all set to 1.0.

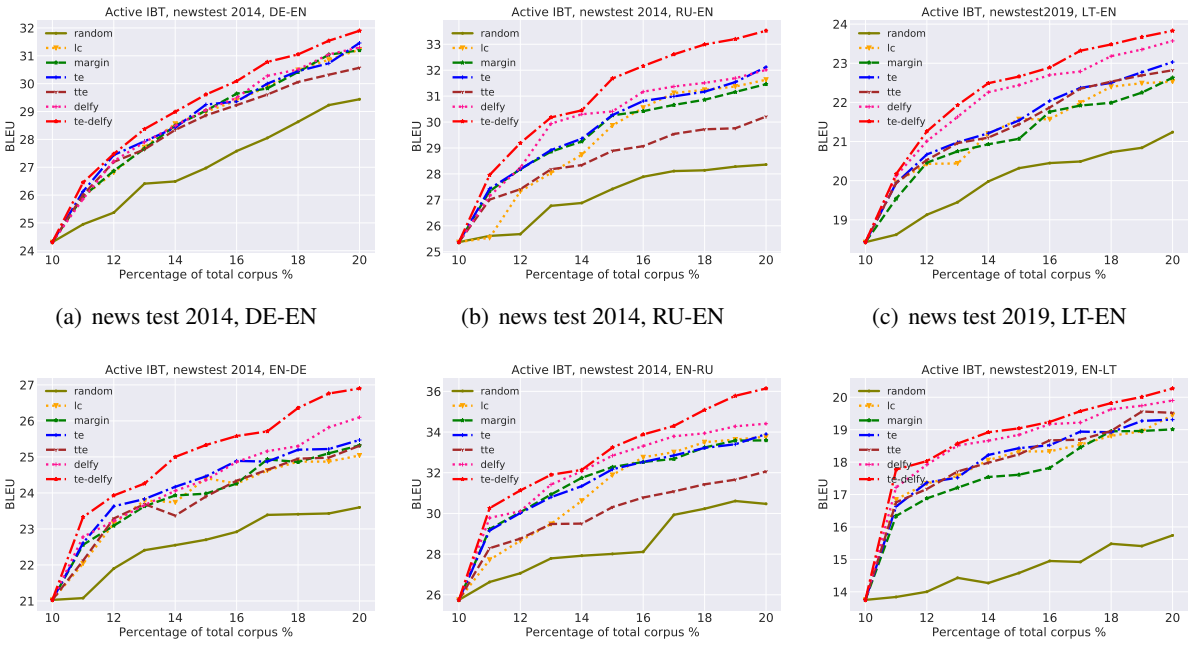
Our implementation is based on pytorch[3]. All models are trained on 8 RTX 2080Ti GPU cards with a mini-batch of 4096 tokens. We stop training if validation perplexity does not decrease for 10 epochs in each active training round.

4.2 Active NMT

As a starting point, we empirically compare different acquisition functions proposed in Section 3.2 and Section 3.3, as well as the uniformly random selection baseline. Twelve rounds of active NMT training are done. In each round, 1.67% of the entire parallel corpus is selected and added into the training corpus. Thus, we ensure the token budget is 20% of the entire parallel corpus in the final round. Training corpus in the first round is identical across different acquisition functions to ensure the fairness of comparison.

Results are shown in Figure 2. Most active acquisition functions can outperform the random selection baseline in all three language pairs. Our model agnostic acquisition function (**delfy**) is also better than the best uncertainty based acquisition function. We try to combine **delfy** with some well-performing uncertainty based acquisition functions since they represent different aspects of the informativeness of a sentence. We choose to combine **delfy** with token entropy (**te**). We add the ranks given by these two acquisition functions to avoid the magnitude problem. For example, if a sentence

[2]https://github.com/moses-smt/mosesdecoder
[3]http://pytorch.org/



(a) news test 2014, DE-EN (b) news test 2014, RU-EN (c) news test 2019, LT-EN

(d) news test 2014, EN-DE (e) news test 2014, EN-RU (f) news test 2019, EN-LT

Figure 4: Active Iterative Back-Translation, BLEU scores on the test dataset.

gets the highest **delfy** score as well as the second-highest **te** score, then its **delfy** rank is 1 and its **te** rank is 2, such that its final score is $1 + 2 = 3$. Since we sort sentences in descending order of their scores, we should multiply the summation of the ranks by -1. This new combined acquisition function is named as **te-delfy**.

Our combined method (**te-delfy**) proves to be more effective, outperforming all the other acquisition functions in each active NMT training round in all three language pairs. To be more specific, in the last active training round, **te-delfy** surpasses the best uncertainty based acquisition function by 1.4 BLEU points in DE-EN, 1.6 BLEU points in RU-EN and 1.1 BLEU points in LT-EN.

4.3 Active NMT with Transfer Learning

To evaluate different acquisition functions in active NMT with transfer learning, we start from a French to English NMT model. The parallel corpus for building this initial model contains 4M sentence pairs which are randomly selected from the WMT 2014 shared task. To share vocabulary between different languages, we latinize all the Russian sentences[4].

Figure 3 shows the results. All the active acquisition functions are still advantageous compared with the random selection baseline except total token entropy (**tte**). Our combined method (**te-delfy**) is also the best in most active training rounds. **Te-delfy** yields the best final results, beating the best uncertainty based acquisition function by 0.5 BLEU points in DE-EN, 0.3 BLEU points in RU-EN and 0.5 BLEU points in LT-EN. However, in active NMT with transfer learning, the performance gains brought by different acquisition functions are not as much as it is in active NMT (Section 4.2).

4.4 Active Iterative Back-Translation

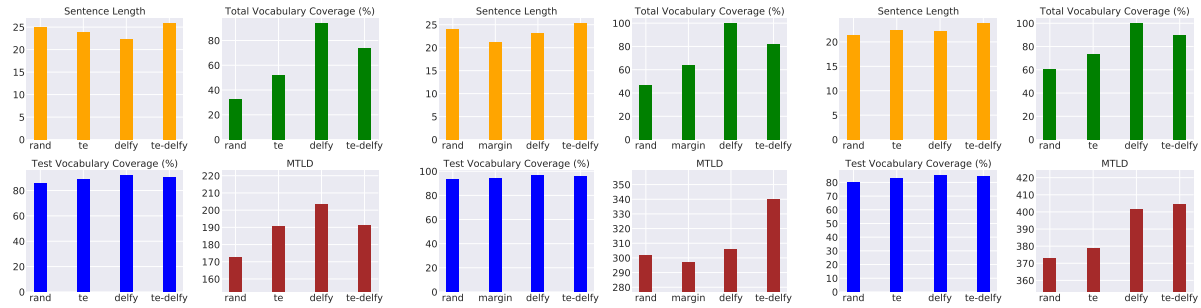
For active iterative back-translation (IBT), we randomly select 10% of the entire parallel corpus to train an initial NMT model. The initial model is shared across different acquisition functions. We do 10 rounds of Active IBT training. In each round, 1% of the entire parallel corpus is added into the training corpus. The total token budget is still 20% as in Section 4.2 and Section 4.3. For α in Algorithm 2, we use as many as half of the amount of the authentic parallel corpus in this Active IBT round. k_1, k_2 in Algorithm 3 are set to 3 and 6 respectively.

Results are summarized in Figure 4. Our combined method (**te-delfy**) becomes even more powerful than it is in active NMT, leading all the way until the final round in all the experiments. All active acquisition functions we try surpass the random

[4]https://github.com/barseghyanartur/transliterate

Method	Setting	DE→EN	EN→DE	RU→EN	EN→RU	LT→EN	EN→LT
Transformer Base	Entire Corpus	32.5	27.3	33.9	36.6	24.2	20.3
Random	Active IBT	29.4	23.6	28.4	30.5	21.2	15.7
Best Uncertainty	Active IBT	31.5	25.5	32.1	33.9	23.0	19.5
Delfy (Ours)	Active IBT	31.3	26.1	32.0	34.4	23.6	20.0
Te-delfy (Ours)	Active IBT	**31.9**	**26.9**	**33.5**	**36.1**	**23.8**	**20.3**
Te-delfy (Ours)	Active IBT++	**32.8**	**27.4**	**35.0**	**37.4**	**25.4**	**21.3**

Table 1: Comparison between Active IBT models in the final round, Active IBT++ models and the full supervision Transformer. Best results are all achieved by **Te-delfy**. The token budget is 20% of the entire parallel corpus.



(a) DE-EN (b) RU-EN (c) LT-EN

Figure 5: Text analysis of selected sentences, including average sentence length, vocabulary coverage and MTLD score.

baseline by a large margin, with a minimum performance gain of 1.1 BLEU points. We argue that synthetic sentence pairs need more sophisticated selection criteria than the authentic ones. Low-quality pseudo-parallel data can damage rather than help the model performance.

We make a comparison between the actively learned models and the full supervision Transformer in Table 1. The best results are all achieved by **te-delfy** which further proves its superiority. Active IBT++ (Algorithm 3) is applied with **te-delfy**. With a token budget of 20% of the entire parallel corpus, we can surpass the vanilla Transformer in every translation direction. These results show that Active IBT and Active IBT++ are promising approaches for enhancing NMT models.

5 Analysis

5.1 Linguistic Features

In order to find the common features of the beneficial sentences in translation, we analyze the final parallel corpus constructed by different acquisition functions in active NMT from four aspects. All the analyses are done on word level instead of the subword level. First, we study the impact of the average sentence length. Second, we study the vocabulary coverage by calculating the ratio of the vocabulary size of the selected corpus to the total/test vocabulary size. Finally, the lexical diversity of the selected corpus is analyzed based on the MTLD metric (McCarthy and Jarvis, 2010). Analyses are done on random selection, the best uncertainty based method, **delfy** and **te-delfy**. The results are shown in Figure 5.

Most algorithms tend to choose some medium-length sentences, rather than the extremely long or short ones. We also use sentence length as our acquisition function (choosing the longest or shortest sentences), which proves to be terrible (Appendix A). Vocabulary coverage varies among different acquisition functions, with random selection always being the lowest one. Higher vocabulary coverage means fewer unseen words which might create a more knowledgeable model. Also, **delfy** and **te-delfy** always achieve higher MTLD scores than the other two methods do. Note that a higher vocabulary coverage does not necessarily mean a higher diversity score. In LT-EN and RU-EN, **delfy** always has a larger vocabulary size than **te-delfy**, but its selected corpus is less diverse. In general, a good acquisition function should favor medium-length sentences as well as having a large vocabulary cov-

erage. Meanwhile, diversified training corpus is also beneficial to model performance.

Methods	Easy→Hard	Hard→Easy
lc	16.0	17.5
margin	16.3	18.3
te	15.9	18.7
tte	16.1	18.6
delfy	16.9	19.1
te-delfy	16.0	19.8

Table 2: We validate the necessity of active learning when there is a limited human translation budget. Hard → Easy corresponds to active learning. Easy → Hard represents reverse active learning. We experiment on EN-LT with a token budget of 20% of the entire parallel corpus. Active learning results are always better than reverse active learning results.

5.2 Reverse Active learning

Active learning chooses difficult samples for the model. Instead, several curriculum learning methods (Zhang et al., 2018; Platanios et al., 2019; Liu et al., 2020; Zhou et al., 2020) accelerates model convergence, which starts training with easy data samples and gradually moves to hard ones. Curriculum learning's success makes it reasonable to think about whether the reverse of active learning is also beneficial. Reverse active learning selects sentences with the lowest acquisition function scores in each round. We make a comparison between active learning and reverse active learning in Table 2. Reverse active learning lags behind active learning with all acquisition functions we try. Also, reverse active learning can not beat the random baseline of 18.5 BLEU points. Curriculum learning emphasizes the training process of networks (easy to hard), which might accelerate convergence. However, when the amount of training data is limited, active learning is a better choice.

6 Conclusion

Various acquisition functions are conducted on active NMT, active NMT with transfer learning and active iterative back-translation (IBT). Our experiment results strongly prove that active learning is beneficial to NMT. Our combined method (**te-delfy**) achieves the best final BLEU score in every experiment we do. Also, the proposed Active IBT++ framework efficiently exploits the selected parallel corpus to further enhance the model accuracy. These techniques may also be useful for unsupervised NMT. Active pre-training is worth trying and active IBT has already proven its capability. We leave it for future work to study more acquisition functions in more NMT scenarios.

Acknowledgments

Yuekai Zhao and Zhihua Zhang have been supported by the Beijing Natural Science Foundation (Z190001), National Key Research and Development Project of China (No. 2018AAA0101004), and Beijing Academy of Artificial Intelligence (BAAI).

References

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

Tianchi Bi, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Multi-agent learning for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 856–865, Hong Kong, China. Association for Computational Linguistics.

Miguel Domingo, Mercedes García-Martínez, Amando Estela, Laurent Bié, Alexandre Helle, Álvaro Peris, Francisco Casacuberta, and Manuerl Herranz. 2019. Demonstration of a neural machine translation system with online learning for translators. *arXiv preprint arXiv:1906.09000*.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. *ArXiv*, abs/2004.03672.

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

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? *arXiv preprint arXiv:1804.05958*.

Julia Kreutzer and Stefan Riezler. 2019. Self-regulated interactive sequence-to-sequence learning. *arXiv preprint arXiv:1907.05190*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

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.

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

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.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2019. Data diversification: An elegant strategy for neural 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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.

Pavel Petrushkov, Shahram Khadivi, and Evgeny Matusov. 2018. Learning from chunk-based feedback in neural machine translation. *arXiv preprint arXiv:1806.07169*.

Minh Quang Pham, Josep M Crego, Jean Senellart, and François Yvon. 2018. Fixing translation divergences in parallel corpora for neural mt. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2967–2973.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Adaptation of machine translation models with back-translated data using transductive data selection methods. *arXiv preprint arXiv:1906.07808*.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4058–4068, Hong Kong, China. Association for Computational Linguistics.

Dongyu Ru, Yating Luo, Lin Qiu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2020. Active sentence learning by adversarial uncertainty sampling in discrete space. *ArXiv*, abs/2004.08046.

Dana Ruiter, Cristina Espana-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834.

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

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, Berlin, Germany. Association for Computational Linguistics.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, page 1070–1079, USA. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

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.

Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2018a. Dynamic sentence sampling for efficient training of neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 298–304.

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

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018b. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5823–5828, Florence, Italy. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *CoRR*, abs/1901.10430.

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

Ye Zhang, Matthew Lease, and Byron C Wallace. 2017. Active discriminative text representation learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.