A Longest and Shortest

Two other simple acquisition functions are favoring the longest or the shortest sentences, which proves to be a bad idea. We try these two acquisition functions in active NMT and find out that they underperform all other methods including the random selection baseline. These poor results may come from the deviation of the training corpus from the test corpus. Readers can refer to Figure 1 for some details. Due to their poor performance, we do not try longest or shortest on active NMT with transfer learning and active iterative back-translation.

B Non-Selective Iterative Back-Translation

We try non-selective iterative back-translation (IBT). The NMT models of different translation directions are pre-trained with 10% of the entire parallel corpus. The rest of the parallel corpus is used as the monolingual corpus to simulate IBT. Seven rounds of IBT training is done. In each IBT round, the NMT model of the opposite translation direction generates a synthetic parallel corpus that was merged into the 10% authentic parallel corpus. We train the NMT model for one epoch in each IBT round. Results are shown in Figure 2.

Due to the poor quality of the synthetic parallel corpus, the model performance of all translation directions suffers from IBT. This phenomenon indicates that synthetic parallel corpus needs to be chosen carefully. By using active acquisition functions to filter the synthetic parallel corpus, IBT can benefit model performance instead of hurting it.

C Analysis Results for All Acquisition Functions

Text analysis results are summarized in Table 1, Table 2 and Table 3. **Avg Len** means the average sentence length. **Total Cov** means the total vocabulary coverage and **Test Cov** means the test vocabulary coverage. **MTLD** scores are given in the last column of each table.

In general, the average sentence length given by each active acquisition function does not have much difference with the random selection baseline. However, when considering vocabulary coverage, random selection lags behind all other methods. A larger vocabulary means less unseen words for the NMT model, which may enhance the translation performance. As for the MTLD score, we use it to measure text diversity in the constructed parallel corpus. In most cases, **delfy** and **te-delfy** have the highest MTLD score, which may be the reason for their good performance. Note that **delfy** always has the highest vocabulary coverage, but it often underperforms **te-delfy**. Also, **te-delfy** tends to have better MTLD scores than **delfy** does. That is maybe because vocabulary coverage and text diversity are both important factors for designing a good acquisition function.

Method	Ave Len	Total Cov	Test Cov	MTLD
rand	24.9	32.9%	85.4%	172.5
lc	25.3	47.7%	87.8%	194.2
margin	23.1	49.9%	88.0%	194.8
te	24.0	51.9%	88.5%	190.7
tte	46.9	45.1%	87.5%	132.6
delfy	22.3	93.7%	91.6%	203.3
te-delfy	25.8	73.5%	90.4%	191.1

Table 1: Analysis Results for German.

Method	Ave Len	Total Cov	Test Cov	MTLD
rand	24.1	46.6%	93.4%	301.8
lc	24.9	65.9%	94.4%	299.7
margin	21.2	64.0%	94.3%	296.8
te	23.6	66.1%	94.6%	290.8
tte	47.3	57.0%	93.4%	174.1
delfy	23.1	99.7%	96.2%	306.1
te-delfy	25.3	82.4%	95.7%	339.9

Table 2: Analysis Results for Russian.

Method	Ave Len	Total Cov	Test Cov	MTLD
rand	21.5	60.6%	80.3%	373.3
lc	23.0	73.8%	82.7%	383.3
margin	20.7	72.6%	82.6%	390.2
te	22.4	73.9%	82.9%	379.1
tte	36.5	66.6%	81.7%	302.8
delfy	22.2	99.8%	84.8%	401.7
te-delfy	23.8	89.6%	84.2%	404.5

Table 3: Analysis Results for Lithuanian

D Experimental Detail

We give some experimental details in this section.

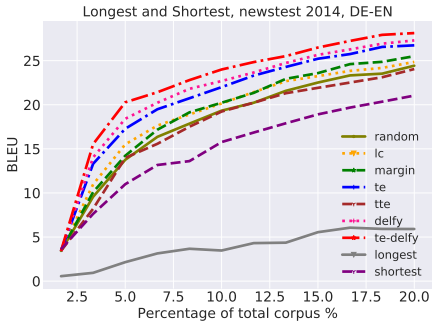
Reloading Optimizer States In each active training round, we get some new training corpus. Model parameters are reloaded without question, but whether should we reload optimizer states? We try reloading and not reloading optimizer states with random selection in active NMT. Results show that reloading optimizer states is always better.

Synthetic Sentence Length When evaluating uncertainty based acquisition functions as well as training with Active IBT or Active IBT++, we all need to generate synthetic translations. It is of critical importance to control the generated sentence length. Otherwise, the model performance will severely fluctuate. Assuming the original sentence length is l, the maximum generated sentence length is $1.3 * l + 5$.

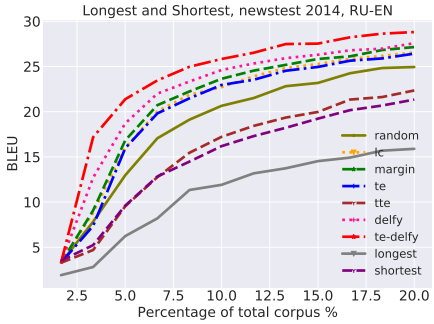
Greedy Decoding vs. Beam Search When calculating uncertainty based acquisition function scores, we use greedy decoding to generate synthetic translation. In Active IBT, we also generate synthetic training corpus by greedy decoding. Meanwhile, in Active IBT++, we use beam search to generate diversified translations for the final parallel corpus. We also use beam search when testing model performance. We use a beam size of 5 and a length penalty of 0.7.

Train Validation Splits 5000 sentence pairs are randomly sampled from the entire parallel corpus as the validation set for each language pair. The rest of the parallel corpus is used to simulate active NMT training.

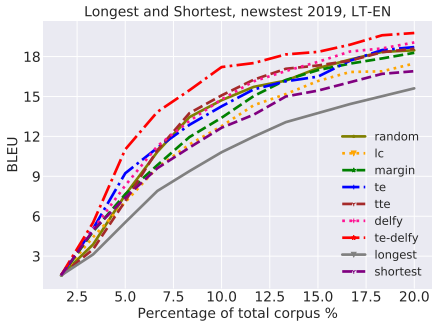
Runtime for each method All our experiments are done on 8 RTX 2080Ti GPU cards. For each acquisition function in active NMT and active NMT with transfer learning, the average runtime is 1.5 days for DE-EN and RU-EN, 0.5 day for LT-EN. For Active IBT, the average runtime is 2.5 days for DE-EN and RU-EN, 1 day for LT-EN.



(a) news test 2014, DE-EN

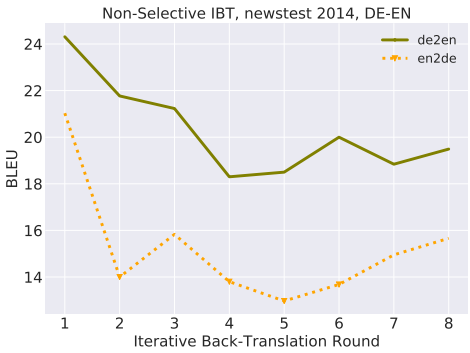


(b) news test 2014, RU-EN

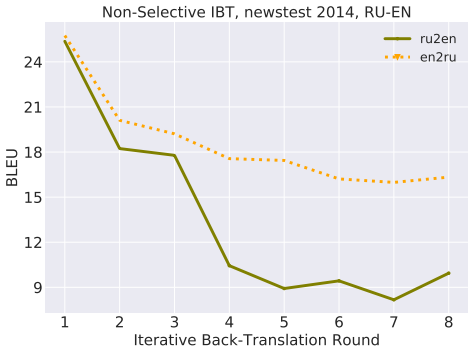


(c) news test 2019, LT-EN

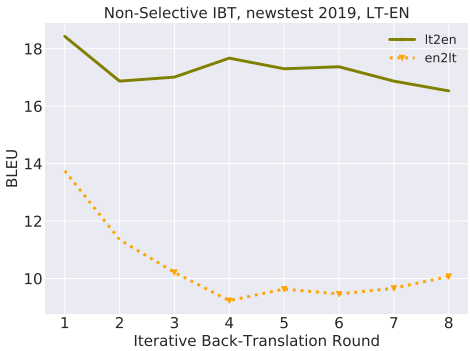
Figure 1: Longest and shortest compared with other acquisition functions.



(a) news test 2014, DE-EN



(b) news test 2014, RU-EN



(c) news test 2019, LT-EN

Figure 2: Non-selective iterative back-translation results.