Investigating Active Learning in Interactive Neural Machine Translation

Kamal Kumar Gupta, Dhanvanth Boppana, Rejwanul Haque,† Asif Ekbal and Pushpak Bhattacharyya

Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
†School of Computing, National College of Ireland, Dublin, Ireland
kamal.pcs17,boppana.cs17,asif,pb@iitp.ac.in
†rejwanul.haque@adaptcentre.ie

Abstract

Interactive-predictive translation is a collaborative iterative process, where human translators produce translations with the help of machine translation (MT) systems interactively. Various sampling techniques in active learning (AL) exist to update the neural MT (NMT) model in the interactive-predictive scenario. In this paper, we explore term based (named entity count (NEC)) and quality based (quality estimation (QE), sentence similarity (Sim)) sampling techniques – which are used to find the ideal candidates from the incoming data – for human supervision and MT model's weight updation. We carried out experiments with three language pairs, *viz.* German-English, Spanish-English and Hindi-English. Our proposed sampling technique yields 1.82, 0.77 and 0.81 BLEU points improvements for German-English, Spanish-English and Hindi-English, respectively, over random sampling based baseline. It also improves the present state-of-the-art by 0.35 and 0.12 BLEU points for German-English and Spanish-English, respectively. Human editing effort in terms of number-of-words-changed also improves by 5 and 4 points for German-English and Spanish-English, respectively, compared to the state-of-the-art.

1 Introduction

Neural machine translation (NMT) requires a significantly large amount of in-domain data for building the robust systems. Absence of sufficient training samples often result in the generation of erroneous output samples. Post-editing could be an effective solution in this situation, where human interference may help to rectify the errors in the output samples. However, there are two problems, *viz. (i)* post-editing a large number of output samples is time consuming and not very efficient in terms of productivity and *(ii)* not including all the post-edited examples might pose the risk of encountering the same mistakes in future. Hence, there is a necessity that instead of post-editing all the output samples, we explore effective sampling techniques for selecting important samples for post-editing, and further these post-edited samples are used to update the model's parameter following an active learning technique that makes the translation model learns from these (new) samples. This essentially increases ability of MT models to change in response to customer's data, which can counter the risk of encountering the same mistakes in future.

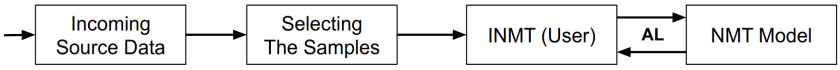


Figure 1: A pipeline showing the flow of data through sampling module, model updation through active learning.

Interactive MT (IMT) is viewed as an effective mean to increase the productivity in the translation industry. In principle, IMT aims to reduce human effort in automatic translation workflows by employing an iterative collaborative strategy with its two most important components, the human agent and the MT engine. As of today, NMT models (Bahdanau et al., 2015; Vaswani et al., 2017) represent state-of-the-art in MT research. This has led researchers to test interactive-predictive protocol on NMT too. Papers (Knowles and Koehn, 2016; Peris et al., 2017) that pursued this line of research suggest that NMT is superior than phrase-based statistical MT (Koehn et al., 2003). So use of interactive NMT (INMT) for output sample correction can significantly reduce the overall translation time and active learning strategy can use human corrected samples for adapting the underlying NMT model so that in future, the model does not repeat previous errors and improves the translation quality.

The contributions of our current work are stated as follows:

- We propose term based (NEC) and quality based (QE and Sim) sampling techniques that provide us with the ideal source samples which are first post-edited using interactive NMT (INMT) and then used to update the Transformer (Vaswani et al., 2017) based NMT model.

- With the help of the proposed sampling techniques, we significantly reduce human efforts in correcting the hypothesis in terms of token replacements using this proposed INMT model.

2 Related Work

In a case, where an MT model is not providing high quality translation due to low resource or out-of-domain scenarios, it could be beneficial to update the model with new samples while preserving the previous knowledge too. There has been some works which deal with the large input data streams but generally adopt the incremental learning approaches (e.g. updating the model as the labelled data become available) rather than the active learning approach (where labelled data stream is not guaranteed). In the literature (Levenberg et al., 2010; Denkowski et al., 2014), authors used incremental learning to update the translation model but these were with respect to the statistical machine translation (SMT) model. Turchi et al. (2017) applied incremental learning over the NMT model where they used the human post-edited data to update the initially trained models which make it very costly and time consuming due to human-edited data. Nepveu et al. (2004) and Ortiz-Martínez (2016) used an interactive paradigm for updating the SMT model on the iteratively corrected outputs.

As for active learning, it has also been well adopted for model learning. The unbounded and unlabelled large data streams is well suited to the objective of active learning (Olsson, 2009; Settles, 2009). This unbounded data stream scenario was explored by Haffari et al. (2009); Bloodgood and Callison-Burch (2010), where a pool of data was edited and the SMT model was updated using this data. González-Rubio et al. (2011) used the stream data to update the SMT model. Further, *interactive paradigm* of *SMT* was introduced in González-Rubio et al. (2012) and González-Rubio and Casacuberta (2014).

Source	aunque nunca jugué un juego de beber basado en el tema nazi .
Reference	never played a Nazi themed drinking game though .
Initial Hypothesis	never played a Nazi drinking play there .
Hypo-1	never played a Nazi themed play though .
Hypo-2	never played a Nazi themed drinking though .
Hypo-3	never played a Nazi themed drinking game though .

Table 1: Hypothesis correction and translation in INMT process. Here, **Hypo-** shows the step by step correction by user to achieve reference/desired sentence

Later, NMT became more prominent and efficient in the interactive paradigm of *MT* (Knowles and Koehn, 2016; Peris et al., 2017). Peris and Casacuberta (2018) explored the application of active learning and IMT on the NMT model. They performed the experiments over the attention based encoder-decoder NMT model (Bahdanau et al., 2015). To handle the incoming and unlabelled data stream, they introduced the sampling techniques which are majorly attention and alignment based. We explore the sampling criteria on the basis of lexical properties (term-based) and semantic properties (quality-based). We observe the impact of the proposed sampling techniques over the Transformer-based NMT.

3 Interactive Neural Machine Translation

In INMT (Knowles and Koehn, 2016; Peris et al., 2017), human translators correct errors in automatic translations in collaboration with the MT systems. Here, users read tokens of the generated hypothesis from left to right and modifies (insert/replace) his/her choice of words in the hypothesis generated by the NMT model. From the start index to the right most token position where the user make change is considered as the 'validated prefix'. After the user makes any change, the model regenerates a new hypothesis by preserving the validated prefix and new tokens next to it. Multiple attempts of token replacements may be required by a user to get the desired output as shown by an example in Table 1.

For an input-output sentence pair [x, y], where $x = (x_1, x_2, ..., x_m)$ being a sequence of input tokens and $y = (y_1, y_2, ..., y_n)$ being a sequence of output tokens, the probability of the ith translated word y_i is calculated as in (1):

$$p(y_i|y_1, ..., y_{i-1}, x) = f(y_{i-1}, s_i, c_i) \tag{1}$$

Here, s_i and c_i are the i^{th}decoder hidden state and context vector, respectively. As shown in Eq. (1), in NMT, during decoding, next predicted output y_i depends on model's previous output $y_1, ..., y_{i-1}$. In INMT, y_i will be generated by considering $y_1^*, ..., y_{i-1}^*$ as the previous tokens, where y_{i-1}^* is actually the token of user's choice at sequence position $i-1$. Eq. (2) shows the conditional probability of generating y_i in the INMT scenario.

$$p(y_i|y_1^*, ..., y_{i-1}^*, x) = f(y_{i-1}^*, s_i, c_i) \tag{2}$$

4 Sampling Techniques

From Figure 1, we see that the sampling module selects and recommends the incoming inference samples to the INMT for supervision. The purpose of a sampling technique is to filter out the ideal candidate from the incoming inference samples for which the trained NMT model is most uncertain and by supervising that sample it should increase the NMT performance using the technique of AL. Let *S* be the input sentences for inference, *B* be the block of sentences that are taken from *S* iteratively. From the block *B*, *C* a chunk, the size of which depends on the

	English-German	English-Spanish	English-Hindi
Train	1.26m (Europarl)	1.9m (Europarl)	1.6m (IITB corpus)
Dev	1,057 (Europarl)	2000 (Europarl)	599 (IITB corpus)
Testset	59,975 (newscommentary)	51,613 (newscommentary)	47,999 (ILCI corpus)

Table 2: Size of the corpora used for the experiments

percentage (%) of the samples from *B* are taken, is used to be supervised from the human. We take the size of *B* as 10,000 samples and the chunk size from *B* can be 20, 40, 60 and 80%. The amount of samples is measured by the count of sentence pairs. The sampling techniques which are implemented are pool based, and basically belong to two categories, namely uncertainty sampling (which labels those instances for which the model is least certain about the correct output to be generated) and query-by-committee (QbC) (where a variety of models are trained on the labeled data, and vote on the outputs of unlabeled data; label those instances for which the committee disagrees the most). Hence, the objective of the sampling techniques as mentioned below is to select from the unbounded data stream S, those sentences $S'(\subset S)$ which are worth to be used to update the parameters p of the *NMT* model.

4.1 Random Sampling (RS)

In RS, samples from the unlabelled block are taken without any criteria or uncertainty metric. Even though random sampling has no logically involved concept still it is expected to produce good and diverse samples from this sampling. We consider random sampling as the baseline for the proposed sampling techniques.

4.2 Quality Estimation (QE)

Quality estimation (QE) is the process of evaluating the MT outputs without using gold-standard references. This requires some kind of uncertainty measure which indicates the confidence that the model has in translating the sentences. It uses human translation edit rate (HTER) score evaluation metric. The HTER score is generally used to measure human effort in editing (insert/replace/delete) the generated hypothesis (Specia et al., 2018). We use this as a confidence score of the translation model. A translation with high HTER score can be viewed as a bad translation since it requires more human effort for editing, and a translation with low HTER score can be viewed as a good translation since it requires less human effort for editing. We performed QE sampling using the *Openkiwi* toolkit (Kepler et al., 2019). Openkiwi offers pre-trained QE models for English–German. We use one of the pre-trained models to obtain the HTER (uncertainty measure or score s_i) for every sentence S_i in S. In our case, high HTER score represents the sampling criteria. For each input sentence, this tool takes two inputs, i.e. source sentence and its translation generated by the initial NMT model, and gives us an estimated HTER score for the sentence. For a test sentence S_i in S where $(1 \leq i \leq |S|)$ ($|S| =$ number of sentences in S), quality estimation (QE) pre-trained model takes S_i and its generated translation T_i, and returns the corresponding HTER score HTER_i.

4.3 Sentence Similarity (SS)

Our second sampling strategy is based on sentence-similarity measure. We calculated similarity between a source sentence and its round-trip translation (RTT) (Moon et al., 2020). RTT, also known as back-and-forth translation, is the process of translating text into another language (forward translation), then translating the result back into the original language (back translation), using machine translation (MT) systems. Naturally, quality of a round-trip translation depends on two consecutive translation processes, i.e. forward translation by the source-to-target MT

system, and back-translation by the target-to-source MT system. When an NMT system is less confident about translating a source sentence (e.g. translating out-of-domain data), it is likely that a generated round-trip translation would not be closer to the source sentence. As for our *RTT*-based sampling setup, we had to prepare an additional MT system, i.e. back-translation MT system, and a low similarity score is regarded as the criteria for sampling. We calculated similarity between a source sentence and its RTT in two different ways: (i) semantic similarity and (ii) surface-level similarity of the source sentence and its RTT.

4.3.1 Similarity Based on Sentence Embeddings (Sim_{emb})

An RTT of a source sentence could be significantly different from the source sentence at lexical level; however, it could be semantically similar to the sentence, and this is not usually captured by surface-level similarity metrics such as BLEU (Papineni et al., 2002). We used semantic forms of the source sentence and its RTT in order to see how similar they are semantically. 'Similarity based on sentence embeddings' (Sim_{emb}) as the name itself suggests, this sampling technique uses a cosine similarity measure based on sentence embeddings. For each input sentence, two embeddings are generated: (i) embedding of the source sentence and (ii) embedding of the RTT of of the source sentence. These embeddings are generated using S-BERT[1] (Reimers and Gurevych, 2019). Sentences having the least similarity scores in the block are sampled and supervised by the user.

4.3.2 Similarity Based on Edit distance Between Sentences (Sim_{fuzzy})

This is a surface level similarity measure and it does not take into account the semantics of the source sentence and its RTT. In this sampling technique the similarity measure is based on 'levenshtein-distance' between the source sentences and their RTTs. In fact, for each sentence of test set the similarity score (Sim_{fuzzy}) between the sentence and its RTT is calculated using 'fuzzywuzzy' [2]. More specifically, this generates a score in the range of 0–100 (0 and 100 represent lowest and highest similarity scores, respectively). The sentences for which we obtained least similarity scores in the block are considered for supervision.

4.4 Named Entity Counting (NEC)

The NMT model suffers with the vocabulary restriction problem due to the limitation over the decoder side vocabulary size (Sennrich et al., 2016). Named entities (NEs) are open vocabularies and it is not possible for the NMT model to have all the NEs in the decoder vocabulary. Therefore, we considered presence of NEs as one of the sampling criteria. In other words, we took inability of the NMT model to translate the NEs perfectly into account for sampling. We count the NE tokens in each source sample of the incoming inference data and the sentences having the highest number of NE tokens in the block are considered as "difficult to translate" by the NMT model, and hence filtered for supervision. We use Spacy[3] named entity recognizer (NER) for marking NEs in sentences from English, German and Spanish languages.

4.5 Query-by-committee (QbC)

We combined opinions of random and our proposed sampling techniques to filter out the input samples for human supervision. Like Peris and Casacuberta (2018), we use a voted entropy function as in (3) to calculate the highest disagreement among the sampling techniques for a sample x. In (3), *#V(x)* is the number of sampling techniques voted for x to be supervised. *C*

[1] https://github.com/BinWang28/SBERT-WK-Sentence-Embedding
[2] https://github.com/seatgeek/fuzzywuzzy
[3] https://spacy.io/usage/linguistic-features#named-entities

denotes the number of all the sampling techniques participating in the voting process.

$$C_{QbC}(x) = \frac{-\#V(x)}{|C|} + log\frac{\#V(x)}{|C|} \tag{3}$$

4.6 Attention Distraction Sampling (ADS)

Attention distraction sampling (ADS) was proposed by Peris and Casacuberta (2018). Attention network of NMT (Bahdanau et al., 2015) distributes weights over tokens of source sentence based on their contribution in generating every target token. If the system finds the translation of a sample uncertain then the attention probability distribution follows the uniform distribution. It shows that NMT model is having difficulty in distributing weights over the source tokens based on their contribution in target generation. The samples having highest distraction are selected for active learning. The kurtosis of weights given by the attention model while generating y_i is calculated to measure the attention distraction, as in (4):

$$Kurt(y_i) = \frac{\frac{1}{|x|}\sum_{j=1}^{|x|}(\alpha_{i,j} - \frac{1}{|x|})^4}{(\frac{1}{|x|}\sum_{j=1}^{|x|}(\alpha_{i,j} - \frac{1}{|x|})^2)^2} \tag{4}$$

where $\alpha_{i,j}$ is the attention weight between the j-th source word and i-th target word. Note that, fraction $\frac{1}{|x|}$is equivalent to the mean of the attention weights of the word y_i. Finally, The kurtosis values for all the target words are used to obtain the attention distraction score.

5 Dataset

We carried out our experiments on three language pairs using three benchmark datasets. Table 2 shows the statistics of training, development and test sets used for our experiments. In order to measure performance of the proposed sampling techniques, we use different domain datasets for training and testing. For German–English and Spanish–English, we use Europarl corpus (Koehn, 2005) for training and News-Commentary (NC) corpus for testing. This gives us a clear indication whether the translation models trained over Europarl corpus are able to adapt over the sampled examples from NC corpus using active learning. Likewise, for English–Hindi translation, we use the IITB corpus (Kunchukuttan et al., 2018) for training which is a combination of sentences from government sites, TED talks, administration books etc. As for evaluation, we use the ILCI corpus (Jha, 2010) which is a combination of sentences from the health and tourism domain.

6 Experimental Setup

Our MT systems are Transformer models (Vaswani et al., 2017). We used 6 layered Encoder-Decoder stacks with 8 attention heads. Embedding size and hidden sizes were set to 512, dropout rate was set to 0.1. Feed-forward layer consists of 2,048 cells. Adam optimizer (Kingma and Ba, 2015) was used for training with 8,000 warm up steps. We used BPE (Sennrich et al., 2016) with a vocabulary size of 40K. Models were trained with OpenNMT toolkit[4] (Klein et al., 2020) with batch size of 2,048 tokens till convergence and checkpoints were created after every 10,000 steps. During inference, beam size is set to 5. We measured BLEU (calculated with *multi-bleu.pl* script) of the trained models on the test sets.

7 Results and Analysis

We evaluate the impact of the proposed sampling techniques for active learning in NMT in two different ways. Firstly, we test whether the proposed techniques help the NMT model to

[4]https://opennmt.net/

En-to-De	20%	40%	60%	80%
Random	23.88	24.26	24.67	25.31
ADS	24.36	25.69	26.24	26.78
Quality estimation	24.02	24.98	25.61	26.17
Fuzzy	24.55	25.66	26.21	26.68
Sentence Similarity	24.35	25.73	26.47	26.9
NE Counting	25.22	26.14	26.31	26.84
QbC	25.51	26.08	26.69	27.13

En-to-Hi	20%	40%	60%	80%
Random	25.84	26.08	26.41	26.83
ADS	25.90	26.81	27.1	27.58
Fuzzy	25.97	26.67	27.03	27.52
Sentence Similarity	25.88	26.44	26.91	27.28
NE Counting	25.92	26.75	27.2	27.64
QbC	26.18	26.87	27.15	27.42

De-to-En	20%	40%	60%	80%
Random	25.19	26.32	27.11	27.05
ADS	25.80	26.58	27.39	27.98
Fuzzy	25.98	26.64	27.29	27.85
Sentence Similarity	26.18	26.73	27.52	28.11
NE Counting	25.50	26.38	27.26	27.48
QbC	26.53	26.83	27.62	28.13

Es-to-En	20%	40%	60%	80%
Random	39.16	39.52	40.19	40.87
ADS	39.50	39.85	40.51	41.52
Fuzzy	39.28	40.25	40.85	41.27
Sentence Similarity	39.74	39.91	40.75	41.64
NE Counting	39.43	39.74	40.36	41.38
QbC	39.78	40.26	40.97	41.68

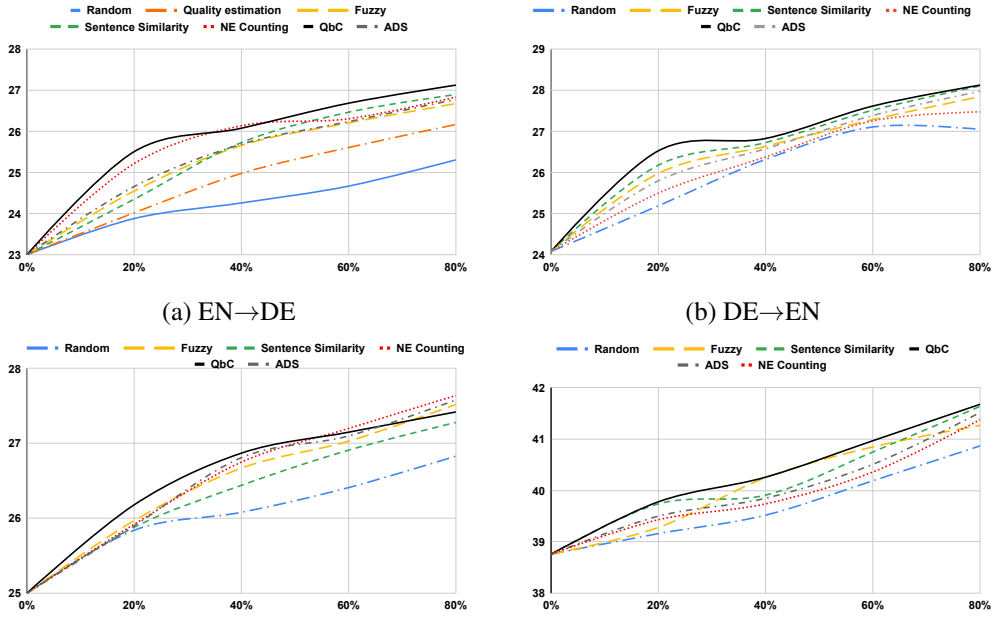
Table 3: BLEU scores of the hypothesis generated by NMT model based on samples selected by different sampling techniques and % of data used to adapt it. For each translation direction, the initial BLEU score before applying the sampling techniques is: *En-to-De:* 23.28, *De-to-En:* 24.08, *En-to-Hi:* 25.76 and *Es-to-En:* 38.76

improve its translation performance in terms of the BLEU score. Secondly, in order to see whether the proposed techniques are able to reduce the human efforts (number of token correction required) in correcting the hypothesis, we compare the performance of the proposed sampling techniques with the baseline i.e random sampling and the state-of-the-art sampling *i.e. attention distraction sampling (ADS)* (Peris and Casacuberta, 2018) methods.

7.1 Effect on Translation Quality

We considered the random sampling-based method as our baseline model. By increasing the amount of the samples of the block to be supervised recommended by the proposed sampling techniques with 20, 40, 60 and 80%, we recorded changes in the BLEU scores. The BLEU scores presented are calculated based on a single block of 10,000 sentences. Table 3 shows the BLEU scores for different translation directions. We also present charts (see Figure 2) to illustrate the effect of the sampling techniques on the translation quality of the NMT models for the specific translation directions using AL. As can be seen from Figure 2, for English-to-German translation, the initial BLEU score of the trained NMT model before employing active learning was 23.28. When we adapt the trained NMT model to the new samples recommended by the random sampling, the BLEU score increases to 25.31 (when 80% of the samples of block are supervised) which is a 2.03 BLEU points improvement over the initial score. When we compare our proposed sampling methods with the random sampling-based method, we see that QE, Sim_{emb}, Sim_{fuzzy} and NEC brought about 26.17, 26.90, 26.68 and 26.84 BLEU points, respectively, on the test set. Note that these scores were obtained when we used 80% of the samples of the block for supervision. Interestingly, we can see from the figure that Sim_{emb} is the best performing method and provides us 26.90 BLEU points on the test set, which is 1.59 BLEU more than one that we obtained with the random sampling method (baseline). We also tested a combined opinion of sampling techniques (i.e. QbC) and it outperformed the other methods and produced 27.13 BLEU points, which is a 1.82 BLEU improvement over the one that we obtained after applying the random sampling method.

For the German-to-English translation task we obtained 24.08 BLEU points on the test set for the initial setup, i.e. without applying active learning. The baseline INMT system (i.e. based on random sampling method) brought about 27.05 BLEU points on the test set. The INMT system with sentence-similarity sampling feature (i.e. Sim_{emb}) surpassed the baseline



(a) EN→DE (b) DE→EN

(c) EN→HI (d) ES→EN

Figure 2: Presenting the BLEU score improvements of NMT model based on the new learned samples chosen by different sampling techniques and data size used to adapt it.

by 0.94 BLEU points. Furthermore, the QbC method outperforms all the other sampling methods, and with this, we achieve 28.13 BLEU points (an improvement of 1.08 BLEU points over the baseline (i.e. random sampling technique)) on the test set.

As for the English-to-Hindi translation task, the initial BLEU score on the test set was 25.76. In this translation task, NEC was found to be the best performing sampling method. The INMT setup with this sampling method statistically significantly outperforms the baseline INMT system (built on the random sampling method), and we obtain an improvement of 0.81 BLEU points over the baseline. The statistical significance test is performed using the bootstrap resampling method (Koehn, 2004).

In the Spanish-to-English translation task, for the initial setup, we obtained 38.76 BLEU points on the test set. The baseline sampling strategy provided us with 40.87 BLEU points. As in English-to-German, QbC is found to be the best performing sampling method, and provided us a gain of 0.81 BLEU points over the baseline. When we compare Sim_{emb} and QbC, we see that they are comparable as far as BLEU scores are concerned.

Furthermore, in Figure 2, we demonstrate the performance of different sampling techniques in AL (active learning) for the German-to-English, English-to-German, English-to-Hindi and Spanish-to-English translation tasks. The x-axis of the graphs in Figure 2 represents the amount (%) of the samples supervised in the block and the y-axis represents the BLEU scores. For English-to-Hindi, the baseline INMT model (i.e. random sampling) produces 26.83 BLEU points on the test set, which corresponds to an absolute improvement of 1.07 BLEU points over the vanilla NMT system (i.e. 25.76 BLEU points). NEC is found to be the best-performing sampling technique, and yields 27.64 BLEU points with an absolute improvement of 0.82 BLEU points over the baseline (random sampling).

As for Spanish-to-English translation, we see that Sim_{emb} significantly outperforms the

	Random Sampling	QbC
En-to-De	52.06	57.73
De-to-En	45.60	50.45
En-to-Hi	37.82	46.14
Es-to-En	49.37	53.61

Table 4: Word prediction accuracy (WPA) of the NMT models for different translation directions with 80% samples supervised.

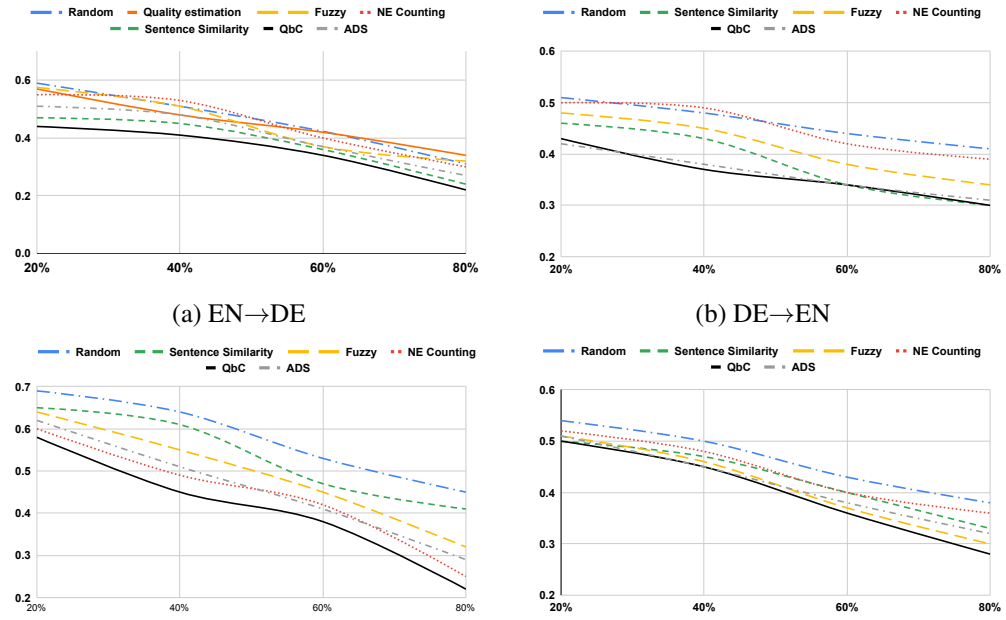
random sampling by 0.77 BLEU points. Furthermore, for English-to-German, English-to-Hindi and Spanish-to-English, the respective best-performing sampling techniques, which are our proposed methods, bring about gains over ADS (Peris and Casacuberta, 2018) by 0.35, 0.06 and 0.12 BLEU scores. These improvements are very small and except English-to-German, the remaining two improvements are not statistically significant. However, in the next section, we will see that our proposed sampling techniques outperform ADS significantly in terms of human effort reduction.

7.2 Effect on Human Effort

We wanted to see whether the proposed sampling techniques in AL are helpful in reducing human effort in correcting a translation. Human translator is provided with the IMT system to correct the model generated hypothesis and calculate the effort in correcting it. AL based on different sampling techniques is applied over the IMT to make a comparison among the effect of sampling techniques in human effort reduction. Since it is quite expensive to use human translators in INMT performance evaluation, we measured human effort in a reference-simulated environment, where the reference sentences are considered as the user's choice of sentences. The idea is to correct the hypothesis until it matches the reference sentence. Using different sampling techniques, we aimed at improving the translation quality of the NMT system. We recorded performance of the INMT system in terms of the model's ability to predict the next word at decoding. Every time the *user modified hypothesis* is fed to the NMT model, the model predicts next correct token based on the modifications made by the user. We calculate the model's accuracy in predicting the next words using a commonly-used metric: word prediction accuracy (WPA) metric. WPA is the ratio of the number of correct tokens predicted and the total number of tokens in the reference sentences (Peris et al., 2017). Higher the WPA scores of the NMT model means the lesser human efforts in correcting the hypothesis. We also calculated human efforts using another metric: word stroke ratio (WSR). WSR is the ratio of the number of tokens corrected by the user and the total number of tokens present in the reference sentences (Knowles and Koehn, 2016). In our case, we investigated whether the proposed sampling techniques are able to reduce human efforts in translation (i.e. lower WSR and higher WPA scores are better).

Table 4 shows WPA scores of our INMT systems in different translation tasks. Here, we show the WPA scores only when 80% of the samples in the block are supervised. We considered random sampling as the baseline and compared it with the QbC since we found that it is the best performing approach out of all proposed sampling techniques (i.e. Sim, NEC, Fuzzy) as far as WPA is concerned. In sum, the interactive-predictive translation setup with QbC surpassed the baseline setup by 5.67%, 4.85%, 8.32% and 4.24% accuracy in terms of WPA for the English-to-German, German-to-English, English-to-Hindi and Spanish-to-English translation tasks, respectively.

In Figure 3, we show WSR scores obtained by the different sampling techniques. As above, we considered varying sizes of samples for supervision, i.e. 20, 40, 60 and 80% of



(a) EN→DE

(b) DE→EN

(c) EN→HI

(d) ES→EN

Figure 3: Human effort reduction in terms of token replacement in Interactive NMT

the samples are supervised in a block. We calculate the fraction of tokens replaced in the hypothesis correction. These hypothesises are generated by the NMT models adapted over the samples recommended by the different sampling techniques. The x-axis of the graphs shows the % of samples supervised and y-axis shows the average number of tokens replaced. As can be seen from the graphs, for English-to-German translation, QbC achieves statistically significantly absolute improvement of 1.82 BLEU points over the baseline. As for English-to-Hindi and Spanish-to-English, NEC and Sim_{emb} yield 0.81 and 0.77 BLEU improvements over the baseline. We also observed the reduction of human efforts in terms of word stroke ratio (WSR). For English-to-German, English-to-Hindi and Spanish-to-English, we achieve a reduction in WSR of 9%, 23% and 10% over the baseline. We also present the scores that were shown in graphs in Table 3. We see that for English-to-German translation, QbC is the best-performing approach in terms of WSR. For German-to-English, QbC and Sim_{emb} are found to be the best-performing strategies. For English-to-Hindi and Spanish-to-English, along with the QbC, the second best-performing sampling techniques are NEC and Sim_{emb}, respectively. Unlike German-to-English and Spanish-to-English, for English-to-Hindi, Sim_{emb} is not the best-performing method. We observed that there may be some reasons for this: (i) morphological richness of Hindi, and (ii) syntactic divergence of English and Hindi languages. These might introduce more challenges in RTT in case of Sim_{emb}. We also compared the amount of human effort reduction by the proposed techniques and ADS. For English-to-German, English-to-Hindi and Spanish-to-English translation, we observed the reduction in WSR by 5.13, 6.72 and 4.38 points, respectively, over the ADS.

8 Conclusion

In this paper, we have explored the applicability of various sampling techniques in active learning to update NMT models. We selected incoming source samples using different sampling techniques, translate them using a NMT model, corrected them via interactive NMT protocol, and subsequently updated the NMT model using the corrected parallel samples. It helped the MT model to adapt over the new parallel samples which results in improving the translation quality and reducing the human effort for further hypothesis correction. We proposed term based (NEC) and quality based (QE, Sim_{emb}, Sim_{fuzzy}) sampling techniques to pick the source samples from a large block of input sentences for correction and subsequently updating the NMT models. Since it is not feasible for a human to supervise (modify) a large set of input data coming for the translation, the proposed sampling techniques help to pick and recommend the suitable samples from large input data to the user for supervision. We measure the impact of sampling techniques by two criteria: *first*, improvement in translation quality in terms of BLEU score and *second*, reduction in human effort (i.e. number of tokens in generated outputs needed to correct).

We performed experiments over three language pairs i.e. English-German, English-Spanish and English-Hindi. We use different domain data for training and testing the NMT model to see if the NMT model trained over the data from one domain can successfully adapt to the different domain data. We empirically showed that the proposed term and quality based sampling techniques outperform the random sampling and outperformed the *attention distraction sampling (ADS)* method

9 Acknowledgement

The research reported in this paper is an outcome of the generous support received from the project "Hindi to English Machine Aided Translation for the Judicial Domain (HEMAT)", sponsored by the Technology Development in Indian Language (TDIL), MeiTy, Government of India.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA.

Bloodgood, M. and Callison-Burch, C. (2010). Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 854–864, Uppsala, Sweden. Association for Computational Linguistics.

Denkowski, M., Dyer, C., and Lavie, A. (2014). Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 395–404, Gothenburg, Sweden. Association for Computational Linguistics.

González-Rubio, J. and Casacuberta, F. (2014). Cost-sensitive active learning for computer-assisted translation. *Pattern Recognition Letters*, 37:124–134.

González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. (2011). An active learning scenario for interactive machine translation. In *Proceedings of the 13th international conference on multimodal interfaces - ICMI '11*. ACM Press.

González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. (2012). Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254, Avignon, France. Association for Computational Linguistics.

Haffari, G., Roy, M., and Sarkar, A. (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 on - NAACL '09*. Association for Computational Linguistics.

Jha, G. N. (2010). The tdil program and the indian langauge corpora intitiative (ilci). In *LREC*.

Kepler, F., Trénous, J., Treviso, M., Vera, M., and Martins, A. F. T. (2019). OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klein, G., Hernandez, F., Nguyen, V., and Senellart, J. (2020). The opennmt neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 102–109.

Knowles, R. and Koehn, P. (2016). Neural interactive translation prediction. In *Proceedings of AMTA 2016, vol. 1: MT Researchers' Track*, pages 107–120. Association for Machine Translation in the Americas, AMTA. Twelfth Conference of The Association for Machine Translation in the Americas, AMTA 2016 ; Conference date: 28-10-2016 Through 01-11-2016.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, pages 48–54, Edmonton, AB.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Levenberg, A., Callison-burch, C., and Osborne, M. (2010). Stream-based translation models for statistical machine translation. In *In Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010*.

Moon, J., Cho, H., and Park, E. L. (2020). Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Nepveu, L., Lapalme, G., and Foster, G. (2004). Adaptive language and translation models for interactive machine translation. In *In Proc. of EMNLP*, pages 190–197.

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing.

Ortiz-Martínez, D. (2016). Online learning for statistical machine translation. *Computational Linguistics*, 42(1):121–161.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Peris, Á. and Casacuberta, F. (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.

Peris, Á., Domingo, M., and Casacuberta, F. (2017). Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). 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.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Specia, L., Scarton, C., and Paetzold, G. H. (2018). Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Turchi, M., Negri, M., Farajian, M. A., and Federico, M. (2017). Continuous learning from human post-edits for neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):233–244.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.