Phrase-level Active Learning for Neural Machine Translation

Junjie Hu[*]
University of Wisconsin-Madison
junjie.hu@wisc.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

Abstract

Neural machine translation (NMT) is sensitive to domain shift. In this paper, we address this problem in an active learning setting where we can spend a given budget on translating in-domain data, and gradually fine-tune a pre-trained out-of-domain NMT model on the newly translated data. Existing active learning methods for NMT usually select sentences based on uncertainty scores, but these methods require costly translation of full sentences even when only one or two key phrases within the sentence are informative. To address this limitation, we re-examine previous work from the phrase-based machine translation (PBMT) era that selected not full sentences, but rather individual phrases. However, while incorporating these phrases into PBMT systems was relatively simple, it is less trivial for NMT systems, which need to be trained on full sequences to capture larger structural properties of sentences unique to the new domain. To overcome these hurdles, we propose to select *both* full sentences and individual phrases from unlabelled data in the new domain for routing to human translators. In a German-English translation task, our active learning approach achieves consistent improvements over uncertainty-based sentence selection methods, improving up to 1.2 BLEU score over strong active learning baselines.[1]

1 Introduction

Machine translation (MT) models are very sensitive to domain shift (Koehn and Knowles, 2017; Chu and Wang, 2018), and one typical way to address this problem is adding in-domain data to the MT training process (Luong and Manning, 2015; Chu et al., 2017). However, this data may not be available *a priori*, and hiring professional translators with knowledge of specific domains (such as medicine or law) is usually costly.

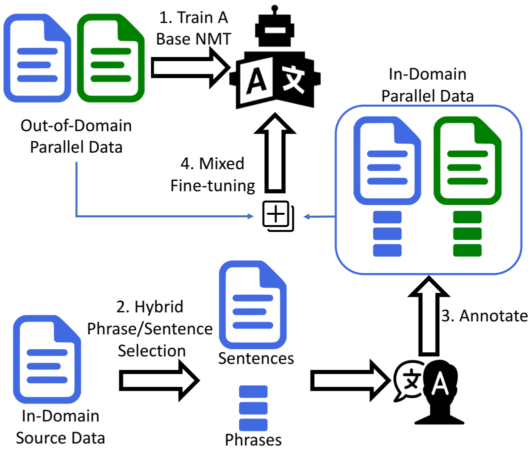


Figure 1: Overview of the active learning process

As a result, active learning approaches (Gangadharaiah et al., 2009; Haffari et al., 2009; Bloodgood and Callison-Burch, 2010) have been widely adopted to reduce the annotation cost by translating a smaller representative subset of the in-domain data, with the hope that models trained on this translated subset approximate those trained on a much larger labeled set. In general, active learning (AL) approaches iterate between two steps: *data selection/annotation*, and *model update*. With regards to data selection for machine translation, most existing works (Haffari et al., 2009; Peris and Casacuberta, 2018; Zeng et al., 2019) focus on selecting *sentences* that are most useful for training either phrase-based machine translation (PBMT) or neural machine translation (NMT) models.

However, even the most informative sentences inevitably involve segments that the MT system can already translate well, and asking the translator to also translate these segments is not cost-effective. There have been a few works used in conjunction with older PBMT models that ameliorate this problem through phrase-based selection techniques (Bloodgood and Callison-Burch, 2010; Daumé III and Jagarlamudi, 2011; Miura et al.,

[*] Work done at Carnegie Mellon University

[1] Code/data is released at https://github.com/JunjieHu/phrase-al-nmt.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 1087–1099
November 10–11, 2021. ©2021 Association for Computational Linguistics

2016), which select only *individual phrases*, maximizing information gain. However, while these translated phrases can be easily integrated into PBMT by adding them to the existing phrase table, incorporating them into NMT models is less simple because NMT has no concept of a "phrase table" and must be trained on full sentences similar to those that must be translated.

In this paper, we propose a method for incorporating phrase-based active learning into NMT. Specifically, we first describe sentence-based and phrase-based selection strategies, then propose a hybrid strategy that combines both methods. We also describe several ways to incorporate this translated data into the training of NMT systems. We conduct experiments on German-English translation by adapting NMT models trained on WMT parallel data to the medicine and IT domains. Experimental results show that the hybrid selection strategy obtains more stable translation performance than either phrase-based or sentence-based selection strategy.

2 Problem Definition

In the setting of active learning for domain adaptation, we are given an out-of-domain labelled corpus $(x, y) \in \mathcal{L}$ and an in-domain unlabelled corpus $x \in \mathcal{U}$. We define a phrase as a contiguous sequence of words up to some length limit N, and denote a set of possible phrases in a sentence x by $\cup_{n\in[1,N]} n\text{-gram}(x)$, where we set $N = 4$ in all experiments below. To obtain translations of unlabelled data, we assume access to professional translators $\mathcal{O}(\cdot)$ who can translate source-side sentences $\mathcal{S}$ and/or phrases $\mathcal{P}$ selected from $\mathcal{U}$, i.e., $\mathcal{O}(x)\ \forall x \in \mathcal{S} \subset \mathcal{U}$, and $\mathcal{O}(p)\ \forall p \in \mathcal{P} \subset \mathcal{P}_{\mathcal{U}} = \cup_{x\in\mathcal{U}} \cup_{n\in[1,N]} n\text{-gram}(x)$. We assume that translating sentences or phrases requires cost $c(\cdot)$, and annotation must be performed within a fixed budget $B = \sum_{x\in\mathcal{S}} c(x) + \sum_{p\in\mathcal{P}} c(p)$. This active learning procedure consists of two main steps: selection/translation (§3) and fine-tuning (§4).

3 Active Selection Strategies

3.1 Sentence Selection Strategies

Existing sentence-based active learning methods usually define a sentence-level scoring function $\phi(x, \cdot)$, and select sentences with the top scores. Following Zeng et al. (2019), we categorize these methods into two classes: data-driven and model-driven methods. Data-driven methods only rely on the unlabeled data $\mathcal{U}$ and the labeled data $\mathcal{L}$, i.e., $\phi(x, \mathcal{U}, \mathcal{L})$, and usually score sentences based on the trade-off between the density and diversity of the selected sentences. In contrast, model-driven approaches usually estimate the prediction uncertainty of a source sentence given the current MT model θ, i.e., $\phi(x, \theta, \mathcal{U}, \mathcal{L})$, and select sentences with high uncertainty for training the model. Before getting to our proposed phrase-based strategies in §3.2 we highlight several existing sentence selection strategies.

Random Sampling: One easy strategy is randomly sampling sentences from the unlabeled data $\mathcal{U}$ for annotation. Although it is simple, this method is an unbiased approximation of the data distribution in $\mathcal{U}$. Therefore, this method remains a strong baseline in the active learning literature (Gangadharaiah et al., 2009; Miura et al., 2016; Zeng et al., 2019) if the annotation budget is sufficiently large.

Margin-based Ratio Score (MRS): Zhang et al. (2018) propose to measure the distance between sentence embeddings. This method takes each unlabeled sentence, estimates its distance in embedding space from the labeled sentences in the out-of-domain corpus, and iteratively selects sentences that are more distant from sentences in the labeled data. In our instantiation of this method, we leverage the pre-trained mBERT model (Devlin et al., 2019) to extract sentence representation $\mathbf{e}_x$ of a particular sentence x.[2] Instead of using a cosine similarity function, we measure a ratio-based score which is the ratio between the cosine similarity of $(\mathbf{e}_x, \mathbf{e}_{x'})$ and the average cosine similarity with their k nearest neighbors in Eq. (1), because the margin-based ratio score has been shown effective in sentence retrieval in (Artetxe and Schwenk, 2019).

$$\begin{aligned} &\text{ratio}(\mathbf{e}_x, \mathbf{e}_{x'}) \qquad (1)\\ &= \frac{\cos(\mathbf{e}_x, \mathbf{e}_{x'})}{\sum\limits_{z\in \text{NN}_k(x)} \frac{\cos(\mathbf{e}_x, \mathbf{e}_z)}{2k} + \sum\limits_{z\in \text{NN}_k(x')} \frac{\cos(\mathbf{e}_{x'}, \mathbf{e}_z)}{2k}}, \end{aligned}$$

where k is the number of nearest neighbors.

We then compute the distance between each in-domain sentence and its nearest out-of-domain

[2] We average the word representations from the 7th layer of the mBERT model as the sentence embedding, because the middle-layer representations have proven effective in cross-lingual retrieval tasks (Pires et al., 2019; Hu et al., 2020).

neighbor within a randomly sampled subset of labeled sentences $\mathcal{L}'$:

$$\phi(x,\cdot) = \text{dist}(x,\mathcal{L}') = 1 - \max_{x' \in \mathcal{L}'} \text{ratio}(\mathbf{e}_x, \mathbf{e}_{x'}).$$

We approximate the distance between x and out-of-domain corpus $\mathcal{L}$ using a subset $\mathcal{L}'$ for efficiency purposes, because the out-of-domain $\mathcal{L}$ is usually large. Next we use the distance $\text{dist}(x,\mathcal{L}')$ as our scoring function $\phi(x,\cdot)$, and select the unlabeled sentence with the largest distance from (subsampled) sentences in the out-of-domain corpus.

Round Trip Translation Likelihood (RTTL): One model-driven method is based on a method referred to as "round trip translation" (Haffari et al., 2009; Zeng et al., 2019). The labeled data $\mathcal{L}$ is used to train two MT models $\theta_{\text{src-tgt}}, \theta_{\text{tgt-src}}$ that translate between the source and target languages in two directions. Each unlabeled source sentence $x \in \mathcal{U}$ is first translated to $\hat{y}$ in the target language by $\theta_{\text{src-tgt}}$, and then $\hat{y}$ is translated to $\hat{x}$ by $\theta_{\text{tgt-src}}$. This method assumes that if this round-trip translation process fails to recover some of the content on the source side then this is an indication that the sentence may be difficult for the current model and is a good candidate for human annotation. Haffari et al. (2009) use a scoring function that computes the similarity between the original sentence x and $\hat{x}$ using the sentence-level BLEU score (Chen and Cherry, 2014), while Zeng et al. (2019) estimate the likelihood of the original source sentence x given $\hat{y}$ by the reverse MT model $\theta_{\text{tgt-src}}$.

$$\hat{y} \approx \operatorname*{argmax}_y P_{\theta_{\text{src-tgt}}}(y|x) \tag{2}$$

$$\phi(x,\cdot) = \log P_{\theta_{\text{tgt-src}}}(x|\hat{y}) \tag{3}$$

3.2 Phrase Selection Strategies

A few existing phrase-based active learning methods (Bloodgood and Callison-Burch, 2010; Miura et al., 2016) have been proposed to improve PBMT systems. These methods first determine the possible set of phrases in a sentence, select phrases to be translated according to a scoring metric, and incorporate these in the training of the PBMT system. In the following paragraphs, we introduce two phrase-based selection strategies, and discuss how to integrate this data into NMT in §4. Similar to the sentence selection strategies, we define a phrase-level scoring function $\phi(p,\cdot)$ and select phrases with the top scores.

n-gram Frequency (NGF) (Bloodgood and Callison-Burch, 2010): The most straightforward phrase selection strategy is to select the most frequent phrases in the unlabelled data that *do not* appear in the already labeled data. First we extract two sets of possible n-grams ($n \leq 4$) from sentences in $\mathcal{U}$ and $\mathcal{L}$, which are defined as $\mathcal{P}_\mathcal{U} = \cup_{x \in \mathcal{U}} \cup_{n \in [1,N]} n\text{-gram}(x)$, and $\mathcal{P}_\mathcal{L} = \cup_{(x,y) \in \mathcal{L}} \cup_{n \in [1,N]} n\text{-gram}(x)$. We then score each phrase as follows:

$$\phi(p,\cdot) = \begin{cases} occ(p,\mathcal{U}), & \text{if } p \in \mathcal{P}_\mathcal{U}, p \notin \mathcal{P}_\mathcal{L} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\text{occ}(p)$ counts the occurrences of p in $\mathcal{U}$. We then select the top frequent phrases until we use up the budget for annotating phrases.

Semi-Maximal Phrases (NGF-SMP): The two phrase sets $\mathcal{P}_\mathcal{U}, \mathcal{P}_\mathcal{L}$ extracted by the n-gram Frequency method contain many substrings that also occur in some longer strings. For example, p = "eines der" always co-occurs with the longer p' = "eines der besten" in the WMT14 German-English dataset. To identify the longer strings, Miura et al. (2016) proposed the following semi-order relation, which defines the relation between a phrase p' and its sub-string p satisfying the condition that p' occurs at least half the time of p in the corpus $\mathcal{U}$.

$$\begin{aligned} p \overset{*}{\preceq} p' \Leftrightarrow \exists \alpha, \beta :\ \alpha p \beta = p' \qquad\qquad & \qquad (5) \\ \wedge \frac{\text{occ}(p,\mathcal{U})}{2} < \text{occ}(p',\mathcal{U}) & \end{aligned}$$

A phrase p is called a semi-maximal phrase if there does not exist a phrase p' in $\mathcal{U}$ such that $p \overset{*}{\preceq} p'$. Therefore, a compact subset of phrases $\mathcal{P}'_\mathcal{U}$ can be constructed by containing only semi-maximal phrases in the phrase set $\mathcal{P}_\mathcal{U}$ in $\mathcal{U}$:

$$\mathcal{P}'_\mathcal{U} = \{p | \nexists p' \in \mathcal{P}_\mathcal{U}, p \overset{*}{\preceq} p' \wedge p \in \mathcal{P}_\mathcal{U}\}. \tag{6}$$

By using semi-maximal phrases in $\mathcal{P}'_\mathcal{U}$ rather than all phrases in $\mathcal{P}_\mathcal{U}$, we remove a large number of phrases that are included in a longer phrase more than half the time, and reduce the redundancy of the selected phrases. Next we can select phrases similarly using Eq. (4) by replacing the original phrase set $\mathcal{P}_\mathcal{U}$ with the sub-set $\mathcal{P}'_\mathcal{U}$.

Notably, we select representative phrases by their occurrences instead of using a similarity function between phrase embeddings. Because it is easy to count the phrase occurrences by extract string-match while it is infeasible to do so for sentences.

As for sentence selection, measuring a similarity between sentence embeddings (e.g., MRS) provides an alternative way of matching sentences.

3.3 Hybrid Selection Strategy

Phrase-based selection has its benefits, such as efficient annotation of core vocabulary from the target domain. However, at the same time it lacks the ability to identify larger sentence structure that may nonetheless be unique to the target domain. Modeling this structure is particularly important for NMT (in constrast to PBMT), as NMT directly learns both lexical and syntactic transformations within the same model.

Because of this, we propose a simple yet novel hybrid selection strategy that leverages the benefits of both sentence-based and phrase-based selection strategies. Specifically, we allocate our budget in a way to annotate sentences with B_s words from our set of sentences and B_p words from our set of phrases. Depending on the specific sentence-based and phrase-based selection strategies chosen in the hybrid selection strategy, it is non-trivial to determine which selection strategy improves the in-domain translation performance more than the other one before actual finetuning. Therefore, in our implementation, we assume that we have no prior knowledge about which selection strategies will be most effective, and simply evenly distribute the annotation budget into the sentence-based and phrase-based strategies. We leave more sophisticated allocation strategies as future work, and we discuss some potential avenues briefly in §7.

4 Training with Sentences and Phrases

After data selection, we fine-tune the base NMT model on the newly translated data. This is essentially an extreme form of domain adaptation where we adapt a base NMT model trained on out-of-domain data to a new domain. Specifically, we adapt a strategy of *mixed fine-tuning* (Luong and Manning, 2015), which continues training a pre-trained out-of-domain model on both in-domain data and a certain amount of out-of-domain data to prevent overfitting to relatively small in-domain data. Compared to the standard domain adaptation setting where we have only a small number of in-domain sentences, our phrase-level active learning setting has the additional difficulty of having to use short translations of individual phrases. In the following, we describe both methods to choose which data to use in mixed fine-tuning, and how to incorporate phrasal translations.

4.1 Data Mixing

For data mixing, we sample a subset $\mathcal{L}_r$ of data directly from the labeled set $\mathcal{L}'$, and concatenate $\mathcal{L}_r$ with the newly annotated sentences $\mathcal{L}_s$ and phrases $\mathcal{L}_p$ for mixed fine-tuning (Line 8 in Algorithm 1). Specifically, we define a distribution function ψ over $\mathcal{L}'$, and either sample by $(x, y) \sim \psi$ or greedily take the most likely data by $(x, y) = \text{argmax}_{(x,y)\in\mathcal{L}'} \psi(x, y)$ iteratively for M times to obtain the subset $\mathcal{L}_r$ of M parallel data.

Random Sampling: The most simple way to select out-of-domain data is to randomly sample sentences from the out-of-domain corpus $\mathcal{L}'$, i.e., $(x, y) \sim \text{Uniform}(\mathcal{L}')$. Although it is simple, this has been popularly used in the literature of domain adaption for NMT (Chu and Wang, 2018).

Retrieve Similar Sentences: Recently, Aharoni and Goldberg (2020) showed that pre-trained language models implicitly learn sentence embeddings that cluster by domains, and proposed a data selection method that has proven more effective than methods based on the likelihood of an in-domain language model (Moore and Lewis, 2010). Since our base NMT model is pre-trained on out-of-domain corpus, we need to adapt the model to the domain of the unlabeled data. Instead of random sampling, we adopt the selection method in Aharoni and Goldberg (2020) to retrieve parallel sentences from $\mathcal{L}'$ that are close to the in-domain sentences in $\mathcal{U}$. To do so, we leverage the contextualized sentence representations, and measure the distance of a source sentence in $\mathcal{L}'$ w.r.t. the unlabeled corpus $\mathcal{U}$ by $\text{ratio}(x,\mathcal{U})$, $\forall x \in \mathcal{L}'$. Next, we iteratively retrieve labeled data from $\mathcal{L}'$ that have the smallest distance scores to their nearest neighbors, i.e., $(x, y) = \text{argmax}_{(x,y)\in\mathcal{L}'} \text{ratio}(x,\mathcal{U})$.

4.2 Incorporating Phrasal Translations

In addition to obtaining real parallel data from $\mathcal{L}'$ for mixed fine-tuning, we create synthetic parallel data $(\hat{x}, \hat{y})$ by incorporating phrasal translations into existing context from $\mathcal{L}'$. Specifically, for an unlabeled sentence $x \in \mathcal{U}$ containing a newly annotated phrase p_x, we retrieve the most similar sentence pair (x^*, y^*) from $\mathcal{L}'$ by

$$(x^*, y^*) = \underset{(x',y')\in\mathcal{L}'}{\text{argmax}} \, \text{ratio}(\mathrm{e}_x, \mathrm{e}_{x'}) \qquad (7)$$

We then alter (x^*, y^*) with the newly annotated phrase pair (p_x, p_y) to create synthetic sentence pair $(\hat{x}, \hat{y})$. Similar to data mixing, we concatenate the set of synthetic data with the annotated sentences $\mathcal{L}_s$ and phrases $\mathcal{L}_p$ for mixed fine-tuning.

Switch Phrases: Inspired by existing data augmentation methods (Fadaee et al., 2017), we examine a data augmentation method that switches out phrases in the out-of-domain sentence pairs in $\mathcal{L}'$ by the newly annotated phrase pairs from $\mathcal{U}$. First, we define the following operation Switch(x, p, i) that returns a new sentence by substituting the phrase at the i-th position in x^* by p_x.

$$\text{Switch}(x^*, p_x, i) = [x^*_{<i}; p_x; x^*_{\geq i+|p_x|}] \tag{8}$$

Next, we enumerate all possible positions in x^* for switching phrases, and then apply the in-domain language model trained on $\mathcal{U}$ to select the most probably synthetic sentence by

$$\hat{x} = \underset{\substack{x'=\text{Switch}(x^*,p_x,i) \\ \forall 0 \leq i < |x^*|-|p_x|, \\ p_x \in \cup_{n\in[1,N]} n\text{-gram}(x)}}{\text{argmax}} P_{\text{LM}}(x'), \tag{9}$$

where p_x is a phrase in the unlabeled sentence x. Notably, we use a 4-gram language model implemented in KenLM[3]. Since sentences are usually short (average length of 10-25 words), creating a synthetic sentence takes $O(|x^*||x|)$ scoring operations by the language model.

To synthesize the corresponding $\hat{y}$ from the retrieved target sentence y^*, we apply a word alignment model trained on $\mathcal{L}$ to find the index j for the translation of the replaced phrase $x^*_{i:i+|p_x|}$ in y^*, and substitute the phrase at the j-th position in y^* by p_y to obtain $\hat{y} = \text{Switch}(y^*, p_y, j)$.

Contextualized Phrases: The other idea is to augment the context of a newly annotated phrase pair (p_x, p_y), since a phrase p_x lacks larger sentence structure. Specifically, we define the contextualized operation that augments a phrase p_x in x by appending it to the retrieved sentence x^*.

$$\text{Contextualize}(x^*, p_x) = [x^*, p_x] \tag{10}$$

We then enumerate all annotated phrases in x, and apply an in-domain language model to find the most probable annotated phrase pair (p_x, p_y) that synthesizes $\hat{x}$. The corresponding $\hat{y}$ can be obtained by Contextualize(y^*, p_y).

$$\hat{x} = \underset{\substack{x'=[x^*,p_x] \\ \forall p_x \in \cup_{n\in[1,N]} n\text{-gram}(x)}}{\text{argmax}} P_{\text{LM}}(x') \tag{11}$$

5 Experiments

5.1 Experimental Setting

We use the WMT14 German-English data as the out-of-domain labeled data for training our base NMT model, and take the source sentences of two parallel corpora in the medicine and IT domains (Koehn and Knowles, 2017) as the unlabeled data. More details can be found in Appendix B.1.

As our NMT model, we use a 6-layer 512-unit Transformer (Vaswani et al., 2017) implemented in Fairseq,[4] and use a subword vocabulary of 50,000 for both languages constructed by Byte Pair Encoding (Sennrich et al., 2016). We train the base model with Adam for 10 epochs with 4K warmup steps and a peak learning rate of 1e-3, and decay the learning rate based on the inverse square root of the number of update steps (Vaswani et al., 2017).

For active learning, we set our annotation budgets by number of words translated (following the prevailing translation market practice to charge for jobs by the word), and investigate the budgets from 2.5K words up to 40K words.[5] After data selection (§3), we obtain a set $\mathcal{L}_r$ of M parallel sentences (§4), and set the size $M = |\mathcal{L}_p|$ where $\mathcal{L}_p$ is selected by NGF-SMP. We then fix $\mathcal{L}_r$ for mixed fine-tuning in all experiments, and continue fine-tuning the base model on a mixture of the newly-translated data and $\mathcal{L}_r$ for 5 more epochs.

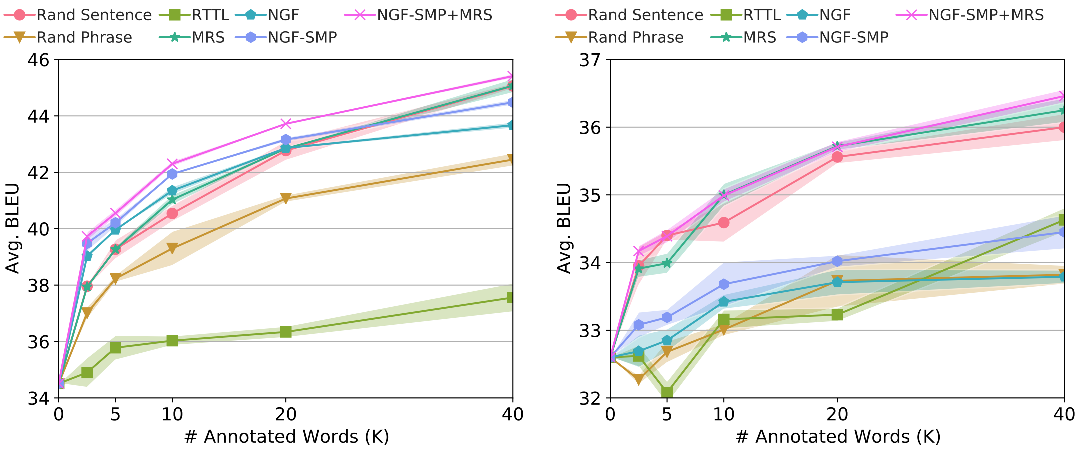
5.2 Word-level Translation Accuracy

Since our selection and mixed fine-tuning methods focus on leveraging phrasal translations for domain adaptation, we perform a fine-grained analysis on the word-level translation accuracy of the NMT systems due to the domain shift. A source word is defined as an unseen in-domain word when it never appears in the out-of-domain corpus. If phrase selection strategies select more in-domain words, we would expect a higher translation accuracy of such in-domain words by the adapted NMT systems using phrase selection. As a result, we compare the

[3] https://github.com/kpu/kenlm

[4] https://github.com/pytorch/fairseq

[5] At current market rates, this would cost from 491 to 7,092 USD for German-English translation by professional translators at https://translated.com/.



(a) WMT14→Medicine

(b) WMT14→IT

Figure 2: Average BLEU score over 3 runs for adapting a base NMT to the Medicine and IT domains.

translation accuracy of in-domain words by the NMT models using different selection strategies in Figure 3. As shown in the figure, NGF-SMP significantly improves the translation accuracy of the in-domain words with a small annotation budget. In contrast, MRS falls short of the other compared methods when the annotation budget is less than 80K words. Moreover, we find that the hybrid selection strategy of NGF-SMP and MRS can combine the merits of both methods, and obtain an even higher accuracy when the budget is greater than 40K annotated words. Qualitatively, the example in Table 1 shows the translations for a source sentence with all words appearing in the medical domain. The NMT model adapted by MRS translates the first half of the source sentence by picking the correct word "exercised", while the NMT model adapted by NGF-SMP generates the correction translation "somnolence" in the second half of the output. The NMT model using the hybrid of NGF-SMP and MRS strategies translates both words correctly (more examples in Appendix B.2).

5.3 How Does Each Selection Strategy Help?

We examine the question of which selection strategy (§3) best improves accuracy on in-domain test data. For mixed fine-tuning, in this section we use the retrieved out-of-domain parallel data for a fair comparison among all active selection strategies. Figure 2 shows the average BLEU score and the standard deviation of the adapted MT systems to two new domains over 3 independent runs.[6]

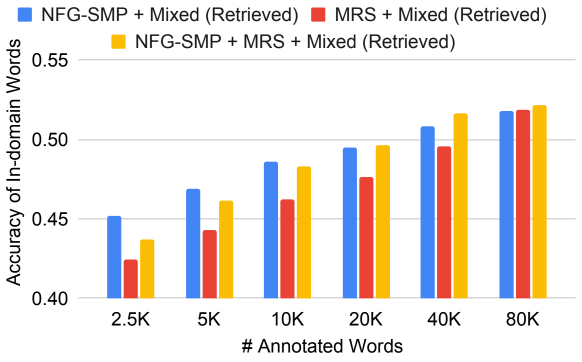


Figure 3: Translation accuracy of in-domain words in the test set from the medicine domain

Comparing among sentence selection strategies in Figure 2, MRS performs slightly better than the random sentence selection baseline on adapting the NMT model to the IT domain with smaller standard deviation values, and performs comparably on adapting to the medicine domain. However, we observe that RTTL performs worst, and we conjecture that this is due to the usage of the base NMT models that are trained on the out-of-domain parallel data in both directions. The errors accumulated from the round trip translation process lead to an inaccurate estimation of the uncertainty score for a source sentence. Table 2 shows the top 5 sentences selected by RTTL. The selected sentences in the medicine domain are short phrases rather than complete sentences, and those selected in the IT domain contain duplicate phrases such as "bewerten mitâ".

[6] To obtain a stable result, we independently run the active learning procedure with different selection strategies 3 times, collect new translation data, and concatenate them with the same data retrieved from out-of-domain labeled data

	Output	S-BLEU
Source	Jedoch ist Vorsicht geboten, da Berichten zufolge Verwirrung und Somnolenz während der Behandlung auftreten können.	
Reference	However, caution should be exercised as confusion and somnolence have been reported.	
NGF-SMP	However, caution is required, as there are reports of confusion and somnolence during the treatment.	15.71
MRS	However, caution should be exercised, as confusion and drowsiness may occur during the treatment.	15.62
NGF-SMP+MRS	However, caution should be exercised as confusion and somnolence may occur during the treatment.	15.71

Table 1: Translations generated by NMT models using different selection strategies. The last column shows the sentence BLEU score of the translations. Translation errors are highlighted in red.

MED	Portugal Lundbeck Portugal Lda Quinta da Fonte Edifício D. Bronchitis Gastrointestinaltrakt : Neugebore 139 B.
IT	Eigenschaften des Stichwortes â % 1â bewerten mitâ Drei Sternenâ keine Speicherplatzinformation aufâ procfsâ bewerten mitâ Einem Sternâ neue und einzelne auswÃ Â hlen

Table 2: Top 5 sentences selected by RTTL

For phrase-based selection methods, NGF-SMP significantly outperforms the random phrase selection strategy. Further, NGF-SMP even outperforms sentence selection methods when the annotation budget is small (less than 20k words) for adaption to the medicine domain. As we increase the annotation budget to 40K annotated words, sentence selection strategies outperform phrase selection strategies. This indicates that if we keep training NMT systems on shorter phrase pairs when the annotation budget is sufficient, the NMT systems would be limited by lack of longer sentence structures. In Figure 2b, we also find that NMT models trained with phrasal translations fall short of those trained with sentence translations when adapting to the IT domain. It is hard to train the NMT systems to translate certain phrases correctly without the sentence context. For example, "Persönlichen Ordner" in the IT domain is translated to "home directory" rather than "personal folder" in the sentence "jedes Skript dieses Dialogs hat Schreib-Zugriff auf Ihren Persönlichen Ordner ".

Finally, the hybrid selection of NGF-SMP and MRS strategies outperforms the individual selection strategies over every budget in our set of budgets, i.e., 2.5K, 5K, 10K, 20K, 40K annotated words, improving the best phrase selection strategy NGF-SMP by 0.49 average BLEU points, and the best sentence selection strategy MRS by 1.11 average BLEU points in the medicine domain. Notably, the phrase-based selection strategy especially helps in the scenario where the context is not required to translate domain-specific words, for example, the name of a medicine or a disease in the medicine domain (See the first example in Appendix B.2). For the adaptation scenario that requires a longer context in some domains such as IT, the hybrid strategy can also significantly outperforms the best phrase-based strategy NGF-SMP by 1.2 average BLEU points, and the best sentence selection strategy MRS by 0.15 BLEU points. Overall, our hybrid selection strategy is effective to combine the merits of both sentence and phrase selection strategies in the domain adaptation setting.

5.4 How Representative Are the Selected Data?

If the selected data has a significant overlap of segments with the in-domain test data, we would expect a better adaptation performance of the NMT trained on the selected data. Therefore we investigate the n-gram overlap between the selected data and the test data when we annotate 5K words from the medicine corpus, and report the average BLEU score of the adapted NMT models trained on the selected data in Table 3. Interestingly, we find that there exists a high correlation ($\rho \approx 0.8$) between the n-gram overlap and the average BLEU score, which indicates that the n-gram overlap with the test set can be used as a good measure of whether the selected data is useful for improving the NMT model in the new domain. Compared to the random phrase selection, NGF-SMP selects phrases with a high overlap with the test data. We also observe that sentence selection strategies cover fewer phrases in the test data than phrase selection strategies. This also corroborates our assumption that asking translators to annotate phrases that the MT system can already translate well is not cost-effective to improve the in-domain translation performance.

Methods	uni-gram	bi-gram	tri-gram	4-gram	Avg. BLEU
OoD Data	79.33	32.65	7.30	1.10	34.51
+ Random Sentence	82.81	38.45	11.62	3.73	39.27
+ RTTL	80.70	35.76	9.85	3.04	35.78
+ MRS	82.74	38.83	12.01	4.05	39.27
+ Random Phrase	82.36	35.84	7.98	1.15	38.23
+ NGF	84.45	41.82	14.94	6.17	39.96
+ NGF-SMP	85.80	43.13	16.15	7.11	40.21
+ NGF-SMP + MRS	84.48	41.89	14.98	6.48	40.55
ID Training Data	98.58	87.30	67.61	52.11	57.59
Pearson Correlation	0.90	0.83	0.80	0.78	/

Table 3: Percentage of the n-gram in the test sentences that are covered by the selected data with 5K words, the out-of-domain training data and the in-domain training data. The last row shows the Pearson correlation coefficient between n-gram overlap and avg. BLEU score.

5.5 How Redundant Are the Selected Data?

To answer this question, we first define “in-domain words” as words that only appear in the in-domain test set but do not exist in the out-of-domain data. We report the statistics of the in-domain word types word counts in the selected data with 10K annotated words in Table 5. We find that phrase selection strategies select more unique in-domain word types and counts than the sentence selection strategies. This indicates that phrase selection strategies leverage the same amount of budget effectively to annotate more diverse in-domain words than sentence selection strategies.

5.6 How Do Phrasal Translations Help in Mixed Fine-tuning?

We further investigate the effect of mixed fine-tuning using the newly annotated in-domain data and sub-sampled out-of-domain data when comparing with fine-tuning only on the newly annotated data. Table 4 shows the average BLEU score and the standard deviation values over 3 independent runs. Compared to fine-tuning on only annotated data, adding randomly sampled sentence pairs from the out-of-domain data helps when the annotation budget is less than 5K annotated words, but hurts when we increase the budget. In contrast, adding sentences retrieved by the similarity in the sentence embedding space not only outperforms fine-tuning only on annotated data and mixed fine-tuning with randomly sampled sentences, but also achieves smaller standard deviation values. On the other hand, mixed fine-tuning on synthetic data by switching phrases performs slightly worse than the mixed fine-tuning on real retrieved data, but outperforms the fine-tuning without any out-of-domain data, especially when the annotation budget is small, e.g., 5K annotated words. Combining synthetic data by switching phrase and real retrieved data for mixed fine-tuning also improves the translation performance over the training only on synthetic data. However, the contextualized method performs worst among all mixed fine-tuning methods, which indicates that simply appending existing sentence context to phrasal translations might potentially introduce noise to the training data.

6 Related Work

Active Learning for Machine Translation Pioneering works on active learning for machine translation focus on selecting sentences that are most useful for training PBMT. This includes sentence selection strategies based on maximizing the percentage of unseen n-gram (Eck et al., 2005), n-gram frequency, lexical diversity (Haffari et al., 2009), or in-domain coverage (Ananthakrishnan et al., 2010). These sentence selection strategies have been used in active learning algorithms to deal with static data in the batch mode (Ananthakrishnan et al., 2010), or steaming data in the interactive setting (González-Rubio et al., 2012; Peris and Casacuberta, 2018; Lam et al., 2019).

For phrase-level annotations, there have been a few works applying phrase-based selection (Bloodgood and Callison-Burch, 2010; Miura et al., 2016) to PBMT. While the annotated phrases can be easily integrated by adding them with estimated translation probability to the existing phrase table in PBMT, it it less trivial to integrate these phrase-level annotations in NMT. Arthur et al. (2016) integrated the word-level translations to NMT by interpolating the probability of the NMT decoder with the estimated lexical probability. However, this approach requires a modification of the NMT model. Our paper investigates data-driven approaches that augment the training data by leveraging annotated phrases and existing parallel data.

Word/Phrase-based Data Augmentation The other line of research investigates data augmentation methods that leverage word or phrase translations to create synthetic parallel data for training MT models. This includes augmentation methods that replace a word in the existing parallel data with a low-frequency word sampled from the frequency distribution of the vocabulary (Xie et al., 2017) or from the probability of language models in both directions (Fadaee et al., 2017; Kobayashi, 2018). Wang et al. (2018) proposed an effective method

Out-of-domain Data				In-domain Data		2.5K	5K	10K	20K	40K
Sampled	Retrieved	Switched	Contextualized	NGF-SMP	MRS					
				✓		39.39 ± 0.14	39.22 ± 0.00	40.56 ± 0.02	41.19 ± 0.25	44.07 ± 0.33
					✓	37.94 ± 0.08	38.68 ± 0.54	40.62 ± 0.59	42.62 ± 0.03	45.00 ± 0.11
				✓	✓	38.94 ± 0.02	39.60 ± 0.09	41.34 ± 0.12	42.44 ± 0.15	44.90 ± 0.06
✓				✓	✓	39.46 ± 0.14	40.51 ± 0.23	40.62 ± 0.49	41.82 ± 0.26	43.78 ± 0.57
	✓			✓	✓	**39.73** ± 0.16	40.55 ± 0.14	**42.30** ± 0.10	**43.72** ± 0.04	**45.41** ± 0.08
		✓		✓	✓	38.93 ± 0.36	40.59 ± 0.17	41.82 ± 0.29	42.70 ± 0.37	45.33 ± 0.04
			✓	✓	✓	35.36 ± 0.38	37.85 ± 0.68	39.96 ± 0.35	42.83 ± 0.11	44.14 ± 0.15
	✓	✓		✓	✓	39.61 ± 0.06	**40.95** ± 0.06	42.19 ± 0.08	43.42 ± 0.17	45.06 ± 0.19
	✓		✓	✓	✓	37.88 ± 0.25	39.52 ± 0.32	41.17 ± 0.28	42.80 ± 0.21	44.28 ± 0.13

Table 4: Comparison between mixed fine-tuning methods. Bold indicates highest average BLEU by column.

Methods	IDWT	WT	$\frac{\text{IDWT}}{\text{WT}}$	IDWC	WC	$\frac{\text{IDWC}}{\text{WC}}$
Random Phrase	787	2206	35.68	860	5003	17.19
NGF	489	1053	46.44	889	5002	17.77
NGF-SMP	796	1492	53.35	1076	5001	21.52
Random Sentence	631	1984	31.80	712	5023	14.17
RTTL	592	1338	44.25	961	5023	19.13
MRS	647	2056	31.47	721	5023	14.35
NGF-SMP + MRS	667	1755	38.01	859	5035	17.06

Table 5: Statistics of the unique in-domain word types and word counts in the selected data with 10K annotated words.

that randomly replaces words in parallel sentences with other random words from the in-domain vocabulary. A more recent work on dictionary-based data augmentation (Peng et al., 2020) proposed to use an existing high-quality in-domain dictionary, and replaced a source word in the existing parallel data by the most similar word in the dictionary according to the cosine similarity metric in the embedding space. In contrast, we select noisy in-domain phrases using different phrase-based selection strategies (§3.2) to ensure the selection quality in an active learning process.

7 Discussion and Future Work

In this paper, we investigate ways to incorporating phrasal translations into training NMT for domain adaptation in the active learning setting. We find that phrasal translation is particularly useful in the adaptation scenario where longer sentence context is not necessarily required to translate in-domain words correctly. In contrast, NMT systems can benefit from learning sentence structure with sentence-based selection strategies. The hybrid selection strategies can combine the merits of both sentence-based and phrase-based selection strategies. Nonetheless, there are several future directions. (1) It is worth exploring how different annotation strategies may result in a difference in cost or time. (2) Although several findings could be generalized to other language pairs, testing our methods on morphologically rich languages is our next step. (3) Our current hybrid strategy simply allocates the annotation budget evenly without assuming any prior knowledge of the strategies and the translation performance. Techniques in multi-armed bandit problems (Gittins et al., 2011) can be used to learn a good allocation strategy.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. 2010. A semi-supervised batch-mode active learning strategy for improved statistical machine translation. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 126–134, Uppsala, Sweden. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Michael Bloodgood and Chris Callison-Burch. 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.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

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

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, Portland, Oregon, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Rashmi Gangadharaiah, Ralf D. Brown, and Jaime Carbonell. 2009. Active learning in example-based machine translation. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 227–230, Odense, Denmark. Northern European Association for Language Technology (NEALT).

John Gittins, Kevin Glazebrook, and Richard Weber. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 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.

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the International Conference on Machine Learning 1*, pages 7449–7459.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2019. Interactive-predictive neural machine translation through reinforcement and imitation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 96–106, Dublin, Ireland. European Association for Machine Translation.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Akiva Miura, Graham Neubig, Michael Paul, and Satoshi Nakamura. 2016. Selecting syntactic, non-redundant segments in active learning for machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–29, San Diego, California. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (short papers), Uppsala, Sweden. Association for Computational Linguistics.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. Dictionary-based data augmentation for cross-domain neural machine translation. *arXiv preprint arXiv:2004.02577*.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. In *International Conference on Learning Representations (ICLR)*, Toulon, France.

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

Pei Zhang, Xueying Xu, and Deyi Xiong. 2018. Active learning for neural machine translation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 153–158. IEEE.

Appendix

A Pseudo code

Algorithm 1 shows the active learning procedure for machine translation, which consists of two main steps: selection/translation (§3) and fine-tuning (§4).

Algorithm 1 Active Learning for Domain Adaptation of Machine Translation

```
procedure ActiveAdaptation($\mathcal{U}, \mathcal{L}, B$)
    Inputs: the unlabelled set $\mathcal{U}$, the labelled set $\mathcal{L}$, and a budget $B$.
    Train a MT model $\theta$ on $\mathcal{L}$.
    $\mathcal{S}, \mathcal{P} \leftarrow$ Selection($\mathcal{U}, \mathcal{L}, B$)
    Translate $\mathcal{S}$ by $\mathcal{L}_s = \{(x, \mathcal{O}(x)) | x \in \mathcal{S}\}$
    Translate $\mathcal{P}$ by $\mathcal{L}_p = \{(p, \mathcal{O}(p)) | p \in \mathcal{P}\}$
    $\mathcal{L}_r \leftarrow$ Obtain parallel data from $\mathcal{L}$ (§4)
    Fine-tune $\theta$ on $\mathcal{L}_s \cup \mathcal{L}_p \cup \mathcal{L}_r$
return $\theta$
```

Algorithm 2 Hybrid Phrase/Sentence Selection

```
procedure Selection($\mathcal{U}, \mathcal{L}, B$)
    Inputs: the unlabelled set $\mathcal{U}$, the labelled set $\mathcal{L}$, and a budget $B$.
    Initialize $\mathcal{S} = \{\}$, $\mathcal{P} = \{\}$
    Allocate the budget: $B_s, B_p \leftarrow B$
    while $\sum_{x \in \mathcal{S}} c(x) < B_s$ do
        $x \leftarrow \mathrm{argmax}_{x \in \mathcal{U}}\, \phi(x, \cdot)$
        $\mathcal{U} = \mathcal{U} \setminus \{x\}$
        $\mathcal{S} = \mathcal{S} \cup \{x\}$
    Construct $\mathcal{P}_{\mathcal{U}}, \mathcal{P}_{\mathcal{L}}$ by strategies (§3.2)
    while $\sum_{p \in \mathcal{P}} c(p) < B_p$ do
        $p \leftarrow \mathrm{argmax}_{p \in \mathcal{P}_{\mathcal{U}}}\, \mathbf{occ}(p, \mathcal{U})$
        $\mathcal{P}_{\mathcal{U}} = \mathcal{P}_{\mathcal{U}} \setminus \{p\}$
        $\mathcal{P} = \mathcal{P} \cup \{p\}$
    return $\mathcal{S}, \mathcal{P}$
```

B Experiments

B.1 Experimental Details for Reproducibility

Dataset: As pointed out in Aharoni and Goldberg (2020), there is overlap between the training data and the test data in the original split of the two corpora provided by Koehn and Knowles (2017), so we follow them in removing the duplicated sentences in the in-domain data, and re-splitting two new test sets in order to prevent the model from memorizing the selected in-domain training data that could potentially be included in the test data. Table 6 shows the data statistics.

Data	Domain	Lang	#Sentences	#Words	Vocab	Avg Len
$\mathcal{L}$	WMT14	De	4.4M	108.0M	1.9M	24.4
		En		114.5M	955.3K	25.8
$\mathcal{U}$	Medicine	De	227.2K	3.8M	114.3K	16.8
	IT	De	190.6K	2.1M	114.6K	11.5

Table 6: Data statistics of the out-of-domain labeled data in WMT14 and the in-domain unlabeled data in the medicine and IT domains.

Model: As our NMT model, we use a 6-layer 512-unit Transformer (Vaswani et al., 2017) implemented in `Fairseq`,[7] and use a subword vocabulary of 5,000 for both languages constructed by Byte Pair Encoding (Sennrich et al., 2016). The model has 45M parameters.

Training: We train the base model with Adam for 10 epochs with 4K warmup steps and a peak learning rate of 1e-3, and decay the learning rate based on the inverse square root of the number of update steps (Vaswani et al., 2017). We save the last checkpoint as our base model, and continue fine-tuning the base model on a mixture of the newly-translated data and the retrieved out-of-domain data for 5 more epochs.

Training/Inference Time: We train each model on one NVIDIA RTX 2080Ti GPU for all our experiments. Training the base NMT model takes less than 1 days, and fine-tuning the base NMT model on selected data takes less than 4hours. The decoding of 2000 sentences can be finished within 5 minutes.

B.2 Qualitative Analysis

In the first example of Table 7, the NMT model adapted by NGF-SMP can predict most words correctly while the NMT model adapted by MRS generate a random sentence.

B.3 Do Phrasal Annotations Bias NMT?

Since phrasal annotations are short and do not contain complex sentence structure, we hypothesis that NMT systems trained on phrasal annotations would be biased towards generating shorter sentences or sentences in different grammatical order w.r.t. the reference sentence. To understand this question, we analyze the length ratio between the translation outputs and the reference sentences in Figure 4.

[7] https://github.com/pytorch/fairseq

	Output	S-BLEU
Source	Schwindel, Parästhesie, Geschmacksstörung	
Reference	Dizziness, paraesthesiae, taste disorder	
NGF-SMP	Dizziness, paraesthesia, taste disturbance	23.27
MRS	The room was very small and the bathroom was very small.	0.00
NGF-SMP+MRS	Dizziness, paraesthesia, taste disturbance	23.27
Source	Über Hospitalisierung oder Todesfälle in Verbindung mit Infektionen wurde berichtet.	
Reference	Hospitalisation or fatal outcomes associated with infections have been reported.	
NGF-SMP	There have been reports of Hospitalisation or death associated with infections.	29.79
MRS	Hospitals or deaths associated with infections have been reported.	54.63
NGF-SMP+MRS	There have been reports of Hospitalisation or fatality associated with infections.	29.79

Table 7: Translations generated by NMT models using different selection strategies. The last column shows the sentence BLEU score of the translations. Translation errors are highlighted in red.

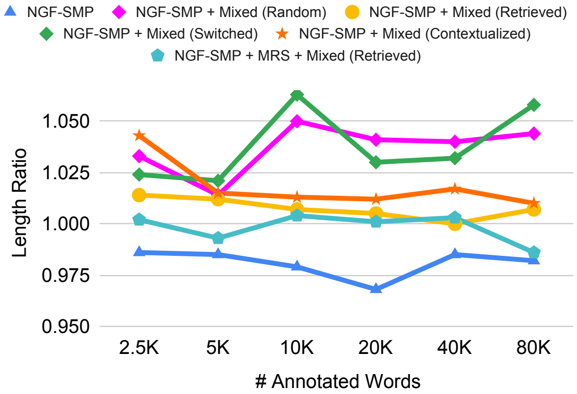


Figure 4: Length ratio between the NMT outputs and the reference sentences.

We find that the NMT model trained only on annotated phrases selected by NGF-SMP generates shorter sentences than reference sentences. In contrast, adding sentences randomly sampled from the labeled corpus $\mathcal{L}$ make the NMT model generate longer sentences than the reference sentences, while retrieving sentences from $\mathcal{L}$ that are similar to the sentences in $\mathcal{U}$ makes the model produces translation outputs with closed lengths as the reference sentences. Qualitatively, we also show the problem of generating sentences with different structures as the reference sentences in the third example in Table 1. In the third example, the NMT model trained with NGF-SMP produces a translation in an active voice, while the reference sentence uses a passive voice.