To Label or Not to Label: Hybrid Active Learning for Neural Machine Translation

Abdul Hameed Azeemi and Ihsan Ayyub Qazi and Agha Ali Raza
Lahore University of Management Sciences
Pakistan
{abdul.azeemi, ihsan.qazi, agha.ali.raza}@lums.edu.pk

Abstract

Active learning (AL) techniques reduce labeling costs for training neural machine translation (NMT) models by selecting smaller representative subsets from unlabeled data for annotation. Diversity sampling techniques select heterogeneous instances, while uncertainty sampling methods select instances with the highest model uncertainty. Both approaches have limitations - diversity methods may extract varied but trivial examples, while uncertainty sampling can yield repetitive, uninformative instances. To bridge this gap, we propose Hybrid Uncertainty and Diversity Sampling (HUDS), an AL strategy for domain adaptation in NMT that combines uncertainty and diversity for sentence selection. HUDS computes uncertainty scores for unlabeled sentences and subsequently stratifies them. It then clusters sentence embeddings within each stratum and computes diversity scores by distance to the centroid. A weighted hybrid score that combines uncertainty and diversity is then used to select the top instances for annotation in each AL iteration. Experiments on multi-domain German-English and French-English datasets demonstrate the better performance of HUDS over other strong AL baselines. We analyze the sentence selection with HUDS and show that it prioritizes diverse instances having high model uncertainty for annotation in early AL iterations.

1 Introduction

The training process for Machine Translation (MT) models typically requires large, high-quality parallel corpora to achieve robust performance. However, acquiring such data is costly and laborious, especially for low-resource languages and domain-specific data which requires professional translators with domain knowledge. For instance, training an MT model to translate specific medical conditions requires extensive labeling by medicine experts. Active learning (AL) is a well-known strategy that helps to reduce the labeling requirements by selecting a smaller representative subset for annotation, thereby reducing the overall cost. AL follows an iterative procedure that involves (1) querying instances from an unlabeled pool; (2) obtaining annotations for unlabeled instances through a human expert; and (3) adding the annotated examples to the labeled dataset and retraining.

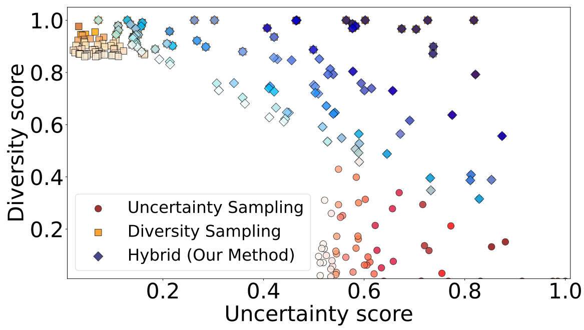

Figure 1: Comparison of the instances selected through different active learning strategies on the task of English to German (En → De) translation using the Law domain dataset and BART-base model. The color represents the active learning iteration; from the darkest (first iteration) to the lightest (tenth iteration). Our proposed hybrid active learning method (HUDS) selects diverse yet challenging sentences for annotation in early iterations.

In the context of Neural Machine Translation (NMT), active learning aids in selecting a small number of sentences for labeling that are likely to bring a similar performance if the NMT model is trained on a larger labeled set. AL methods for NMT can be broadly classified into diversity and uncertainty sampling methods, often described as the "two faces of active learning" (Dasgupta, 2011). Diversity sampling ensures that a heterogeneous set of instances from the unlabeled set is selected for annotation. In contrast, uncertainty sampling attempts to select the instances with the highest model uncertainty, i.e., the ones which are the most difficult for the next AL iteration. Both these methods have their limitations, such that the methods

Proceedings of the 31st International Conference on Computational Linguistics, pages 3071–3082
January 19–24, 2025. ©2025 Association for Computational Linguistics

focusing on diversity can extract varied but trivial examples, whereas uncertainty sampling may lead to the selection of instances with high uncertainty but which are repetitive and not useful. To mitigate this, hybrid sampling approaches have been proposed for other NLP tasks, including named entity recognition (Kim et al., 2006), abstractive text summarization (Tsvigun et al., 2022a), and text classification (Yuan et al., 2020). No hybrid AL strategy for efficiently acquiring domain-specific data in NMT has been proposed yet which successfully incorporates model uncertainty and data diversity in the sampling procedure.

To develop a hybrid AL strategy for NMT, we should consider a multi-step procedure involving uncertainty computation, diversity sampling, and a weighted hybrid acquisition function that leverages both. In this work, we develop HUDS (Hybrid Uncertainty and Diversity Sampling), a hybrid active learning strategy for NMT. We first compute uncertainty scores for unlabeled sentences and stratify them. We then obtain embeddings for each sentence in a stratum, cluster them using k-MEANS and compute the diversity score for each instance by their distance to the centroid. Finally, we compute a hybrid sampling score through a weighted sum of the uncertainty and diversity scores and select the *top*-k instances with the highest scores for annotation in each iteration of AL. This weighted sampling ensures the selection of diverse yet challenging instances for annotation (Fig. 1), which contributes to the model's robust performance on domain-specific datasets.

1.1 Contributions

- To the best of our knowledge, we present the first hybrid active learning strategy (HUDS) for NMT.
- We evaluate HUDS on German-English and French-English datasets spanning different domains (Medicine, Law, IT, Ted talks) with the BART language model and demonstrate that HUDS outperforms other AL baselines in NMT.
- We examine the examples selected for annotation by HUDS and demonstrate that it prioritizes diverse yet challenging instances for annotation in early AL iterations. We also find that HUDS selects varied instances with higher unigram coverage compared to other AL strategies.

2 Preliminaries

2.1 Neural Machine Translation

The primary objective of a Neural Machine Translation (NMT) model is to convert a source sentence $\mathbf{X} = \{x_1, \ldots, x_S\}$ into a corresponding target sentence $\mathbf{Y} = \{y_1, \ldots, y_T\}$, where $\mathbf{X}$ and $\mathbf{Y}$ each contain S and T tokens, respectively. The likelihood of each token in the target sentence, given the source sentence, is depicted using the chain rule below:

$$P(\mathbf{Y} \mid \mathbf{X}; \theta) = \prod_{i=1}^{T} p\left(y_i \mid y_{0:i-1}, \mathbf{X}; \theta\right) \quad (1)$$

In the above equation, θ symbolizes the parameters of the model. NMT models strive to optimize the cross-entropy (CE) loss, achieving this by decreasing the negative log-likelihood of the training examples, represented as:

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{i=1}^{T} \log p\left(y_i \mid y_{0:i}, \mathbf{X}; \theta\right) \quad (2)$$

When the model performs inference, it produces probabilities for the target tokens via an auto-regressive process. These probabilities then aid in choosing high-probability tokens with the help of search strategies such as beam search.

2.2 Active Learning for NMT

We consider a labeled corpus $\mathcal{D}_l$ and an unlabeled corpus $\mathcal{D}_u$ for active learning in NMT. In the context of domain adaptation, $\mathcal{D}_l$ is a larger out-of-domain labeled corpus (e.g., WMT14 German-English corpus), and $\mathcal{D}_u$ is a small in-domain unlabeled corpus (e.g., Medicine domain German-English corpus). For acquiring the translation of the sentences selected by the query strategy, we consider a human oracle who can translate any sentence in the unlabeled data. We consider a fixed budget c for annotation such that all the annotation through must occur within that budget.

2.3 Contrasting Uncertainty and Diversity

Active learning strategies can be broadly classified into diversity-sampling methods and uncertainty-sampling methods, also known as "two faces of active learning" (Dasgupta, 2011). Diversity sampling methods aim to select diverse instances for annotation, whereas uncertainty sampling techniques use the model's uncertainty as a proxy to select

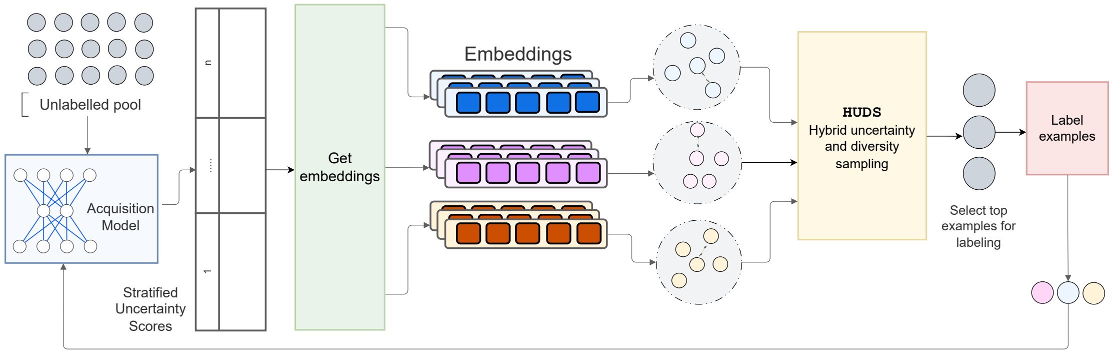

Figure 2: A representation of hybrid uncertainty and diversity sampling (HUDS) for active learning in NMT.

the hardest examples that should be annotated. To achieve optimal results, any robust AL strategy should leverage both uncertainty and diversity. However, designing such a hybrid approach is not trivial (Hsu and Lin, 2015) and is an open research problem (Yuan et al., 2020).

We now discuss two promising hybrid active learning strategies BADGE (Ash et al., 2019) and ALPS (Yuan et al., 2020) for classification tasks that combine the benefits of uncertainty and diversity sampling. Inspired by these two strategies, we then present a new hybrid active learning strategy for the task of neural machine translation.

2.3.1 BADGE

The primary objective of the BADGE algorithm is to select heterogeneous and uncertain instances for annotation in each iteration of AL. This is achieved by clustering special representations called *gradient embeddings* that encapsulate model certainty. Once the gradient embeddings are computed, BADGE uses k-MEANS++ seeding algorithm (Arthur and Vassilvitskii, 2007) to select a subset of examples S_t for annotation. The examples included in this subset are diverse in terms of uncertainty and data distribution due to the way gradient embeddings are calculated.

2.3.2 ALPS

Improving upon BADGE, ALPS attempts to estimate the uncertainty of instances using the masked language model (MLM) loss of a pre-trained BERT within the downstream classification task. Thus, ALPS works in a cold-start setting by leveraging the loss from a pre-trained BERT in contrast to BADGE which estimates uncertainty using the classification model that is being trained, i.e, in a warm-start setting. ALPS achieves this by creating a surprisal embedding s_x for each input x. This process involves feeding the non-masked input x into the BERT MLM head and calculating the cross-entropy loss for 15% randomly selected tokens. These embeddings are then clustered and a diverse subset of examples are selected for the next iteration.

3 Hybrid Uncertainty and Diversity Sampling

We now present a novel hybrid strategy for active learning in NMT named hybrid uncertainty and diversity sampling (HUDS). HUDS is a multi-step procedure involving uncertainty computation, diversity sampling, and a weighted hybrid acquisition function that leverages both of these. The complete workflow is shown in Fig. 2.

Uncertainty scores. We first estimate the uncertainty for an unlabeled sentence s through the normalized negative log-likelihood: $NNLL(X) = -\frac{1}{S}\sum_{j=1}^{S}\log p(x_j|x_1, x_2, \ldots, x_{j-1})$, where S is the sentence length and x_j is the j^{th} token in X. Lower NLL values indicate higher model confidence in the prediction, while higher values indicate less confidence thus greater uncertainty. We then perform stratification on uncertainty scores, with the range for ith stratum defined as,

$$s_i = \left(s_{min} + \frac{i-1}{n}(r), s_{min} + \frac{i}{n}(r)\right) \quad (3)$$

where s_{min} is the minimum uncertainty, s_{max} is the maximum uncertainty, r is the range specified by $s_{max} - s_{min}$, and n is the total number of strata.

Diversity scores. We then utilize a pre-trained BERT to obtain the embeddings of sentences within each stratum. Subsequently, we cluster the embeddings using k-MEANS and compute the diversity score for each instance by the cosine distance between that instance and the cluster centroid. The

Algorithm 1 Single iteration of HUDS

Input: Unlabeled pool $\mathcal{D}_u$, Number of strata n, Pre-trained encoder $\mathcal{P}$, Acquisition Model $\mathcal{M}$, Hybrid parameter λ, Instances to select k

$\mathcal{U}_s \leftarrow \mathcal{M}(u), u \in \mathcal{D}_u$ ▷ Compute uncertainty scores for each sentence in $\mathcal{D}_u$

$s_{min} \leftarrow min(\mathcal{U}_s), s_{max} \leftarrow max(\mathcal{U}_s), r \leftarrow s_{max} - s_{min}, \mathcal{N} \leftarrow \emptyset$

for $i = 1 \rightarrow n$ **do**

  $\mathcal{N} \leftarrow \mathcal{N} \cup \mathcal{U}_s \left(s_{\min} + \frac{i-1}{n}(r), s_{\min} + \frac{i}{n}(r)\right)$ ▷ Perform stratification on uncertainty scores $\mathcal{U}_s$

end

$\mathcal{C} \leftarrow \emptyset$

for $\mathcal{X}$ *in* $\mathcal{N}$ **do**

  $\mathcal{E} \leftarrow \emptyset$

  for *sentences* $x \in \mathcal{X}$ **do**

    $\mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{P}(x)$ ▷ Obtain embeddings for each sentence in stratum $\mathcal{X}$

  end

  $c \leftarrow$k$-$MEANS$(\mathcal{E})$ ▷ Perform k-MEANS clustering on embeddings

  $\mathcal{C} \leftarrow \mathcal{C} \cup \lambda \cdot d(c, e) + (1 - \lambda) \cdot \mathcal{U}_s(e), e \in \mathcal{E}$ ▷ Compute hybrid score for each sentence

end

$\mathcal{Q} \leftarrow topK(\mathcal{C}, k)$ ▷ Select top k instances for annotation

Output: $\mathcal{Q}$

goal of clustering is only to determine the distance of each example from the centroid of the (already defined) stratum and thus a value of $k = 1$ suffices here. In other words, diversity is represented with distance from the average vector, over sentences grouped by uncertainty strata. These diversity scores represent the heterogeneity of the examples. Therefore, the semantically dissimilar (and potentially harder) instances in each stratum have higher scores than the prototypical examples.

Hybrid scores. We compute a hybrid sampling score $H(x)$ for a sentence x through the weighted sum of the uncertainty and diversity scores (Eq. 4) and select the *top*-k instances with the highest scores for annotation.

$$H(x) = \lambda \cdot d(x, c_i) + (1 - \lambda) \cdot u_x \tag{4}$$

where $d(x, c_i)$ is the diversity score represented by cosine distance between the embedding of sentence x and the centroid of the cluster c_i of stratum s_i (to which the sentence x belongs), while u_x is the uncertainty score of sentence x. Algorithm 1 summarizes the complete procedure.

Stratification of uncertainty scores allows the selection of diverse sentences from each subpopulation, i.e., examples with a low uncertainty can also be selected as long as they are diverse. This is in contrast to the other pure uncertainty sampling methods which prevent the selection of informative sentences that have a low uncertainty. We hypothesize that hybrid sampling aids in selecting sentences with low overall uncertainty but having harder (dissimilar) segments that the NMT model cannot translate well.

4 Experimental Setup

We mirror the conventional active learning setup followed in previous studies (Shelmanov et al., 2021; Dor et al., 2020; Hu and Neubig, 2021a; Tsvigun et al., 2022a). We first train the NMT model on the out-of-domain labeled corpus. We then start the AL procedure with a randomly selected small subset of labeled sentences from in-domain data and in the first iteration fine-tune the NMT model on this subset. In each subsequent iteration, we use the query strategy $\mathcal{Q}$ to select the examples for annotation, add them to the annotated pool of examples (i.e., emulate the manual labeling process of translation without human involvement), and remove them from the unlabeled set. The NMT model is then trained on the sentence pairs in the labeled pool and evaluated on an unseen validation set. We use SACREBLEU (Post, 2018) to evaluate the performance on the validation set in each iteration, according to the protocol followed in previous works on AL in NMT (Hu and Neubig, 2021b; Zhao et al., 2020). SACREBLEU uses the standard WMT tokenization, addresses several issues in the reporting of regular BLEU scores, and is more reproducible (Post, 2018).

Dataset	# Train Pairs	# Val Pairs	# Test Pairs
Medicine	248,099	2,000	2,000
Law	467,309	2,000	2,000
IT	222,927	2,000	2,000

Table 1: Data statistics of the multi-domains German-English corpora in the Medicine, IT, and Law domains.

4.1 Datasets

We utilize the WMT14 En-De (English-German) parallel data as our out-of-domain labeled dataset. It contains 4.5M training pairs and 3000 validation and test pairs. For the in-domain data, we use the multi-domains dataset containing German-English parallel data for five domains (Koehn and Knowles, 2017), deduplicated and resplit by Aharoni and Goldberg (2020). This dataset has been frequently used for the evaluation of domain adaptation and active learning approaches in neural machine translation (Dou et al., 2019; Hu and Neubig, 2021a). We use three corpora for our experiments: Medicine, Law, and IT. The statistics are shown in Table 1.

4.2 Model

We use the BART model[1] for our experiments which works well for natural language generation tasks including translation (Lewis et al., 2020). BART is a transformer encoder-decoder model having a bidirectional encoder and an autoregressive decoder. Our choice of BART is motivated by recent works on active learning for text generation that have employed encoder-decoder architectures such as BART and demonstrated their effectiveness (Xia et al., 2024; Tsvigun et al., 2022a). We did not use a large language model (LLM) for the domain adaptation experiments, since LLMs are typically pre-trained on massive, diverse corpora that often overlap with the test domain. Such pre-training raises the risk of data contamination (Sainz et al., 2023), making it difficult to accurately assess the model's ability to learn from newly introduced, domain-specific data. As a result, improvements observed in an LLM would not reliably reflect genuine gains from the active learning process, but rather the model's prior exposure to similar samples.

We conduct experiments using the BART-base variant pre-trained on Wikipedia and the BookCorpus. This variant has 140 million parameters and six blocks in the encoder and decoder. We train the model using the AdamW optimizer with a training batch size of 16. The learning rate is set to 2e-5 with a weight decay of 0.028 and gradient clipping of 0.28. The beam size is set to 4 for evaluation. We use $\lambda = 0.5$ for HUDS sampling (after the first query iteration), giving equal weightage to uncertainty and diversity in the computation of hybrid scores. Uncertainty is computed for 20,000 randomly selected examples from the unlabeled pool as it is computationally expensive to be performed for the complete dataset. 1000 sentences are queried in each iteration. The experiments are run on a single 48GB NVIDIA A6000 GPU.

4.3 Baselines

AL strategies for NMT primarily include a sentence-level selection of examples or phrase-level selection for training the NMT model. Sentence-level strategies focus on selecting *sentences* that are most useful for training while phrase-level strategies choose individual phrases that are the most informative. In some scenarios, phrase-level strategies can be more cost-effective as they provide granular control to avoid the inclusion of otherwise informative sentences which have phrases that the NMT model can already translate well. However, their implementation for NMT is generally more complex than sentence-level selection, involving the creation of synthetic parallel data that incorporates phrasal translations followed by a data mixing procedure that is utilized for mixed fine-tuning (Hu and Neubig, 2021a). Thus, we compare our approach to three sentence-level selection baselines, including Random selection, Normalized Sequence Probability (NSP) (Ueffing and Ney, 2007), and In-Domain Diversity Sampling (IDDS) (Tsvigun et al., 2022a). NSP and IDDS are strong baselines, while random selection is competitive in AL given a sufficiently large annotation budget (Hu and Neubig, 2021b).

While BADGE and ALPS are promising AL methods for classification tasks, applying them to NMT requires significant modification to the acquisition functions. Thus, we do not consider them as baselines for AL in NMT. Additionally, we do not include subset selection approaches as baselines, e.g., Kothawade et al. (2022), as their goal is to select a subset from a *fixed* set of (often *fully labeled*) examples to allow training in resource-constrained environments. In contrast, active learning aims to iteratively select a subset of examples for labeling from an *unlabeled* pool, with the aim of minimiz-

[1] https://huggingface.co/facebook/bart-base

DATASET	AL STRATEGY	AL ITERATION 2	4	6	8	10
MEDICINE	Random	$29.36_{\pm 0.1}$	$30.75_{\pm 0.1}$	$31.44_{\pm 0.1}$	$31.93_{\pm 0.1}$	$32.70_{\pm 0.1}$
	NSP	$29.69_{\pm 0.6}$	$31.39_{\pm 0.5}$	$32.27_{\pm 0.4}$	$32.89_{\pm 0.3}$	$33.68_{\pm 0.2}$
	IDDS	$29.45_{\pm 0.1}$	$31.11_{\pm 0.1}$	$32.08_{\pm 0.3}$	$32.76_{\pm 0.2}$	$33.05_{\pm 0.0}$
	HUDS	**30.76**$_{\pm 0.2}$	**32.03**$_{\pm 0.1}$	**33.14**$_{\pm 0.1}$	**33.65**$_{\pm 0.1}$	**34.58**$_{\pm 0.1}$
IT	Random	$27.64_{\pm 0.3}$	$28.30_{\pm 0.1}$	$28.78_{\pm 0.1}$	$29.28_{\pm 0.1}$	$29.54_{\pm 0.1}$
	NSP	$28.43_{\pm 0.2}$	$29.26_{\pm 0.1}$	$29.70_{\pm 0.2}$	$30.09_{\pm 0.1}$	$30.23_{\pm 0.2}$
	IDDS	$23.81_{\pm 0.1}$	$24.31_{\pm 0.2}$	$24.84_{\pm 0.0}$	$25.21_{\pm 0.1}$	$25.41_{\pm 0.1}$
	HUDS	**29.15**$_{\pm 0.1}$	**29.69**$_{\pm 0.1}$	**30.12**$_{\pm 0.1}$	**30.54**$_{\pm 0.1}$	**30.70**$_{\pm 0.0}$
LAW	Random	$32.60_{\pm 0.3}$	$33.52_{\pm 0.5}$	$34.28_{\pm 0.4}$	$34.80_{\pm 0.5}$	$35.21_{\pm 0.5}$
	NSP	$32.89_{\pm 0.1}$	$34.18_{\pm 0.0}$	$34.73_{\pm 0.1}$	$35.16_{\pm 0.3}$	$35.53_{\pm 0.0}$
	IDDS	$32.91_{\pm 0.0}$	$34.12_{\pm 0.1}$	$34.61_{\pm 0.1}$	$35.07_{\pm 0.0}$	$35.33_{\pm 0.0}$
	HUDS	**33.76**$_{\pm 0.0}$	**34.39**$_{\pm 0.2}$	**34.89**$_{\pm 0.1}$	**35.17**$_{\pm 0.0}$	**35.62**$_{\pm 0.1}$

Table 2: SACREBLEU scores with different active learning strategies for NMT across different domain datasets. `BART-base` model is used in all experiments. Active learning is done over ten iterations with 1000 sentences queried in each iteration. For each strategy, three independent runs are done with different seeds, and the mean SACREBLEU is reported, with the subscript representing the standard error. Best results are given in **bold.** HUDS consistently shows better performance compared to other AL strategies.

ing the labeling cost.

Random Sampling. Random sampling involves selecting a subset of sentences randomly from the unannotated data for labeling. Despite its simplicity, it has shown strong performance in active learning for various NLP tasks (Miura et al., 2016; Tsvigun et al., 2022a) as it reflects the distribution of the complete unlabeled set without bias. Moreover, it is an efficient strategy as no queries are needed for example selection in contrast to the other AL strategies.

Normalized Sequence Probability (NSP). This is an uncertainty-based query strategy that helps to quantify the uncertainty of a model's predictions on a particular sequence. It was introduced by Ueffing and Ney (2007) and is a strong baseline that is the basis for several other AL techniques (Haffari et al., 2009; Tsvigun et al., 2022b). NSP for a sentence $\mathbf{v}$ is defined as $1 - \bar{p}_{\hat{\mathbf{w}}}(\mathbf{y} \mid \mathbf{v})$, where $\bar{p}_{\hat{\mathbf{w}}}(\mathbf{y} \mid \mathbf{v})$ is the geometric mean of the token probabilities predicted by model.

In-Domain Diversity Sampling (IDDS). This strategy (Tsvigun et al., 2022a) attempts to select sentences that are different from the previously annotated ones in an attempt to increase the diversity of the final set. Simultaneously, it prevents the selection of noisier instances (or outliers) that are significantly different semantically from the documents in the in-domain dataset. This mitigates the prevalent issue of selecting noisier instances found in other uncertainty-based AL strategies. The acquisition function of IDDS is given by:

$$IDDS(\mathbf{v}) = \alpha \frac{\sum_{k=1}^{|A|} f(\mathbf{v}, \mathbf{v}_k)}{|A|} - (1-\alpha) \frac{\sum_{m=1}^{|B|} f(\mathbf{v}, \mathbf{v}_m)}{|B|} \quad (5)$$

In this equation, $f(\mathbf{v}, \mathbf{v}')$ is a function measuring text similarity, B represents the labeled set, A represents the unlabeled set, and α is a hyperparameter between 0 and 1. IDDS demonstrates strong results on the Abstractive Text Summarization task, outperforming other AL strategies. Thus, we select this strategy as one of the baselines for comparison.

5 Results

Table 2 shows the BLEU scores of `BART-base` model fine-tuned with different AL strategies across three datasets. We do three runs with different random seeds for each selection strategy and dataset and report the mean SACREBLEU. We observe that HUDS performs better than random sampling, NSP, and IDDS across all iterations. For the medicine dataset, HUDS demonstrates 34.58 SACREBLEU at the end of the final iteration compared to 33.68 for NSP, 33.05 for IDDS, and 32.70 for Random. The differences in the SACREBLEU scores are statistically significant by Wilcoxon's signed-rank test ($p < 0.05$). We also notice that the relative ranking of SACREBLEU for different strategies varies significantly across datasets.

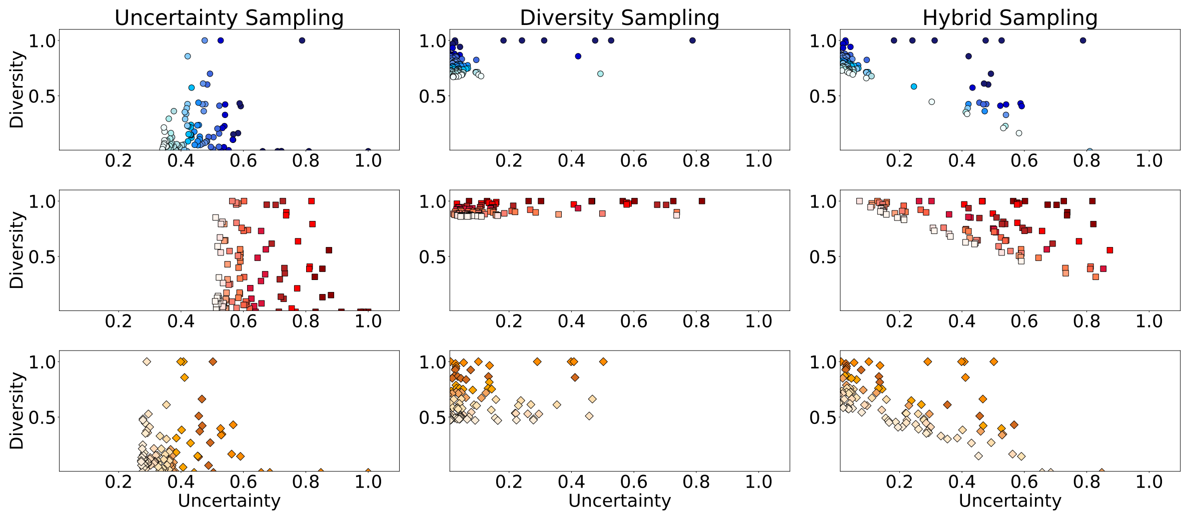

Figure 3: Plot of uncertainty vs diversity scores for Medicine (*Row 1*), Law (*Row 2*), and IT (*Row 3*) datasets. Each plot contains 100 markers representing sentences that were selected during the AL procedure over 10 iterations, with 10 sentences selected in each iteration. The color represents sampling iteration in AL, with the darkest color representing the first iteration and the lightest color representing the tenth iteration. Columns correspond to the sampling strategy. Across all three datasets, hybrid sampling tends to select sentences with high uncertainty and diversity.

We hypothesize that this indicates a difference in the distribution and, more specifically, the uncertainty and diversity of instances that we uncover next. Interestingly, IDDS does not perform well on the IT dataset, with a 13.8% lower SACREBLEU compared to random sampling (25.4 vs 29.5). We conjecture that this is due to the selection of diverse instances in each iteration that are different from the already annotated ones but are uninformative for the model.

5.1 Analyzing HUDS

To understand the sentence selection preference in HUDS during AL iterations, it is important to find out how it compares to regular uncertainty and diversity sampling. We plot uncertainty vs. diversity scores (Fig. 3) and analyze the unlabeled instances selected in ten iterations for annotation by each AL strategy (uncertainty, diversity, HUDS) across three datasets. In the earlier iterations (represented by dark-colored shapes), HUDS selects instances with high values of uncertainty and diversity for Medicine and IT datasets. For the Medicine and Law datasets, a few diverse instances have moderate to low uncertainty (Row 1 and 3, Column 2 of Fig. 3); hence HUDS selects those as well. In contrast, uncertainty sampling always selects uncertain instances regardless of their diversity, and diversity sampling prefers heterogeneous instances only. The difference in the distribution of uncertainty and diversity across datasets also helps explain the difference in the relative ranking of strategies in terms of SACREBLEU in Table 2.

5.2 Evaluation on En-Fr corpus

Random	IDDS	NSP	HUDS
34.8	33.5	35.1	**35.5**

Table 3: Comparison of the mean SACREBLEU after the final AL iteration on the IWSLT 2014 En-Fr validation set.

To confirm the generalization of HUDS to other languages, we conduct an experiment on the English-French (En-Fr) language pair. We utilize the WMT14 (En-Fr) parallel data as the out-of-domain labeled dataset and the IWSLT 2014 Ted Talks (En-Fr) as the in-domain dataset, following the protocol presented in Wang et al. (2017). We use the `BART-base` model for AL. HUDS demonstrates the highest SACREBLEU score on the IWSLT 2014 validation set compared to other strategies, confirming its generalization to En-Fr (Table 3).

5.3 Changing annotation size

To determine the impact of a larger annotation size on the translation quality, we conduct an experiment on the IT dataset with 5000 examples labeled in each iteration (instead of 1000 in the earlier experiments). Ten iterations are performed, which brings the final quantity of selected sentences to

50,000. The results in Table 4 show that HUDS consistently shows a higher SACREBLEU score than other methods with a larger annotation size.

Method	Iteration				
	2	**4**	**6**	**8**	**10**
Random	29.7	30.7	31.5	31.9	32.5
IDDS	25.6	27.1	27.9	28.7	29.1
NSP	30.5	31.2	31.9	32.0	32.3
HUDS	**30.8**	**32.3**	**32.9**	**33.4**	**34.0**

Table 4: Mean SACREBLEU score on the IT dataset. 5000 examples are selected for labeling in each iteration.

5.4 Composition of annotated sentences

To find out how the composition of annotated instances affects the generalization performance of the NMT model, it is useful to analyze the n-gram overlap between the selected sentences and the test set. Prior work (Hu and Neubig, 2021a) has shown that the n-gram overlap correlates highly with test BLEU scores, making it a reliable measure of the usefulness of selected data for domain adaptation. Thus, we conduct an experiment to analyze the unigram overlap between the instances selected by different methods and the validation set of IT corpus, following the protocol presented in Hu and Neubig (2021a). We find that the sentences selected by HUDS have the highest overall unigram overlap with the validation set, followed by random, NSP, and IDDS (Table 5). This indicates that HUDS selects varied instances with higher unigram coverage, potentially leading to better performance. We also compare the overlap of English and German individually, which shows a similar trend, except that NSP has a higher English unigram overlap than the randomly annotated sentences (46.7 vs 46.6).

Method	En Overlap (%)	De Overlap (%)	En-De Overlap (%)
HUDS	**47.2**	**53.2**	**48.0**
IDDS	46.0	52.1	47.3
NSP	46.7	52.2	47.6
Random	46.6	53.0	47.8

Table 5: Percentage overlap between the unigrams of sentences selected by different AL strategies and the validation set of IT corpus. We report the individual overlap for En and De and the En-De combined overlap.

5.5 COMET scores

We additionally assess the translation quality of HUDS using COMET (Rei et al., 2020), a state-of-the-art neural evaluation metric that demonstrates high correlation with human judgment. We use the `Unbabel/wmt20-comet-da` model[2] for scoring the translations on the IT domain dataset. The results in Table 6 show that HUDS outperforms other active learning strategies.

Random	IDDS	NSP	HUDS
42.5	38.0	42.3	**42.6**

Table 6: Comparison of the COMET scores after the final AL iteration for different AL strategies on the medical domain dataset. `Unbabel/wmt20-comet-da` model is used for evaluation. Active learning is done over ten iterations, with 1000 sentences selected in each iteration.

6 Ablation study

An important hyperparameter in the HUDS algorithm is the parameter λ that balances uncertainty and diversity score for a sentence. We now study the effect of varying the value of λ in HUDS on the SACREBLEU score for the multi-domains dataset. We run a separate AL procedure for each value of λ on different datasets and report SACREBLEU score at the end of the first iteration with 1000 sentences annotated and used for training in that iteration. The ablation is *done on a single iteration* to better isolate the impact of individual components and avoid any potential confounding effects. The results in Table 7 demonstrate a higher SACREBLEU score for $\lambda = 0.5$, indicating that a balance between hybrid uncertainty and diversity scores in HUDS leads to better generalization. Additionally, the parameter λ provides a powerful tuning method for uncertainty or diversity sampling using just the validation set.

Dataset	λ						
	0	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**1.0**
Medicine	29.6	29.6	29.7	**29.9**	29.7	29.7	29.5
Law	32.8	32.8	32.7	**33.0**	32.8	32.9	31.2
IT	28.7	28.8	28.6	**29.0**	28.7	28.6	28.5

Table 7: SACREBLEU score after the first AL iteration for different values of λ in HUDS.

7 Related Work

Active Learning in Machine Translation. Active Learning has been widely applied to machine translation (MT) to improve the quality and efficiency of translation systems, especially for low-resource languages or domains (Zhang et al., 2018; Hu and Neubig, 2021a; Gupta et al., 2021; Zhou

[2] `https://github.com/Unbabel/COMET`

and Waibel, 2021; Zhao et al., 2020). Several aspects of AL have been explored for MT, including different sampling techniques that consider uncertainty or diversity (Ambati et al., 2011), different workflows including batch mode vs. online mode (Ananthakrishnan et al., 2010; Lam et al., 2019), leveraging pre-trained or auxiliary models (Domhan and Hieber, 2017; Wang et al., 2020) and exploiting domain-specific and synthetic data (Zhang et al., 2019; Hoang et al., 2018). To our knowledge, no hybrid AL strategy for efficiently acquiring domain-specific data in NMT has been proposed prior to our work.

Uncertainty Sampling Techniques. Some of the early works on AL for MT use uncertainty-based sampling criteria, such as entropy or posterior probability, to select the sentences that the model is most uncertain about for annotation (Haffari et al., 2009; Bloodgood and Vijay-Shanker, 2009; Zhao et al., 2020). Using these sentences will enable the model to generalize better to unseen data. For NMT models, uncertainty is usually measured at the sentence or the beam level, based on the output of an encoder-decoder model or a transformer model (Zhou et al., 2020). However, pure uncertainty sampling techniques suffer from redundancy issues as they tend to select uncertain similar sentences or prefer certain words and structures.

Diversity Sampling Techniques. The main idea of diversity sampling is to select the sentences that are most diverse and representative of the unlabeled data based on some diversity measures such as density, representativeness, or coverage (Bloodgood and Callison-Burch, 2010; Gangadharaiah et al., 2009; Hu and Neubig, 2021a; Pendas et al., 2023). Diversity is usually measured at the word or phrase level based on the frequency or similarity of words or phrases. However, pure diversity sampling methods lead to the selection of noisier or irrelevant instances that are too difficult or too easy for the model.

8 Conclusion

In this work, we proposed HUDS, a novel hybrid active learning strategy that leverages both uncertainty and diversity for neural machine translation. HUDS outperforms other AL methods for NMT, demonstrating the effectiveness of hybrid sampling over other baselines for the same amount of human labeling effort. Future work includes (1) incorporating phrasal selection in HUDS and (2) making HUDS more efficient through precomputation of embeddings and intermediate caching of scores.

9 Limitations

We list the potential limitations of our work below: (1) HUDS involves computation of embeddings, clustering, and calculation of hybrid sampling scores for each unlabeled sentence, which could lead to higher latency compared to other AL methods. For large datasets, the additional computations may affect the *interactivity* of the AL procedure. Precomputation of embeddings and intermediate caching of scores can be leveraged to mitigate this. HUDS can also be augmented with automatic quality estimation methods for NMT (Blain et al., 2023) to increase efficiency. (2) The proposed hybrid sampling strategy HUDS focuses on sentence-level active learning for NMT. Phrase-level active learning could potentially be more cost-effective by selecting only informative phrases within sentences. Extending HUDS to phrase-level selection is left for future work.

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.

Vamshi Ambati, Stephan Vogel, and Jaime G Carbonell. 2011. Multi-strategy approaches to active learning for statistical machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*.

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.

David Arthur and Sergei Vassilvitskii. 2007. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh

Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. 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.

Michael Bloodgood and K. Vijay-Shanker. 2009. Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 137–140, Boulder, Colorado. Association for Computational Linguistics.

Sanjoy Dasgupta. 2011. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for bert: an empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong.

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).

Kamal Gupta, Dhanvanth Boppana, Rejwanul Haque, Asif Ekbal, and Pushpak Bhattacharyya. 2021. Investigating active learning in interactive neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 10–22.

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.

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

Wei-Ning Hsu and Hsuan-Tien Lin. 2015. Active learning by learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Junjie Hu and Graham Neubig. 2021a. Phrase-level active learning for neural machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1087–1099, Online. Association for Computational Linguistics.

Junjie Hu and Graham Neubig. 2021b. Phrase-level active learning for neural machine translation. pages 1087–1099.

Seokhwan Kim, Yu Song, Kyungduk Kim, Jeong-Won Cha, and Gary Geunbae Lee. 2006. MMR-based active machine learning for bio named entity recognition. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 69–72, New York City, USA. 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.

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. 2022. Prism: A rich class of parameterized submodular information measures for guided data subset selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10238–10246.

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: Research Track*, pages 96–106, Dublin, Ireland. European Association for Machine Translation.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Begoa Pendas, Andrés Carvallo, and Carlos Aspillaga. 2023. Neural machine translation through active learning on low-resource languages: The case of spanish to mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 6–11.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

Akim Tsvigun, Ivan Lysenko, Danila Sedashov, Ivan Lazichny, Eldar Damirov, Vladimir Karlov, Artemy Belousov, Leonid Sanochkin, Maxim Panov, Alexander Panchenko, Mikhail Burtsev, and Artem Shelmanov. 2022a. Active learning for abstractive text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5128–5152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Akim Tsvigun, Leonid Sanochkin, Daniil Larionov, Gleb Kuzmin, Artem Vazhentsev, Ivan Lazichny, Nikita Khromov, Danil Kireev, Aleksandr Rubashevskii, Olga Shahmatova, Dmitry V. Dylov, Igor Galitskiy, and Artem Shelmanov. 2022b. ALToolbox: A set of tools for active learning annotation of natural language texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 406–434, Abu Dhabi, UAE. Association for Computational Linguistics.

Nicola Ueffing and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Comput. Linguistics*, 33(1):9–40.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yu Xia, Xu Liu, Tong Yu, Sungchul Kim, Ryan Rossi, Anup Rao, Tung Mai, and Shuai Li. 2024. Hallucination diversity-aware active learning for text summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8665–8677, Mexico City, Mexico. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. 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.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806.

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

Zhong Zhou and Alex Waibel. 2021. Active learning for massively parallel translation of constrained text into low resource languages. *arXiv preprint arXiv:2108.07127*.

A Implementation Details

We extend the `ALToolbox`[3] python package (Tsvigun et al., 2022b) and add support for hybrid uncertainty and diversity sampling (HUDS) for active learning (AL). The implementation of HUDS allows using any translation dataset[4] and model[5] available on HuggingFace (Wolf et al., 2019). The code is submitted as part of the supplementary material.

A.1 Models

We use the `BART-base` model publicly available on HuggingFace transformers (Wolf et al., 2019). The HuggingFace repository is available under the Apache License 2.0 license.

A.2 Datasets

We use the WMT14 `En-De` dataset available under the CC-BY-SA license and the German-English multi-domains parallel data available under the Creative Commons CC0 license.

B Hyperparameters

The hyperparameters used in our experiments are reported in Table 8. We use the validation set of the multi-domains dataset (Medicine, IT, and Law) to tune these parameters.

Hyperparameter	Value
learning rate	$2e-5$
train batch size	16
eval batch size	16
num beams	4
pad-to-max-length	true
gradient-clipping	0.28
scheduler warmup-steps-factor	0.1
weight-decay	0.028
gradient-accumulation	0.1
fp-16	true
number of strata	10

Table 8: Hyperparameters for the experiments.

C Resources

The experiments are run on a single 48GB NVIDIA A6000 GPU on the cloud. Around 1680 GPU hours were used for the entirety of the project which includes the early experiments and the final results.

[3] https://github.com/AIRI-Institute/al_toolbox

[4] https://huggingface.co/datasets?task_categories=task_categories:translation

[5] https://huggingface.co/models?pipeline_tag=translation