The Best of Both Worlds: Combining Human and Machine Translations for Multilingual Semantic Parsing with Active Learning

Zhuang Li[1], **Lizhen Qu**[2],
Philip R. Cohen[1], **Raj V. Tumuluri**[1], **Gholamreza Haffari**[1]
[1]Openstream.ai, [2]Monash University
{zhuang.li, phil.cohen, raj, reza.haffari}@openstream.com
lizhen.qu@monash.edu

Abstract

Multilingual semantic parsing aims to leverage the knowledge from the high-resource languages to improve low-resource semantic parsing, yet commonly suffers from the data imbalance problem. Prior works propose to utilize the translations by either humans or machines to alleviate such issues. However, human translations are expensive, while machine translations are cheap but prone to error and bias. In this work, we propose an active learning approach that exploits the strengths of both human and machine translations by iteratively adding small batches of human translations into the machine-translated training set. Besides, we propose novel aggregated acquisition criteria that help our active learning method select utterances to be manually translated. Our experiments demonstrate that an ideal utterance selection can significantly reduce the error and bias in the translated data, resulting in higher parser accuracies than the parsers merely trained on the machine-translated data.

1 Introduction

Multilingual semantic parsing allows a single model to convert natural language utterances from multiple languages into logical forms (LFs). Due to its wide applications in various research areas, e.g. multilingual question answering and multilingual virtual assistant, multilingual semantic parsing has drawn more attention recently (Zou and Lu, 2018; Sherborne et al., 2020; Li et al., 2021a).

Training a multilingual semantic parser (MSP) requires training data from all target languages. However, there is a severe imbalance of data availability among languages for current multilingual semantic parsing research. The utterances in most current semantic parsing datasets are in English, while non-English data is scarce.

To overcome the data imbalance issue, prior studies translate utterances in the MSP datasets from high-resource languages (e.g. English) to the target low-resource languages of interest by either human translators (Susanto and Lu, 2017; Duong et al., 2017; Li et al., 2021a) or automatic machine translation (MT) (Moradshahi et al., 2020; Sherborne et al., 2020). Unfortunately, human translation (HT), though effective, is cost-intensive and time-consuming. While the cost of MTs is much lower than that of HTs, the low quality of the machine-translated utterances severely weakens the performance of the MSPs in the target languages.

We observe that the quality of MTs is lower than that of HTs, mainly due to translation bias and errors. First, MT systems are likely to be influenced by algorithmic bias. Hence, the outputs of MT systems are generally less lexically and morphologically diverse than human translations (Vanmassenhove et al., 2021). So, there is a lexical distribution discrepancy between the machine-translated and the human-generated utterances. Second, MT systems are prone to generate translations with errors (Daems et al., 2017).

Prior study (Moradshahi et al., 2020) demonstrates that adding only a small portion of human-translated data into the complete set of machine-translated training data significantly improves the MSP performance on the test set of the target language. Given this observation, we propose a novel annotation strategy based on active learning (AL) that benefits from both **H**uman translations and **A**utomatic machine **T**ranslations (HAT). It initially machine-translates all utterances in training sets from the high-resource languages to target languages. Then, for each iteration, HAT selects a subset of utterances from the original training set to be translated by human translators, followed by adding the HT data to the MT training data. The multilingual parser is trained on the combination of both types of translated data.

We further investigate how HAT can select utterances whose HTs maximally benefit the parser performance. We assume the performance improve-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9511–9528

ment is ascribed to the less biased and erroneous training set in a mixture of the MT and HT data. We have found that resolving the bias and error issues for the translations of the most semantically diversified and representative utterances improves the parser performance to the greatest extent. Given this assumption, we provide an **A**ggregated acquisition function that scores the utterances on how much their HTs can mitigate the **B**ias and **E**rror issues for learning the multilingual parsers (ABE). It aggregates four individual acquisition functions, two of which measure the error and bias degree for the translations of the source utterances. The other two encourage the selection of the most representative and semantically diversified utterances.

Our key contributions are as follows:

- We propose a novel AL procedure, HAT, that benefits from two popular annotation strategies for training the MSP. HAT greatly boosts the performance of the parser trained on MT data while it requires only a small extra human annotation cost. With only 16% of total utterances translated by humans, the parser accuracies on the multilingual GEOQUERY (Susanto and Lu, 2017) and NLMAP (Haas and Riezler, 2016) test sets can be improved by up to 28% and 5%, respectively, compared to the accuracies of those trained on machine-translated data, and are only up to 5% away from the ORACLE parsers trained on all human data.

- We propose an aggregated acquisition function, coined ABE, specifically designed to select utterances where their HTs mitigate translation bias and error for learning a good MSP. Compared to other SOTA acquisition baselines, given the same selection budget, our experiments consistently show ABE consistently results in the less biased and erroneous training sets and higher parser accuracies on the multilingual GEOQUERY and NLMAP test sets.

2 Related Work

Multilingual Semantic Parsing. Multilingual semantic parser is an emerging field that parses utterances from multiple languages using one model. Almost all the current MSP data are obtained by translating the utterances in existing semantic parsing datasets in the high-resource languages by the automatic translation services (Moradshahi et al., 2020; Sherborne et al., 2020) or human translators (Susanto and Lu, 2017; Duong et al., 2017; Li et al., 2021a; Li and Haffari, 2023). They don't consider conventional data collection strategies (Wang et al., 2015) for monolingual semantic parsing as they require expert knowledge in LFs, which is more expensive than bilingual knowledge. Therefore, our work follows the same strategies to leverage the knowledge from high-resource to low-resource languages. Moradshahi et al. (2020) tries to mix the human-translated data with machine-translated data to improve the parser accuracies. However, their work is only in a supervised learning setting, while our work studies how to iteratively collect utterances in an AL scenario.

Active Learning. AL is to select the most valuable unlabeled instances to be annotated in order to maximize the model's performance and hence reduce the annotation cost for data-hungry machine learning models. AL has been used to MT (Haffari and Sarkar, 2009), sequence labelling (Vu et al., 2019), text classification (McCallum et al., 1998; Vu et al., 2023), and semantic parsing (Duong et al., 2018; Ni et al., 2020; Li and Haffari, 2023). Following most deep learning AL methods (Duong et al., 2018; Ni et al., 2020; Li and Haffari, 2023), our work also adopts a pool-based query strategy, which means we sample batches from a large pool of unlabelled data instead of evaluating examples one by one from an incoming stream. Among all the AL for semantic parsing works, Li and Haffari (2023) is the one most similar to ours, which selects utterances to be translated. However, they do not utilize MT systems.

3 Multilingual Semantic Parsing with Automatic Machine Translation

An MSP is a parametric model $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ that maps a natural language utterance $\boldsymbol{x} \in \mathcal{X}$ into a formal meaning representation $\boldsymbol{y} \in \mathcal{Y}$, where $\mathcal{X} = \bigcup_{l \in L} \mathcal{X}_l$ includes utterances in different languages L. The standard training objective for a multilingual parser is,

$$\arg\max_{\theta} \prod_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{D}_L} P_\theta(\boldsymbol{y}|\boldsymbol{x}) \tag{1}$$

where $\mathcal{D}_L = \bigcup_{l \in L} \mathcal{D}_l$ includes training data where utterances are from multiple languages L.

Metrics	GeoQuery(De)		GeoQuery(Th)		GeoQuery(El)		NLMap(De)	
	HT	MT	HT	MT	HT	MT	HT	MT
Accuracy↑	78.14	47.21	79.29	56.93	80.57	68.5	81.57	67.86
BT Discrepancy Rate↓	2%	11%	3%	12%	3%	10%	2%	10%
JS↓	36.67	59.95	32.02	73.83	33.67	56.36	33.78	46.84
MAUVE↑	96.01	22.37	97.52	8.48	97.12	45.01	97.34	70.24
MTLD↑	26.02	22.50	20.74	19.07	28.16	27.08	44.80	42.38

Table 1: The scores of five metrics to measure the quality of the HTs and MTs in German (De), Thai (Th) and Greek (El) of the utterances in GeoQuery and NLMap. ↑/↓ means the higher/lower score the better. See **Evaluation** in Sec. 5 for the details of Accuracy, MTLD, JS, MAUVE and BT Discrepancy Rate

3.1 Difficulties for Multilingual Semantic Parsing Utilizing Machine Translation

Although using an MT system to train an MSP is cost-effective, the parser performance is usually much lower than the one trained with human-translated data. For example, as shown in Table 1, the parsers trained on HTs all have significantly higher accuracies than those trained on MTs in different settings. Such performance gaps are due to two major issues of the MT data, discussed below.

Translation Bias. Many existing MT systems amplify biases observed in the training data (Vanmassenhove et al., 2021), leading to two problems that degrade the parsers' performance trained on MT data:

- The MTs lack lexical diversity (Vanmassenhove et al., 2021). As shown in Table 1, MTLD (Vanmassenhove et al., 2021) values show that the HTs of utterances in multilingual GeoQuery and NLMap are all more lexically diversified than MTs. Several studies (Shiri et al., 2022; Xu et al., 2020; Wang et al., 2015; Zhuo et al., 2023; Huang et al., 2021) indicate that lexical diversity of training data is essential to improving the generalization ability of the parsers.

- The lexical distribution of the biased MTs is different to the human-written text. The two metrics, Jensen–Shannon (JS) divergence (Manning and Schutze, 1999) and MAUVE (Pillutla et al., 2021), in Table 1 show the HTs of utterances in GeoQuery and NLMap are more lexically close to the human-generated test sets than MTs.

Translation Error. MT systems often generate translation errors due to multiple reasons, such as underperforming MT models or an absence of contextual understanding (Wu et al., 2023; Wu et al.), leading to discrepancies between the source text and its translated counterpart. One common error type is mistranslation (Vardaro et al., 2019), which alters the semantics of the source sentences after translation. Training an MSP on the mistranslated data would cause incorrect parsing output, as LFs are the semantic abstraction of the utterances. BT Discrepancy Rate in Table 1 demonstrates the mistranslation problem is more significant in the machine-translated datasets.

4 Combining Human and Automatic Translations with Active Learning

To mitigate the negative effect of translation bias and error in the MT data, we propose HAT, which introduces extra human supervision to machine supervision when training the MSPs. Two major intuitions motivate our training approach:

- Adding the HTs to the training data could enrich its lexical and morphological diversity and ensure that the lexical distribution of the training data is closer to the human test set, thus improving the parsers' generalization ability (Shiri et al., 2022; Xu et al., 2020; Wang et al., 2015).

- HTs are less erroneous than MTs (Freitag et al., 2021). The parser could learn to predict correct abstractions with less erroneous training data.

Our HAT AL setting considers only the *bilingual* scenario. One of the languages is in high-resource, and the other one is in low-resource. However, it is easy to extend our method to more than two languages. We assume access to a well-trained black-box multilingual MT system, $g^{mt}(\cdot)$, and a semantic parsing training set that includes utterances in a high-resource language l_s (e.g. English) paired with LFs, $\mathcal{D}_s = \{(\boldsymbol{x}_s^i, \boldsymbol{y}^i)\}_{i=1}^N$, two human-generated test sets $\mathcal{T}_s = \{(\boldsymbol{x}_s^i, \boldsymbol{y}^i)\}_{i=1}^M$ and $\mathcal{T}_t = \{(\boldsymbol{x}_t^i, \boldsymbol{y}^i)\}_{i=1}^M$ with utterances in high and low-resource languages, respectively. Each utterance $\boldsymbol{x}_s$ in $\mathcal{D}_s$ is translated into the utterance $\hat{\boldsymbol{x}}_t = g_{s\to t}^{mt}(\boldsymbol{x}_s)$ in the target language l_t by the MT system, $\hat{\mathcal{D}}_t = \{(\hat{\boldsymbol{x}}_t^i, \boldsymbol{y}^i)\}_{i=1}^N$. The goal of our AL method is to select an optimal set of utterances from the training data in the source language, $\tilde{\mathcal{D}}_s \in \mathcal{D}_s$, and ask human translators to translate them into the target language, denoted by $\bar{\mathcal{D}}_t = g_{s\to t}^{ht}(\tilde{\mathcal{D}}_s)$, for training a semantic parser on the union of $\bar{\mathcal{D}}_t$ and $\hat{\mathcal{D}}_t$. The selection criterion is based on the *acquisi-*

tion functions that score the source utterances. Following the conventional batch AL setting (Duong et al., 2018), there are Q selection rounds. At the qth round, AL selects utterances with a budget size of K_q.

The detailed HAT AL procedure iteratively performs the following steps as in Algorithm. 1.

Algorithm 1: HAT procedure

Input : Initial training set $\mathcal{D}^0 = \mathcal{D}_s \cup \hat{\mathcal{D}}_t$, source utterance pool $\mathcal{D}_s$, budget size K_q, number of selection rounds Q, human annotators $g^{ht}(\cdot)$

Output : A well-trained multilingual parser $P_\theta(\boldsymbol{y}|\boldsymbol{x})$

 # Train the initial parser on the initial data
Update θ of $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ with $\nabla_\theta \mathcal{L}(\theta)$ on $\mathcal{D}^0$
Evaluate $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ on $\mathcal{T}_s$ and $\mathcal{T}_t$
Estimate the acquisition function $\phi(\cdot)$
$\bar{\mathcal{D}}_t^0 = \emptyset$ *# Empty set of human-translated data*
$\bar{\mathcal{D}}_s^0 = \mathcal{D}_s$ *# Initial source utterance pool*
for $q \leftarrow 1$ *to* Q **do**
    # Select a subset $\tilde{\mathcal{D}}_s^q \in \bar{\mathcal{D}}_s^{q-1}$ of the size K_q with the highest scores ranked by the acquisition function $\phi(\cdot)$
    $\tilde{\mathcal{D}}_s^q = \text{TopK}(\phi(\bar{\mathcal{D}}_s^{q-1}), K_q)$
    $\bar{\mathcal{D}}_s^q = \bar{\mathcal{D}}_s^{q-1} \setminus \tilde{\mathcal{D}}_s^q$
    # Translate the utterances in $\tilde{\mathcal{D}}_s^q$ into the target language l_t by human annotators
    $\mathcal{D}_t^q = g^{ht}_{s\rightarrow t}(\tilde{\mathcal{D}}_s^q)$
    # Merge all human-translated data
    $\bar{\mathcal{D}}_t^q = \bar{D}_t^{q-1} \cup D_t^q$
    # Add the human-translated data into the training data
    $\mathcal{D}^q = \mathcal{D}_s \cup \hat{\mathcal{D}}_t \cup \bar{\mathcal{D}}_t^q$
    # Train the parser on the updated data
    Update θ of $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ with $\nabla_\theta \mathcal{L}(\theta)$ on $\mathcal{D}^q$
    Evaluate $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ on $\mathcal{T}_s$ and $\mathcal{T}_t$
    Re-estimate $\phi(\cdot)$
end

4.1 Acquisition Functions

The acquisition functions assign higher scores to those utterances whose HTs can boost the parser's performance more than the HTs of the other utterances. The prior AL works (Sener and Savarese, 2018; Zhdanov, 2019; Nguyen and Smeulders, 2004) suggest that the most representative and diversified examples in the training set improve the generalization ability of the machine learning models the most. Therefore, we provide a hypothesis that *we should select the representative and diversified utterances in the training set, whose current translations have significant bias and errors*. We postulate fixing problems of such utterances improves the parsers' performance the most. We derive four acquisition functions based on this hypothesis to score the utterances. Then, ABE aggregates these acquisition functions to gain their joint benefits. In each AL round, the utterances with the highest ABE scores are selected.

Translation Bias. We assume an empirical conditional distribution, $P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s)$, for each utterance $\boldsymbol{x}_s$ in $\mathcal{D}_s$ at qth AL selection round. Intuitively, the $\boldsymbol{x}_s$ with the most biased translations should be the one with the most skewed empirical conditional distribution. Therefore, we measure the translation bias by calculating the entropy of the empirical conditional distribution, $H(P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s))$, and select the $\boldsymbol{x}_s$ with the lowest entropy. Since the translation space $\mathcal{X}_t$ is exponentially large, it is intractable to directly calculate the entropy. Following (Settles and Craven, 2008), we adopt two approximation strategies, *N-best Sequence Entropy* and *Maximum Confidence Score*, to approximate the entropy.

- *N-best Sequence Entropy:*

$$\phi_b(\boldsymbol{x}_s) = -\sum_{\hat{\boldsymbol{x}}_t \in \mathcal{N}} \hat{P}_e^q(\hat{\boldsymbol{x}}_t|\boldsymbol{x}_s) \log \hat{P}_e^q(\hat{\boldsymbol{x}}_t|\boldsymbol{x}_s) \quad (2)$$

where $\mathcal{N} = \{\hat{\boldsymbol{x}}_t^1, ..., \hat{\boldsymbol{x}}_t^N\}$ are the N-best hypothesis sampled from the empirical distribution $P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s)$. $\hat{P}_e^q(\hat{\boldsymbol{x}}_t|\boldsymbol{x}_s)$ is re-normalized from $P_e^q(\hat{\boldsymbol{x}}_t|\boldsymbol{x}_s)$ over $\mathcal{N}$, which is only a subset of $\mathcal{X}_t$.

- *Maximum Confidence Score (MCS)*:

$$\phi_b(\boldsymbol{x}_s) = \log P_e^q(\boldsymbol{x}_t'|\boldsymbol{x}_s) \quad (3)$$

$$s.t. \boldsymbol{x}_t' = \arg\max_{\boldsymbol{x}_t} P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s) \quad (4)$$

It is difficult to obtain the empirical distribution as we know neither of the two distributions that compose the empirical distribution. Therefore, we use distillation training (Hinton et al.) to train a translation model that estimates $P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s)$ on all the bilingual pairs $(\boldsymbol{x}_s, \boldsymbol{x}_t)$ in the MSP training data $\mathcal{D}^q$. Another challenge is that $\mathcal{D}^q$ is too small to distil a good translation model that imitates the mixture distribution. Here, we apply a bayesian factorization trick that factorizes $P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s) = \sum_{\boldsymbol{y}\in\mathcal{Y}} P_e^q(\boldsymbol{x}_t|\boldsymbol{y}) P_e^q(\boldsymbol{y}|\boldsymbol{x}_s)$, where $\boldsymbol{y}$ ranges over LFs representing the semanics. As there is a deterministic mapping between $\boldsymbol{x}_s$ and the LF, $P_e^q(\boldsymbol{y}|\boldsymbol{x}_s)$ is an one-hot distribution. Thus, we only need to estimate the entropy, $H(P_e^q(\boldsymbol{x}_t|\boldsymbol{y}))$. This has a nice intuition: the less diversified data has less lexically diversified utterances per each LF. Note that if we use this factorization, all $\boldsymbol{x}_s$ that share the same LF have the same scores.

We use the lightweight, single-layer, recurrent neural network-based Seq2Seq model to estimate $P_e^q(X_t|\boldsymbol{x}_s)$ or $P_e^q(\boldsymbol{x}_t|\boldsymbol{y})$. It only takes approximately 30 seconds to train the model on GEOQUERY. Ideally, every time a new source utterance $\boldsymbol{x}_s$ is selected, $P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s)$ should be re-estimated. However, we only re-estimate $P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s)$ once at the beginning of each selection round to reduce the training cost.

Translation Error. Similar to Haffari et al. (2009), we leverage back-translations (BTs) to measure the translation error. We conjecture that if the translation quality for one source utterance $\boldsymbol{x}_s$ is good enough, the semantic parser should be confident in the LF of the source utterance conditioned on its BTs. Therefore, we measure the translation error for each $\boldsymbol{x}_s$ as the expected parser's negative log-likelihood in its corresponding LF $\boldsymbol{y}_{\boldsymbol{x}_s}$ over all the BTs of $\boldsymbol{x}_s$, $\mathbb{E}_{P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s)}[-\log(P_\theta^q(\boldsymbol{y}_{\boldsymbol{x}_s}|g_{t\to s}^{mt}(\boldsymbol{x}_t)))]$, where P_θ^q is the parser trained at qth round. To approximate the expectation, we apply two similar strategies as mentioned in *Translation Bias*.

• *N-best Sequence Expected Error:*

$$\phi_e(\boldsymbol{x}_s) = -\sum_{\hat{\boldsymbol{x}}_t \in \mathcal{N}_{\boldsymbol{y}_{\boldsymbol{x}_s}}} \hat{P}_e^q(\hat{\boldsymbol{x}}_t|\boldsymbol{x}_s) \log P_\theta(\boldsymbol{y}_{\boldsymbol{x}_s}|g_{t\to s}^{mt}(\boldsymbol{x}_t)) \tag{5}$$

where $\mathcal{N}_{\boldsymbol{y}_{\boldsymbol{x}_s}}$ is the set of translations in $\mathcal{D}^q$ that share the same LF $\boldsymbol{y}_{\boldsymbol{x}_s}$ with $\boldsymbol{x}_s$. We only back-translate utterances in $\mathcal{D}^q$ to reduce the cost of BTs.

• *Maximum Error:*

$$\phi_e(\boldsymbol{x}_s) = -\log P_\theta^q(\boldsymbol{y}_{\boldsymbol{x}_s}|g_{t\to s}^{mt}(\boldsymbol{x}'_t)) \tag{6}$$

$$s.t. \boldsymbol{x}'_t = \arg\max_{\boldsymbol{x}_t} P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s) \tag{7}$$

We use the same distilled translation model $P_e^q(\boldsymbol{x}_t|\boldsymbol{x}_s)$ used in *Translation Bias*.

Semantic Density. The previous AL works (Nguyen and Smeulders, 2004; Donmez et al., 2007) have found that the most *representative* examples improve the model performance the most. Therefore we desire to reduce the translation error and bias for the translations of the most representative source utterances. As such, the utterances should be selected from the dense regions in the semantic space,

$$\phi_s(\boldsymbol{x}_s) = \log P(\boldsymbol{x}_s). \tag{8}$$

We use kernel density estimation (Botev et al., 2010) with the exponential kernel to estimate $P(\boldsymbol{x}_s)$, while other density estimation methods could be also used. The feature representation of $\boldsymbol{x}_s$ for density estimation is the average pooling of the contextual sequence representations from the MSP encoder. The density model is re-estimated at the beginning of each query selection round.

Semantic Diversity. The role of the semantic diversity function is twofold. First, it prevents the AL method from selecting similar utterances. Resolving the bias and errors of similar utterances in a small semantic region does not resolve the training issues for the overall dataset. Second, semantic diversity correlates with the lexical diversity, hence improving it also enriches lexical diversity.

$$\phi_d(\boldsymbol{x}_s) = \begin{cases} 0 & \text{if } c(\boldsymbol{x}_s) \notin \bigcup_{\boldsymbol{x}_s^i \in \mathcal{S}} c(\boldsymbol{x}_s^i) \\ -\infty & \text{Otherwise} \end{cases} \tag{9}$$

where $c(\boldsymbol{x}_s)$ maps each utterance $\boldsymbol{x}_s$ into a cluster id and $\mathcal{S}$ is the set of cluster ids of the selected utterances. We use a clustering algorithm to diversify the selected utterances as in (Ni et al., 2020; Nguyen and Smeulders, 2004). The source utterances are partitioned into $|\mathcal{C}|$ clusters. We select one utterance at most from each cluster. Notice the number of clusters should be greater than or equal to the total budget size until current selection round, $|\mathcal{C}| \geq \sum_{i=1}^{q} K_i$. The clusters are re-estimated every round. To ensure the optimal exploration of semantic spaces across different query rounds, we adopt Incremental K-means (Liu et al., 2020) as the clustering algorithm. At each new round, Incremental K-means considers the selected utterances as the fixed cluster centres, and learns the new clusters conditioned on the fixed centres. The feature representation of $\boldsymbol{x}_s$ for Incremental K-means is from MSP encoder as well.

Aggregated Acquisition. We aggregate the four acquisition functions into one,

$$\phi_A(\boldsymbol{x}_s) = \sum_k \alpha_k \phi_k(\boldsymbol{x}_s)$$

where α_k's are the coefficients. Each $\phi_k(\boldsymbol{x}_s)$ is normalized using quantile normalization (Bolstad et al., 2003). Considering the approximation strategies we employ for both *Translation Bias* and *Translation Error*, ABE can be denoted as either ABE(N-BEST) or ABE(MAX). The term ABE(N-BEST) is used when we apply *N-best Sequence*

Entropy and *N-best Sequence Expected Error*. On the other hand, ABE(MAX) is used when we implement *Maximum Confidence Score* and *Maximum Error* strategies.

5 Experiments

Datasets. We evaluate our AL method for MSP on two datasets, GEOQUERY (Susanto and Lu, 2017) and NLMAP (Haas and Riezler, 2016) with multilingual human-translated versions. GEOQUERY includes 600 utterances-LF pairs as the training set and 280 pairs as the test set. NLMAP includes 1500 training examples and 880 test examples.

In our work, we consider English as the *resource-rich* source language and use Google Translate System[1] to translate all English utterances in GEOQUERY into German (De), Thai (Th), Greek (El) and the ones in NLMAP into German, respectively. The AL methods actively sample English utterances, the HTs of which are obtained from the multilingual GEOQUERY and NLMAP.

Active Learning Setting. The HAT active learning procedure performs five rounds of query, which accumulatively samples 1%, 2%, 4%, 8% and 16% of total English utterances in GEOQUERY and NLMAP. We only perform five rounds as we found the performance of the multilingual parser is saturated after sampling 16% of examples with most acquisition functions.

Base Parser. We use BERT-LSTM as our multilingual parser (Moradshahi et al., 2020). It is a Seq2Seq model with the copy mechanism (Gu et al., 2016) that applies Multilingual BERT-base (Devlin et al., 2018) as the encoder and LSTM (Hochreiter and Schmidhuber, 1997) as the decoder.

Baselines. We compare ABE with eight acquisition baselines and an oracle baseline.

1. **Random** randomly selects English utterances in each round.

2. **Cluster** (Ni et al., 2020; Li et al., 2021b) partitions the utterances into different groups using K-means and randomly selects one example from each group.

3. **LCS (FW)** (Duong et al., 2018) selects English utterances for which the parser is least confident in their corresponding LFs, $\boldsymbol{x} = \arg\min_{\boldsymbol{x}} p_\theta(\boldsymbol{y}|\boldsymbol{x})$.

4. **LCS (BW)** (Duong et al., 2018), on the opposite of LCS (BW), trains a text generation model to generate text given the LF. The English utterances are selected for which the text generation model is least confident conditioned on their corresponding LFs, $\boldsymbol{x} = \arg\min_{\boldsymbol{x}} p_\theta(\boldsymbol{x}|\boldsymbol{y})$.

5. **Traffic** (Sen and Yilmaz, 2020) selects utterances with the lowest perplexity and highest frequency in terms of their corresponding LFs.

6. **CSSE** (Hu and Neubig, 2021) combines the density estimation and the diversity estimation metrics to select the most representative and semantically diversified utterances.

7. **RTTL** (Haffari et al., 2009; Haffari and Sarkar, 2009) uses BLEU (Papineni et al., 2002) to estimate the translation information losses between the BTs and the original utterances to select utterances with highest losses.

8. **LFS-LC-D** (Li and Haffari, 2023) is the selection method for MSP, which enriches the diversity of lexicons and LF structures in the selected examples.

9. **ORACLE** trains the parser on the combination of English data, machine-translated data, and the complete set of human-translated data.

Evaluation. We evaluate the AL methods by measuring the accuracy of the MSP, the bias of the training set, and the semantic discrepancy rate between the selected utterances and their BTs.

- **Accuracy:** To evaluate the performance of the MSP, we report the accuracy of exactly matched LFs as in (Dong and Lapata, 2018) at each query round. As the parser accuracies on the English test sets are not relevant to evaluating the active learning method, we only report the accuracies on the test sets in the *target* languages. See Appendix A.2 for the English results.

- **Bias of the Training Set:** We use three metrics to measure the bias of the training data in the target language at each query round.

[1] https://translate.google.com/

1. *Jensen–Shannon (JS) divergence* (Pillutla et al., 2021) measures the JS divergence between the n-gram frequency distributions of the utterances in the training set $\hat{\mathcal{D}}_t \cup \bar{\mathcal{D}}_t^q$ generated by each AL method and test set $\mathcal{T}_t$.
2. *MAUVE* (Pillutla et al., 2021) compares the learnt distribution from the training set to the distribution of human-written text in the test set $\mathcal{T}_t$ using Kullback–Leibler divergence (Kullback and Leibler, 1951) frontiers. Here we use n-gram lexical features from the text when calculating MAUVE. *JS* and *MAUVE* together measure how lexically "human-like" the generated training set is.
3. *MTLD* (McCarthy, 2005) reports the mean length of word strings in the utterances in $\hat{\mathcal{D}}_t \cup \bar{\mathcal{D}}_t^q$ that maintain a given TTR (Templin, 1957) value, where TTR is the ratio of different tokens to the total number of tokens in the training data. *MTLD* evaluate the lexical diversity of the training set.

- **BT Discrepancy Rate:** Since we do not possess bilingual knowledge, we use BT to access the translation quality (Tyupa, 2011). At each query round, we randomly sample 100 utterances from the utterances selected by each acquisition in 5 seeds' experiments. The BT is obtained by using Google Translation to translate the MTs of the 100 sampled utterances back to English. Two annotators manually check the percentage of the BTs which are not semantically equivalent to their original utterances. We only consider a BT discrepancy when both annotators agree. Ideally, the utterances with fewer mistranslations would see fewer semantic discrepancies between the BTs and the original.

5.1 Main Results and Discussion.

Effectiveness of HAT. Fig. 1 shows that HAT significantly improves the parser accuracies on all test sets by adding only a small amount of HTs into the machine-translated training set. For example, with 16% of English utterances translated by humans, HAT improves the parser accuracies by up to 28% and 25%, respectively, on GEOQUERY(DE) and GEOQUERY(TH) test sets. On the other hand, on GEOQUERY(EL) and NLMAP(DE) test sets, the accuracy improvement by HAT is only up to 5% because the parser has already achieved a decent performance after it is trained on the MT data. According to Table 1, we speculate that the training sets of GEOQUERY(EL) and NLMAP(DE) are less biased than those of GEOQUERY(TH) and GEOQUERY(DE). Overall for all dataset settings, if we apply HAT with ABE, the multilingual parsers can perform comparably to the ORACLE parsers with no more than 5% differences in terms of accuracies at an extra expense of manual translating 16% of English utterances.

Effectiveness of ABE. The ABE method has been demonstrated to consistently achieve superior performance over the baselines by utilizing a combination of four important measurements. In contrast, the acquisition baselines focus on only one or two of these measurements, and as a result, they fail to effectively address issues of bias and error across all datasets and languages. Despite this, these baselines may still perform well in certain specific settings, such as LFS-LC-D performing slightly better than ABE on the GEOQUERY(TH) dataset. However, it should be noted that this performance is not consistent across all settings. Three baselines, LCS(FW), LCS(BW), and RTTL, consistently perform lower than the others. LCS(FW) tends to select similar examples, which lack semantic diversity. RTTL is designed to choose the utterances with the most erroneous translations, while such utterances are mostly the tail examples given our inspection. ABE overcomes this issue by balancing the *Translation Error* term with the *Semantic Density*. LCS(BW) has an opposite objective with our *Translation Bias*, which means it selects the utterances with the most translation bias. Therefore, though LCS(BW) performs well in the AL scenario in Duong et al. (2018) for semantic parsing, it performs worst in our scenario.

Bias, Error and Parser Performance. As in Table 2, we also measure the bias of the training set and the BT discrepancy rates of the selected utterances at the final selection round for each acquisition function. We can see that the parser accuracy directly correlates with the training set's bias degree. The bias of the training set acquired by RANDOM, TRAFFIC and CLUSTER, LFS-LC-D, and ABE score better in general than the other baselines in terms of the bias metrics, resulting in a better parser accuracy. RTTL and LCS(FW) that

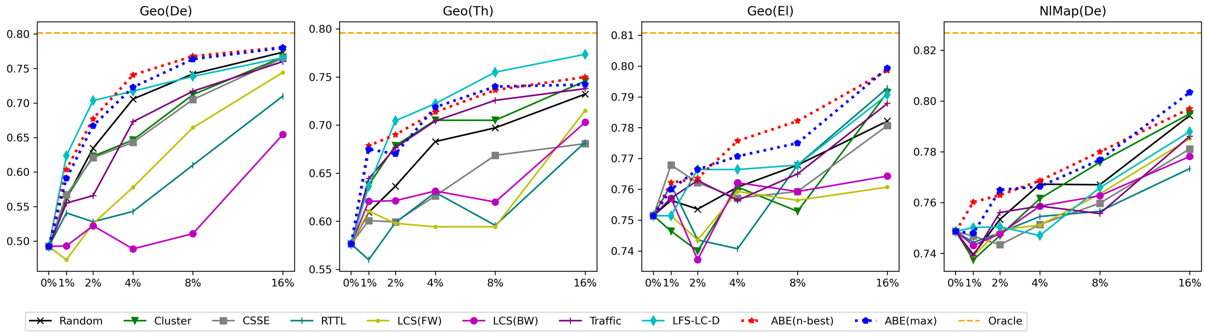

Figure 1: The parser accuracies at different query rounds using various acquisitions on the test sets of GEOQUERY(DE), GEOQUERY(TH), GEOQUERY(EL) and NLMAP(DE). Orange dash lines indicate the accuracies of ORACLE multilingual parsers. All experiments are run 5 times with a different seed for each run.

Metric	No HT	RANDOM	CLUSTER	CSSE	RTTL	LCS(FW)	LCS(BW)	TRAFFIC	LFS-LC-D	ABE(N-BEST)	ABE(MAX)	ORACLE
BT Discrepancy Rate	-	11%	14%	11%	21%	**22%**	8%	14%	10%	17%	18%	-
JS↓	59.95	54.15	54.71	55.53	54.56	54.38	56.13	54.58	54.26	54.16	**53.97**	45.12
MAUVE↑	22.37	36.99	36.12	34.52	35.53	31.61	29.75	35.67	36.87	**38.96**	35.13	73.04
MTLD↑	22.50	23.79	23.32	22.65	22.89	23.00	22.27	23.42	**23.97**	23.80	23.78	24.23

Table 2: Using different acquisitions at the final query round, we depict the scores of the metrics to measure the bias of the training sets in GEOQUERY(DE) and the BT discrepancy rates of the total selected utterances.

select utterances with more erroneous translations do not necessarily guarantee better accuracies for parsers. Our following ablation study shows that the parser performance can be improved by correcting the translation errors for the most representative utterances.

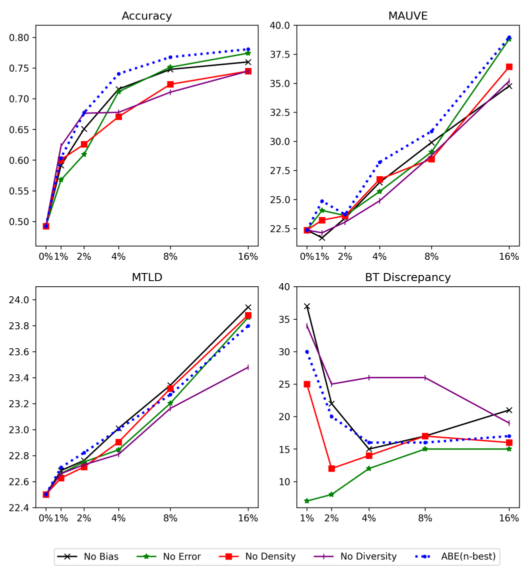

Figure 2: Using ABE(N-BEST) with each term removed, we depict the parser accuracies on the GEOQUERY(DE) test set, the MAUVE and MTLD scores of the GEOQUERY(DE) training sets and the BT Discrepancy rate of the selected utterances from English GEOQUERY at each query round.

5.2 Ablation Study.

Influence of different Acquisition Functions. As in Fig. 2, we evaluate the effectiveness of each acquisition by observing how removing each acquisition from ABE(N-BEST) influences the parser performance, the bias of the training set and the BT Discrepancy rate of the selected utterances. We can see that removing all terms degrades the parser performance. However, each acquisition contributes to the parser accuracy due to different reasons.

Translation Bias and Semantic Diversity contribute to the parser performance mainly due to alleviating the bias of the training set. Excluding Translation Bias does not influence the lexical diversity, while the lexical similarity between the training and test sets becomes lower. Removing Semantic Diversity drops the lexical similarity as well. But it more seriously drops the lexical diversity when the sampling rates are high.

Removing Translation Error significantly decreases the parser accuracy and BT Discrepancy rate in the low sampling regions. However, when the selection rate increases, gaps in parser accuracies and BT Discrepancy rates close immediately. Translation Error also reduces the bias by introducing correct lexicons into the translations.

Removing Semantic Density also drops the parser performance as well. We inspect that Semantic Density contributes to parser accuracy mainly by combing with the Translation Error term. As in

Appendix A.3, using Translation Error or Semantic Density independently results in inferior parser performance. We probe that Translation Error tends to select tail utterances from the sparse semantic region given the TSNE (Van der Maaten and Hinton, 2008) plots at Appendix A.7.

Influence of MT Systems. As in Fig. 3 (Right), at all query rounds, the multilingual parsers perform better with MT data in the training set, showing that MT data is essential for improving the parser's performance when a large number of HTs is not feasible. The quality of the MT data also significantly influences the parser performance when having no HT data in the training set. The accuracy difference between the parsers using Google and Bing translated data is greater than 10% when active learning has not been performed. However, after obtaining the HT data by HAT, the performance gaps close immediately, although the MT data of better quality brings slightly higher performance. When having all utterances translated by humans, the performance differences between parsers with different MT systems can be negligible.

Fig. 3 also demonstrates that ABE(N-BEST) outperforms RANDOM, a strong acquisition baseline, with all three different MT systems. ABE(N-BEST) is also more robust to the MT systems than RANDOM. The performance gaps for the parsers with ABE(N-BEST) are much smaller than those with RANDOM when applying different MT systems.

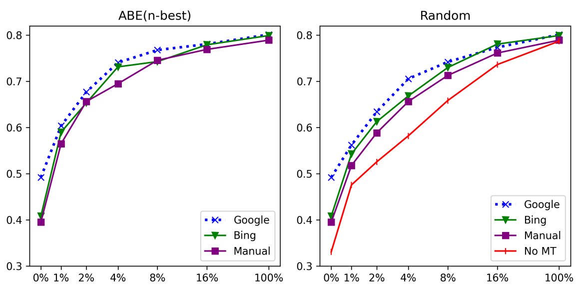

Figure 3: The parser accuracies across various query rounds on the GEOQUERY(DE) test set. We use two selection methods: ABE(N-BEST) (shown on the left) and RANDOM (shown on the right). For each method, we use data from different MT systems - Google, Bing, and our bespoke manually trained MT system. This manual MT system was developed without any pretraining weight, utilizing a limited set of bilingual data. The model architecture was based on the framework proposed by Ott et al. (2018). In addition to these, we also conducted tests without utilizing any MT data.

6 Conclusion

We have tackled the problem of data imbalance when adapting an MSP to a low-resource language. We presented methods to efficiently collect a small amount of human-translated data to reduce bias and error in the training data, assuming a realistic scenario with an MT system and budget constraints for human annotation. Our experiments show that by manually translating only 16% of the dataset, the parser trained on this mixed data outperforms parsers trained solely on machine-translated data and performs similarly to the parser trained on a complete human-translated set.

Limitations

One of the limitations is the selection of hyperparameters. At present, we determine the optimal hyperparameters based on the performance of the selection methods on an existing bilingual dataset. For example, to identify the appropriate utterances to be translated from English to German, we would adjust the hyperparameters based on the performance of the methods on existing datasets in English and Thai. However, this approach may not always be feasible as such a dataset is not always available, and different languages possess distinct characteristics. As a result, the process of tuning hyperparameters on English-Thai datasets may not guarantee optimal performance on English-German datasets. As a future direction, we intend to investigate and develop more effective methods for hyperparameter tuning to address this limitation.

Acknowledgement

I would like to extend my sincere gratitude to Minghao Wu for his invaluable assistance in building the manual MT systems. I am equally grateful to both Minghao Wu and Thuy-Trang Vu for their insightful comments and suggestions during the preparation of this paper.

References

Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Zdravko I Botev, Joseph F Grotowski, and Dirk P Kroese. 2010. Kernel density estimation via diffusion. *The annals of Statistics*, 38(5):2916–2957.

Joke Daems, Sonia Vandepitte, Robert J Hartsuiker, and Lieve Macken. 2017. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in psychology*, 8:1282.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–742.

Pinar Donmez, Jaime G Carbonell, and Paul N Bennett. 2007. Dual strategy active learning. In *European Conference on Machine Learning*, pages 116–127. Springer.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip R Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 379–389.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip R Cohen, and Mark Johnson. 2018. Active learning for deep semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 43–48.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL (1)*.

Carolin Haas and Stefan Riezler. 2016. A corpus and semantic parser for multilingual natural language querying of openstreetmap.

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.

Gholamreza Haffari and Anoop Sarkar. 2009. Active learning for multilingual statistical machine translation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 181–189. The Association for Computer Linguistics.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Shuo Huang, Zhuang Li, Lizhen Qu, and Lei Pan. 2021. On robustness of neural semantic parsers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3333–3342.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021a. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962.

Zhuang Li and Gholamreza Haffari. 2023. Active learning for multilingual semantic parser. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 621–627.

Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2021b. Total recall: a customized continual learning method for neural semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3816–3831.

Juncheng Liu, Yiwei Wang, Bryan Hooi, Renchi Yang, and Xiaokui Xiao. 2020. Active learning for node classification: The additional learning ability from unlabelled nodes. *arXiv preprint arXiv:2012.07065*.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Andrew McCallum, Kamal Nigam, et al. 1998. Employing em and pool-based active learning for text classification. In *ICML*, volume 98, pages 350–358. Madison.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing open-ontology qa semantic parsers in a day using machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983.

Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79.

Ansong Ni, Pengcheng Yin, and Graham Neubig. 2020. Merging weak and active supervision for semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8536–8543.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34.

Priyanka Sen and Emine Yilmaz. 2020. Uncertainty and traffic-aware active learning for semantic parsing. In *Proceedings of the First Workshop on Interactive and Executable Semantic Parsing*, pages 12–17.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

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

Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a crosslingual semantic parser. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 499–517.

Fatemeh Shiri, Terry Yue Zhuo, Zhuang Li, Van Nguyen, Shirui Pan, Weiqing Wang, Reza Haffari, and Yuan-Fang Li. 2022. Paraphrasing techniques for maritime qa system. *arXiv preprint arXiv:2203.10854*.

Raymond Hendy Susanto and Wei Lu. 2017. Neural architectures for multilingual semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–44.

Mildred C Templin. 1957. *Certain language skills in children: Their development and interrelationships*, volume 10. JSTOR.

Sergiy Tyupa. 2011. A theoretical framework for back-translation as a quality assessment tool. *New Voices in Translation Studies*, 7(1):35–46.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Jennifer Vardaro, Moritz Schaeffer, and Silvia Hansen-Schirra. 2019. Translation quality and error recognition in professional neural machine translation post-editing. In *Informatics*, volume 6, page 41. Multidisciplinary Digital Publishing Institute.

Thuy Vu, Ming Liu, Dinh Phung, and Gholamreza Haffari. 2019. Learning how to active learn by dreaming. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 4091–4101.

Thuy-Trang Vu, Shahram Khadivi, Dinh Phung, and Gholamreza Haffari. 2023. Active continual learning: Labelling queries in a sequence of tasks. *arXiv preprint arXiv:2305.03923*.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1332–1342.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462.

Minghao Wu, Lizhen Qu, George Foster, and Gholamreza Haffari. Improving document-level neural machine translation with discourse features. *Available at SSRN 4330827*.

Silei Xu, Sina Semnani, Giovanni Campagna, and Monica Lam. 2020. Autoqa: From databases to qa semantic parsers with only synthetic training data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 422–434.

Fedor Zhdanov. 2019. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*.

Terry Yue Zhuo, Zhuang Li, Yujin Huang, Fatemeh Shiri, Weiqing Wang, Gholamreza Haffari, and Yuan-Fang Li. 2023. On robustness of prompt-based semantic parsing with large pre-trained language model:

An empirical study on codex. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1090–1102.

Yanyan Zou and Wei Lu. 2018. Learning cross-lingual distributed logical representations for semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 673–679.

A Appendix

A.1 Sensitivity Analysis

Fig. 4 shows the parser results on GEOQUERY(DE) test sets when we apply different coefficients or cluster sizes to the four independent acquisitions in ABE(N-BEST). As we can see, tuning the parameters on an existing bilingual dataset does not necessarily bring optimal parser performance, indicating there is still potential in our approach if we can find suitable hyperparameter tuning methods. Another finding is that the ABE(N-BEST) is robust to the hyperparameters changes. Although changing the weights or cluster sizes for each term could influence the parser performances, the parser accuracies do not drop significantly. In addition, we have found that the Semantic Density and Semantic Diversity are more critical to ABE(N-BEST) as there are more fluctuations in the parser accuracies when we adjust the parameters of Semantic Density and Semantic Diversity.

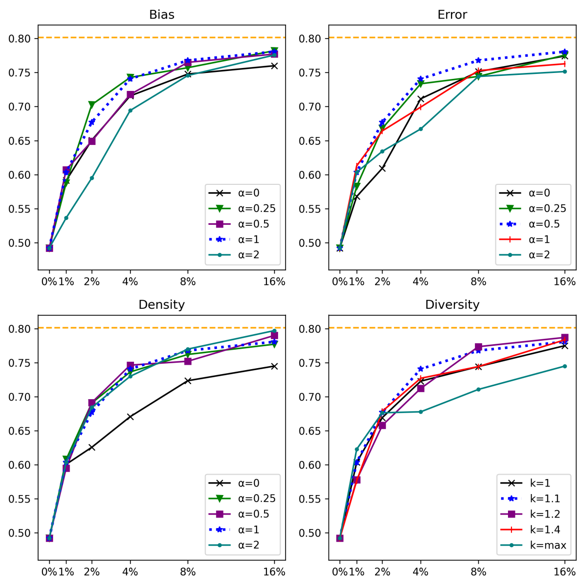

Figure 4: The parser accuracies across various query rounds on the GEOQUERY(DE) test set by employing the ABE(N-BEST) method. This method incorporates varying coefficients, denoted by α, for each term. In addition to this, we also analyze the influence of the proportional rate, denoted by k, which represents the number of clusters in proportion to the budget size at each round.

A.2 Parser Accuracies on English Test Sets

Fig. 5 shows the parser accuracies on the English test sets in different dataset settings. As we can see, the behaviours of the acquisition, ABE(N-BEST), on the target languages do not influence the performance of parsers on the source languages.

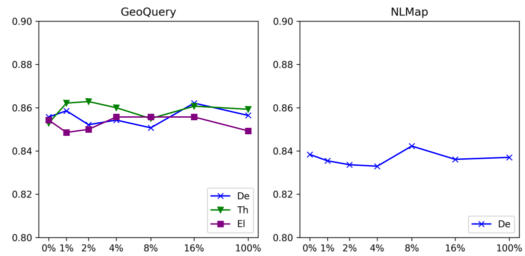

Figure 5: The accuracies on the English test sets after training the parsers on the training sets of GEOQUERY(DE), GEOQUERY(TH), GEOQUERY(EL) and NLMAP(DE) acquired by ABE(N-BEST) at different query rounds.

A.3 Ablation Study of Single Terms

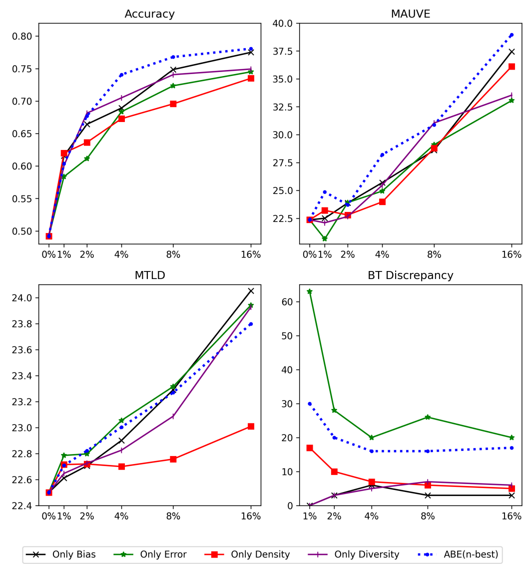

Figure 6: The parser accuracies at different query rounds using each single term in ABE(N-BEST).

A.4 Ablation Study of Factorization

As in Fig. 7, at several query rounds, the parser accuracy can be 3% - 6% higher than that using no factorization in *N-best Sequence Entropy*. But factorization does not help ABE(MAX) improve the parser performance at all.

A.5 Ablation Study of Error Term

As in Fig. 8, we combine Translation Error with different acquisition terms in ABE(N-BEST). Combing Translation Error and Translation Error achieves the best result in the low sampling regions. The accuracy is even 4% higher than the aggregated acquisition, ABE(N-BEST), when the

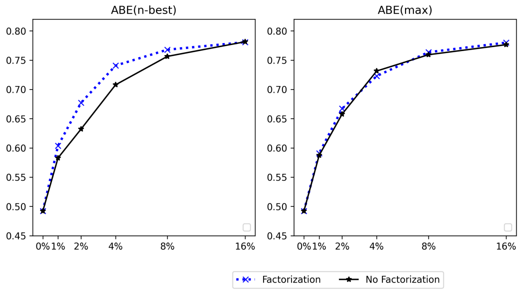

Figure 7: The parser accuracies using ABE(N-BEST) (Left) and ABE(MAX) (Right) with or without factorization.

sampling rate is 1%, suggesting that resolving translation error issues for semantically representative utterances benefits the parser more than resolving issues for tail utterances.

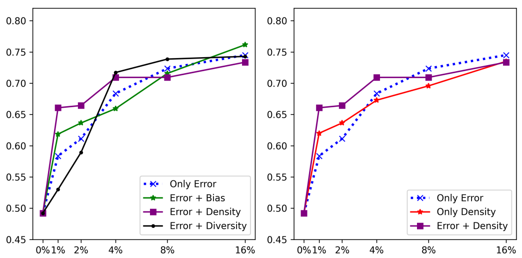

Figure 8: Combining Translation Error with different acquisition terms in ABE(N-BEST), we depict the parser accuracies using different acquisition combinations at each query round.

A.6 BT Discrepancy Pattern

We observe that the BT discrepancy patterns vary as in Fig. 9. For instance, the semantics of the BTs for Thai in GEOQUERY are altered dramatically due to the incorrect reorder of the words. Within NLMAP, the meanings of some German locations are inconsistent after the BT process.

A.7 T-SNE of Ablation Results

Fig. 10 shows the T-SNE plot of the representations of sampled utterances among all the utterances in the training set using ABE(N-BEST) and its various ablation settings. We encode the utterances with the pre-trained language model in the encoder of BERT-LSTM. We can see if we only use Semantic Density alone, the utterances are more likely to be in the dense region while not semantically diversified. On the contrary, the Translation Error tends to select tail utterances in the sparse semantic regions while they also lack semantic diversity. Both terms independently result in inferior parser performances. The Translation Bias and Translation Diversity collect utterances from diverse areas, thus

GeoQuery(De)	
Original	how mani state border at least one other state
MT	wie man an mindestens einen anderen Staat grenzt
BT	how to border at least one other state
GeoQuery(Th)	
Original	which capital are in the state that border s0
MT	ซึ่งเมืองหลวงอยู่ในสถานะที่ชายแดนs0
BT	in which the capital city is in the border state s0
GeoQuery(El)	
Original	how mani people live in c0
MT	πώς ζουν οι Μανιάτες στο γ0
BT	how do the Maniates live in c0
NLMap(De)	
Original	Are there any subway stations in Île-de-France that can be accessed with a wheelchair ?
MT	Gibt es in le-de-France U-Bahn-Stationen, die mit dem Rollstuhl erreichbar sind?
BT	Are there metro stations in le-de-France that are wheelchair accessible?

Figure 9: The original utterances and their corresponding machine translations and back-translations from GEOQUERY(DE), GEOQUERY(EL), GEOQUERY(TH) and NLMAP(DE).

providing better parser accuracies as in Fig. A.3. Removing Semantic Diversity from ABE(N-BEST) drops the parser performance most. As we observe, after removing Semantic Diversity, the utterances become more semantically similar compared to the utterances selected by ABE(N-BEST). Overall, the T-SNE plot can be supplementary proof to our claim that we should retain the representativeness and diversity of the utterances to guarantee the parser performance.

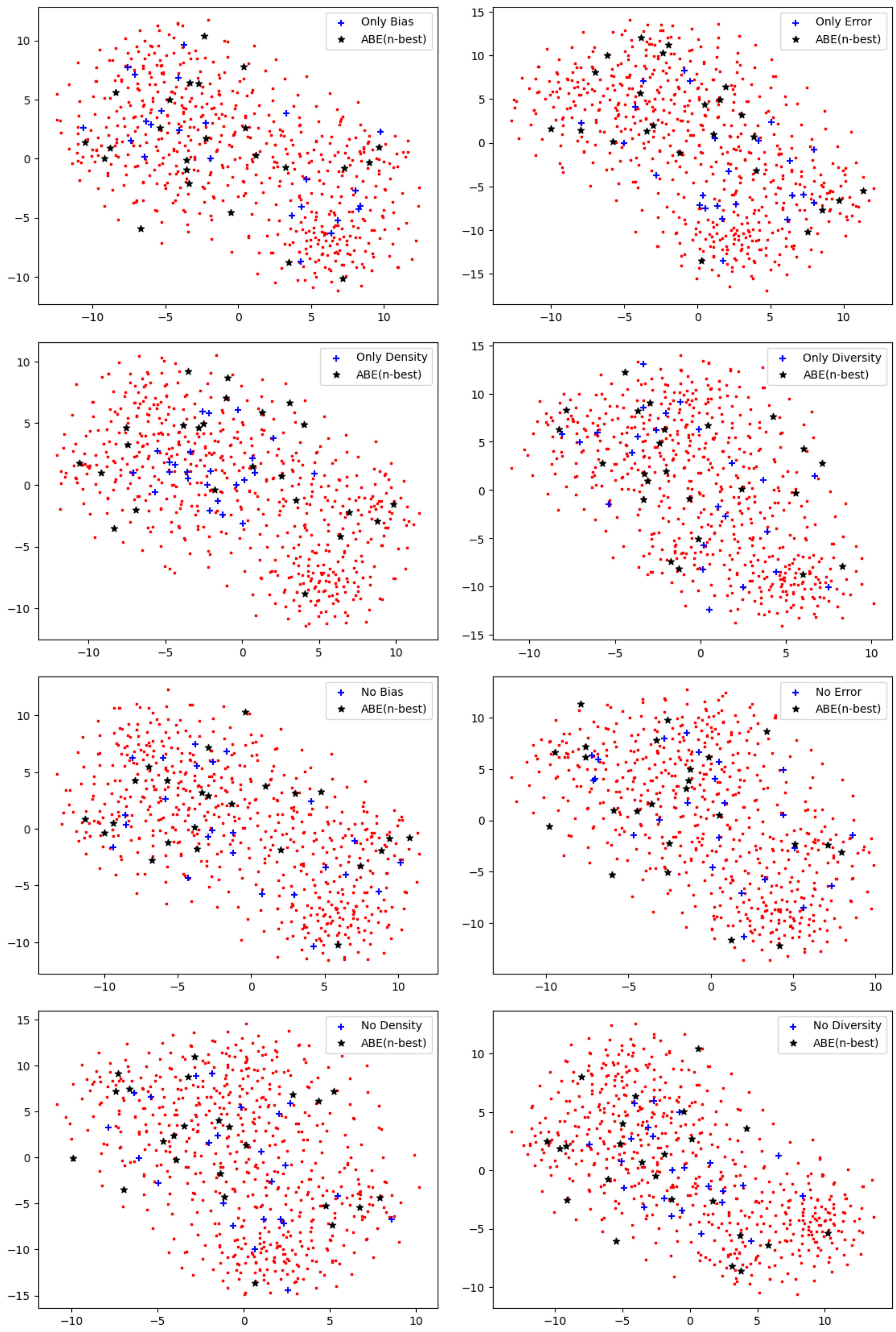

Figure 10: The representations of the English utterances selected by ABE(N-BEST) (black star) and the ablation settings (blue cross). Red dots are the representations of all the English utterances in the training set.

A.8 Evidence Lower Bound (ELBO)

The maximum likelihood estimation objective of our parser is:

$$\arg\max_{\theta}(\iiint_{\boldsymbol{x}_s \in \mathcal{X}_s, \boldsymbol{x}_t \in \mathcal{X}_t, \boldsymbol{y} \in \mathcal{Y}} P_\theta(\boldsymbol{y}, \boldsymbol{x}_t, \boldsymbol{x}_s)\, d\boldsymbol{x}_s\, d\boldsymbol{x}_t\, d\boldsymbol{y}) \tag{10}$$

where $\boldsymbol{x}_t$ is latent for most source utterance $\boldsymbol{x}_s$. We assume $P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)$ is a variational distribution.

$$\begin{aligned}
\log P_\theta(\boldsymbol{y}, \boldsymbol{x}_s) &= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log P_\theta(\boldsymbol{y}, \boldsymbol{x}_s)] \\
&= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log(\frac{P_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{x}_s)}{P_\theta(\boldsymbol{x}_t|\boldsymbol{y}, \boldsymbol{x}_s)})]
\end{aligned}$$

If we assume a conditional independence:

$$\begin{aligned}
&\equiv \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log(\frac{P_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{x}_s)}{P_\theta(\boldsymbol{x}_t|\boldsymbol{x}_s)})] \\
&= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log(\frac{P_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{x}_s)}{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)} \frac{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}{P_\theta(\boldsymbol{x}_t|\boldsymbol{x}_s)})] \\
&= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log(\frac{P_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{x}_s)}{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}] + \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log \frac{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)})] \\
&= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log(\frac{P_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{x}_s)}{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)})] + D_{\mathrm{KL}}(P_e||P_\theta)
\end{aligned} \tag{11}$$

where $\mathbb{E}$ denotes the expectation function over a specified distribution and D_{KL} denotes the Kullback–Leibler divergence between two distributions. Therefore the ELBO of $\log P_\theta(\boldsymbol{y}, \boldsymbol{x}_s)$ is $\mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log(\frac{P_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{x}_s)}{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)})]$.

$$\begin{aligned}
ELBO(P_\theta(\boldsymbol{y}, \boldsymbol{x}_s)) &= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log(\frac{P_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{x}_s)}{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)})] \\
&= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log P_\theta(\boldsymbol{x}_t, \boldsymbol{y}, \boldsymbol{x}_s) - \log P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)] \\
&= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log P_\theta(\boldsymbol{y}|\boldsymbol{x}_t)] - D_{\mathrm{KL}}(P_e||P_\theta) + \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log P_\theta(\boldsymbol{x}_s)] \\
&= \mathbb{E}_{P_e(\boldsymbol{x}_t|\boldsymbol{x}_s)}[\log P_\theta(\boldsymbol{y}|\boldsymbol{x}_t)] - D_{\mathrm{KL}}(P_e||P_\theta) + \log P_\theta(\boldsymbol{x}_s)
\end{aligned} \tag{12}$$

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the limitation section.

☒ A2. Did you discuss any potential risks of your work?
No risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In the abstract and conclusion sections.

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly and Quillbot. It helps me resolve writing issues and fix writing errors. Basically, I used them for all sections.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

In the Experiment section.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The space is limited. I am just using 4 V100 GPUs for all my experiments.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The space is limited.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In the Experiment section.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In the experiment section.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
In the experiment section.

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Space is limited.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We just use local students.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.