Active Learning for Multilingual Semantic Parser

Zhuang Li[*], **Gholamreza Haffari**
Openstream.AI
{zhuang.li,reza.haffari}@openstream.com

Abstract

Current multilingual semantic parsing (MSP) datasets are almost all collected by translating the utterances in the existing datasets from the resource-rich language to the target language. However, manual translation is costly. To reduce the translation effort, this paper proposes the first active learning procedure for MSP (AL-MSP). AL-MSP selects only a subset from the existing datasets to be translated. We also propose a novel selection method that prioritizes the examples diversifying the logical form structures with more lexical choices, and a novel hyperparameter tuning method that needs no extra annotation cost. Our experiments show that AL-MSP significantly reduces translation costs with ideal selection methods. Our selection method with proper hyperparameters yields better parsing performance than the other baselines on two multilingual datasets.

1 Introduction

Multilingual semantic parsing converts multilingual natural language utterances into logical forms (LFs) using a single model. However, there is a severe data imbalance among the MSP datasets. Currently, most semantic parsing datasets are in English, while only a limited number of non-English datasets exist. To tackle the data imbalance issue, almost all current efforts build MSP datasets by translating utterances in the existing datasets from the resource-rich language (e.g. English) into other languages (Duong et al., 2017; Li et al., 2021a). However, manual translation is slow and laborious. In such cases, active learning is an excellent solution to lower the translation cost.

Active learning (AL) is a family of methods that collects training data when the annotation budgets are limited (Lewis and Catlett, 1994). Our work proposes the *first* active learning approach for MSP. Compared to translating the full dataset, AL-MSP aims to select only a subset from the existing dataset to be translated, which significantly reduces the translation cost.

We further study which examples AL-MSP should select to optimize multilingual parsing performance. Oren et al. (2021) demonstrated that a training set with diverse LF structures significantly enhances compositional generalization of the parsers. Furthermore, our experiments show that the examples with LFs aligned with more diversified lexical variants in the training set considerably improve the performance of multilingual parsing during AL. Motivated by both, we propose a novel strategy for selecting the instances which include diversified LF structures with more lexical choices. Our selection method yields better parsing performance than the other baselines. By translating just 32% of all examples, the parser achieves comparable performance on multilingual GEOQUERY and NLMAP as translating full datasets.

Prior works obtain the hyperparameters of the AL methods by either copying configurations from comparable settings or tuning the hyperparameters on the seed evaluation data (Duong et al., 2018). However, the former method is not suitable as our AL setting is unique, whereas the second method requires extra annotation costs. In this work, we provide a cost-free method for our AL scenario for obtaining optimal hyperparameters.

Our contributions are i) the first active learning procedure for MSP that reduces the translation effort, ii) an approach that selects examples for getting superior parsing performance, and iii) a hyperparameter tuning method for the selection that does not incur any extra annotation costs.

2 Background

Multilingual Semantic Parsing. A multilingual semantic parser is a parametric model $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ that estimates the probability of the LF $\boldsymbol{y} \in \mathcal{Y}$ condi-

[*] Most of this author's work was completed during his internship at Openstream.AI.

Findings of the Association for Computational Linguistics: EACL 2023, pages 633–639
May 2-6, 2023

tioned on the natural language utterance $\boldsymbol{x} \in \mathcal{X}_l$ in an arbitrary language from a language set $l \in \mathcal{L}$. The model is trained on the utterance-LF pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N} \in \mathcal{X}_L \times \mathcal{Y}$ where $\mathcal{X}_L = \bigcup_{l \in L} \mathcal{X}_l$ includes multilingual utterances.

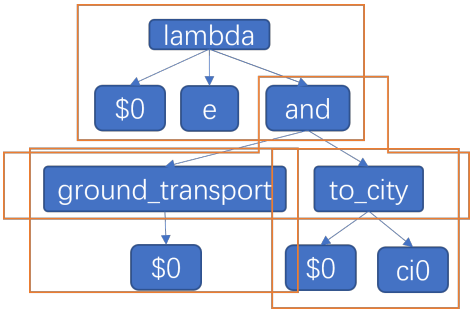

Figure 1: The example of the compounds in an LF tree, *(lambda $0 e (and (ground_transport $0) (to_city $0 ci0))).*

Atoms and Compounds. Each logical form sequence can be represented as a semantic tree, $\boldsymbol{y} = \tau_{\boldsymbol{y}}$. Oren et al. (2021); Shaw et al. (2021) define the nodes and sub-trees in $\tau_{\boldsymbol{y}}$ as the *atoms* and *compounds*, respectively. Increasing the diversity of the atoms and compounds in the training set improves the parser's compositional generalization (Oren et al., 2021; Li et al., 2021b). For example, an LF "*(lambda $0 e (and (ground_transport $0) (to_city $0 ci0)))*" can be expressed as a tree as in Fig. 1. The atoms are nodes such as "*lambda*", "*$0*", "*e*" in the LF tree. In this work, the compounds are defined as two-level sub-trees such as "*(ground_transport $0)*", "*(to _city $0 ci0)*", "*(and ground_transport to_city)*", and "*(lambda $0 e and)*" in the LF tree.

Data Collection for MSP. Prior data collection or active learning works annotates the LFs for the utterances (Duong et al., 2018; Sen and Yilmaz, 2020) or vice versa (Duong et al., 2018; Wang et al., 2015). But most MSP works (Susanto and Lu, 2017; Li et al., 2021a) obtain data by translating existing datasets from high-resource languages into low-resource languages, which is less costly since it does not need annotators' expertise in LFs. Following the same annotation paradigm, our AL does not annotate LFs for multilingual utterances, but instead chooses the utterances to be translated.

3 Active Learning for MSP

AL-MSP considers only a bilingual scenario for the proof of concept, while extending our AL method to more than two languages is easy. The goal of AL-MSP is to minimize the human effort in translating utterances while the semantic parser can still achieve a certain level of performance on the bilingual test sets. Starting from a semantic parser initially trained on the dataset $D_s = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i)\}_{i=1}^{N}$ whose utterances are in the high-resource language s, AL-MSP selects K_q examples $\tilde{D}_s = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i)\}_{i=1}^{K_q}$ from D_s, followed by manually translating the utterances in $\tilde{D}_s$ into a target language t, denoted by $\tilde{D}_t = t_{s \to t}(\tilde{D}_s)$, where $\tilde{D}_t = \{(\boldsymbol{x}_i^t, \boldsymbol{y}_i)\}_{i=1}^{K_q}$. The selection criterion is based on our proposed *acquisition* function $\phi(\boldsymbol{e}_s)$ scoring each example, $\boldsymbol{e}_s = (\boldsymbol{x}_s, \boldsymbol{y})$. The parser is re-trained on the union of $\tilde{D}_t$ and D_s. There will be Q iterations of selection and re-training until the re-trained parser reaches a good performance on the bilingual test sets T_s and T_t. Algorithm 1 describes our experimental settings in detail.

Algorithm 1: AL-MSP

Input : Initial training set $D^0 = D_s^0$, budget size K_q, number of the selection rounds Q
Output : A well-learned multilingual parser $P_\theta(\boldsymbol{y}|\boldsymbol{x})$
Train the parser $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ on the training set D^0
for $q \leftarrow 1$ *to* Q **do**
  Estimate the acquisition $\phi(\cdot)$
  Select a subset $\tilde{D}_s^q \in D_s^{q-1}$ of the size K_q based on the acquisition function $\phi(\cdot)$
  Translate the utterances in $\tilde{D}_s^q$ into the target language, $\tilde{D}_t^q = t_{s \to t}(\tilde{D}_s^q)$.
  Combine the training sets, $D^q = D^{q-1} \cup \tilde{D}_t^q$
  Exclude the selected examples $\tilde{D}_s^q$ from $D_s^q = D_s^{q-1} \setminus \tilde{D}_s^q$
  Re-train the parser $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ on D^q
  Evaluate parser performance on test sets T_s, T_t
end

3.1 Selection Acquisition

Our selection strategy selects the untranslated examples which maximize the acquisition scores. The acquisition comprises two individual terms, LF Structure Diversity and Lexical Choice Diversity.

LF Structure Diversity (LFSD). We give a simple technique to diversify the LF substructures (atoms and compounds) in the instances. At qth iteration, let $D_s^l = \bigcup_{i=1}^{q-1} \tilde{D}_s^i$ denotes all the translated examples and $D_s^u = D_s^{q-1}$ be the untranslated ones. We partition their union $D_s^u \cup D_s^l$ into $|D_s^l| + K_q$ clusters with Incremental K-means (Dataiku Lab, 2022). Each example $\boldsymbol{e}_s = (\boldsymbol{x}_s, \boldsymbol{y})$ is featurized by extracting all the atoms and compounds in the LF tree $\tau_{\boldsymbol{y}}$, followed by calculating the TF-IDF (Salton and McGill, 1986) value for each atom and com-

pound. Incremental K-means considers each example of D_s^l as a fixed clustering centroids and estimates K_q new cluster centroids. For each of the K_q new clusters, we select one example closest to the centroid.

Such selection strategy is reformulated as selecting K_q examples with the highest acquisition scores one by one at each iteration:

$$\phi_s(\boldsymbol{e}_s) = \begin{cases} -||f(\boldsymbol{y}) - \boldsymbol{c}_{m(\boldsymbol{y})}||^2 & \text{if } m(\boldsymbol{y}) \notin \bigcup_{\boldsymbol{e}_s \in D_s^l} m(\boldsymbol{y}) \\ -\infty & \text{Otherwise} \end{cases} \quad (1)$$

where $f(\cdot)$ is the feature function, $m(\cdot)$ maps each LF into its cluster id and $\boldsymbol{c}_i$ is the center embedding of the cluster i. As in Algo. 1, when a new example is chosen, none of its cluster mates will be selected again. The incremental mechanism guarantees the newly selected examples are structurally different from those chosen in previous iterations. Since we use batch-wise AL, we just estimate the clusters once per iteration to save the estimation cost.

Lexical Choice Diversity (LCD). LCD aims to select examples whose LFs are aligned with the most diversified lexicons. We achieve this goal by choosing the example maximizing the average entropy of the conditional probability $p(x_s|a)$:

$$\phi_c(\boldsymbol{e}_s) = -\frac{1}{|A_{\boldsymbol{y}}|} \sum_{a \in A_{\boldsymbol{y}}} \lambda_a \sum_{x_s \in V_s} p(x_s|a) \log p(x_s|a) \quad (2)$$

$$\lambda_a = \begin{cases} 1 & \text{if } a \in A_l \\ \beta & \text{Otherwise, } 0 \leq \beta < 1 \end{cases} \quad (3)$$

where a is the atom/compound, $A_{\boldsymbol{y}}$ is the set of all atoms/compounds extracted from $\boldsymbol{y}$, V_s is the vocabulary of the source language, A_l is the set of atoms/compounds in all selected examples until now, and $p(x_s|a)$ is constructed by counting the co-occurrence of a and x_s in the source-language training set. To prevent selecting structurally similar LFs, the score of each selected atom or compound is penalized by a decay weight β.

Our intuition has two premises. First, the parser trained on example pairs whose LFs have more lexical choices generalizes better. Second, LFs with more source-language lexical choices will have more target-language lexical choices as well.

LF Structure and Lexical Choice Diversity (LFS-LC-D). We eventually aggregate the two terms to get their joint benefits, $\phi(\boldsymbol{x}_s, \boldsymbol{y}) = \alpha\phi_s(\boldsymbol{x}_s, \boldsymbol{y}) + \phi_c(\boldsymbol{x}_s, \boldsymbol{y})$, where α is the weight that balances the importance of two terms. We normalize the two terms using quantile normalization (Bolstad et al., 2003) in order to conveniently tune α.

Hyperparameter Tuning. Because our setup is unique, we can not copy hyperparameters from existing works. The other efforts (Duong et al., 2018) get hyperparameters by evaluating algorithms on seed annotated data. To tune our AL hyperparameters, α and β, a straightforward practice using seed data is to sample multiple sets of examples from the source-language data, the target-language counterparts of which are in seed data, by varying different hyperparameter configurations and reveal their translations in the target language, respectively. The parser is trained on different bilingual datasets and evaluated on the *target-side* dev set. We use the one, which results in the best parsing performance, as the experimental configuration.

Such a method still requires translation costs on the seed data. We assume if the selected examples help the parser generalize well in parsing source-language utterances, their translations should benefit the parser in parsing target languages. Given this assumption, we propose a *novel* cost-free hyperparameter tuning approach. First, we acquire different sets of source-language samples by varying hyperparameters. Then, we train the parser on each subset and evaluate the parser on the *source-side* dev set. Finally, we use the hyperparameters with the best dev set performance.

4 Experiments

Datasets. We experiment with multilingual GEOQUERY and NLMAP. GEOQUERY utterances are in English (EN), German (DE), Thai (TH), and Greek (EL); NLMAP utterances are in English and German. Neither corpora include a development set, so we use 20% of the training sets of GEOQUERY and NLMAP in each language as the development sets for tuning the hyperparameters. To simulate AL process, we consider English as the resource-rich language and others as the target languages. After the examples are selected from the English datasets, we reveal their translations in the target languages and add them to the training sets.

AL Setting. We perform six iterations, accumulatively selecting 1%, 2%, 4%, 8%, 16% and 32% of examples from English GEOQUERY and NLMAP.

Baselines. We compare four selection baselines and the oracle setting: i) *Random* picks English utterances randomly to be translated, ii) *S2S*

(FW) (Duong et al., 2018) selects examples with the lowest parser confidence on their LFs, iii) *CSSE* (Hu and Neubig, 2021) selects the most representative and diversified utterances for machine translation, iv) *Max Compound* (Oren et al., 2021) selects examples that diversify the atoms and compounds in the LFs, v) *ORACLE* trains the parser on the full bilingual training set.

Evaluation. We adopt the exact match accuracy of LFs for all the experiments. We only report the parser accuracy on the target languages as we found the influence of new data is negligible to the parser accuracy on English data (See Appendix A.2).

Base Parser. We employ BERT-LSTM (Moradshahi et al., 2020) as our multilingual parser. Please see Appendix A.1 for its detailed description.

4.1 Hyperparameter Tuning

Table 1 displays the experiment results with the hyperparameters tuned using only English data (EN) and the hyperparameters tuned using seed data on i) English data plus a small subset (10% of train data plus development data) in the target language (EN + 10%), ii) the full bilingual data (EN + full), iii) the same dataset in a different pair of languages from our experiment languages (Diff Lang), iv) a different dataset in the same languages as our experiment (Diff Data).

	GEOQUERY			NLMAP
	DE	TH	EL	DE
EN (Ours)	73.86	74.57	77.57	69.43
EN + 10%	73.86	74.57	77.57	69.02
EN + full	73.86	74.57	77.14	69.43
Diff Lang	73.86	74.04	77.57	-
Diff Data	71.36	-	-	67.72

Table 1: The parsing accuracies on GEOQUERY and NLMAP test sets in various target languages after translating 16% of the English examples selected by LFS-LC-D with the optimal hyperparameters obtained by different tuning approaches.

From Table 1, we can see our approach takes significantly fewer annotation resources than others to find optimum hyperparameters. Adding more target-language data does not help obtain better hyperparameters, validating our assumption that English data is enough for LFS-LC-D to obtain good hyperparameters. Surprisingly, the hyperparameters tuned on a different language pair do not significantly worsen the selection choices. However, tuning hyperparameters from other datasets results in inferior parsing performance, which is anticipated as different datasets include different LFs, but the performance of LFS-LC-D is closely related to the LF structures.

4.2 Active Learning Results

Effectiveness of AL-MSP. Fig. 2 shows that only a small amount of target-language data significantly improves the parsing performance over the zero-shot performance. For example, merely 1% of training data improves the parsing accuracies by up to 13%, 12%, 15% and 6% on GEOQUERY(DE), GEOQUERY(TH), GEOQUERY(EL) and NLMAP(DE), respectively. With the best selection approach LFS-LC-D, translating 32% of instances yields parsing accuracies on multilingual GEOQUERY and NLMAP that are comparable to translating the whole dataset, with an accuracy gap of less than 5%, showing that our AL-MSP might greatly minimize the translation effort.

Effectiveness of LFS-LC-D. LFS-LC-D consistently outperforms alternative baselines on both multilingual datasets when the sampling rate is lower than 32%. In contrast, S2S(FW) consistently yields worse parser performance than the other baselines. Our inspection reveals that the parser is confident in instances with similar LFs. MAX COMPOUND diversifies LF structures as LFS-LC-D, however it does not perform well on GEOQUERY(TH). CSSE diversifies utterances yet performs poorly. We hypothesize that diversifying LF structures is more advantageous to the semantic parser than diversifying utterances. RANDOM also performs consistently across all settings but at a lesser level than LFS-LC-D.

Individual Terms of LFS-LC-D. We also inspect each individual term, LFSD and LCD, in LFS-LC-D. As in Fig. 3, both terms have overall lower performance than LFS-LC-D, indicating the combination of two terms is necessary. Specifically, LFSD performs poorly on NLMAP at the low sampling region. We inspect that NLMAP includes 5x more compounds than GEOQUERY. Therefore, it is difficult for the small number of chosen examples to encompass all types of compounds. LCD performs poorly on GEOQUERY(TH). We notice that Thai is an analytic language linguistically distinct from English, German or Greek, so the entropy values of the probability $p(x_s|a)$ over lexicons in Thai (p=0.03) is statistically more different to the ones over English than German (p=5.80e-30), and Greek (p=1.41e-30)[1]. Overall, the two terms could

[1]We use the Student's t-test (Demšar, 2006).

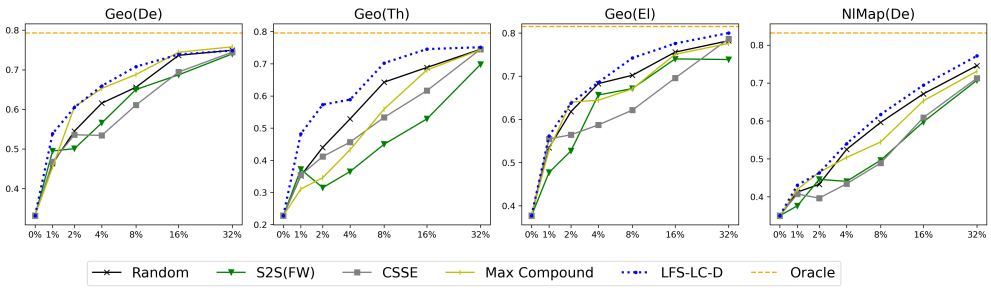

Figure 2: The parsing accuracies at different iterations on the test sets of GEOQUERY and NLMAP in German (De), Thai (Th), and Greek (El) using different selection approaches. All experiments are run five times with different seeds.

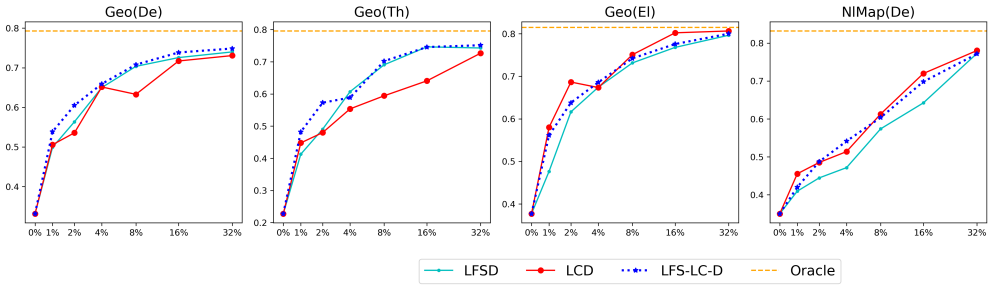

Figure 3: The parsing accuracies at different iterations on the test sets of GEOQUERY and NLMAP in German (De), Thai (Th) and Greek (El) using LFSD, LCD, and LFS-LC-D, respectively.

benefit each other, so LFS-LC-D performs steadily across different settings.

Comparison with Machine Translation. We also evaluate the parsers that utilize machine translation services. The parsers are trained on a combination of English data and data translated into the target language by Google Translation (Wu et al., 2016). The accuracy of parsers evaluated on test sets of Geo(De), Geo(Th), Geo(El), and NlMap(De) was 49%, 58%, 75%, and 75%, respectively. These parsing accuracies are significantly lower than those attained by parsers trained on data provided through human translation, which achieved 80%, 80%, 81%, and 83%, respectively. This suggests that the performance of the parser is tightly correlated to the quality of the employed machine translation system. Clearly, human translation delivers a greater output quality compared to machine translation. In addition, the results reveal that parsers employing AL methods can easily outperform those employing machine translation methods, particularly when the sampling rate for AL is more than 1%, 4%, 8%, and 32% in the four data settings.

5 Conclusion

We conducted the first in-depth empirical study to investigate active learning for multilingual semantic parsing. In addition, we proposed a method to select examples that maximize MSP performance and a cost-free hyperparameter tuning method. Our experiments showed that our method with the proper hyperparameters selects better examples than the other baselines. Our AL procedure with the ideal example selection significantly reduced the translation effort for the data collection of MSP.

Limitations

To reduce annotation costs, existing data collection methods for MSP also utilize machine translation (Moradshahi et al., 2020). Despite the generally lower quality of machine-generated translations compared to human translations, the cost of machine translation services is notably more economical. Our study pioneers the investigation into the feasibility of reducing annotation costs by manually translating only selective portions of the utterance pool. In our work, we provide an initial evaluation of parsers using machine translation versus those using AL methods. Further research is necessary to thoroughly compare these cost-reduction approaches, highlighting their respective advantages and limitations, which we intend to pursue as part of our future work.

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.

Dataiku Lab. 2022. Cardinal.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30.

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.

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.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

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, 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Inbar Oren, Jonathan Herzig, and Jonathan Berant. 2021. Finding needles in a haystack: Sampling structurally-diverse training sets from synthetic data for compositional generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10793–10809.

Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval.

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.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

A Appendix

A.1 Implementation Details

BERT-LSTM BERT-LSTM is a Sequence-to-Sequence model (Sutskever et al., 2014) with the XLM-RoBERTa-base (Liu et al., 2019) as its encoder and an LSTM (Hochreiter and Schmidhuber, 1997) as its decoder.

Hyperparameters of the Parsers We tune the hyperparameters of BERT-LSTM on English data. For a fair comparison, we fix the hyperparameters of the parser while evaluating the active learning methods. Specifically, we set the learning rate to 0.001, batch size to 128, LSTM decoder layers to 2, embedding size for the LF token to 256, and epochs to 240 and 120 for the training on GEOQUERY and NLMAP, respectively.

Hyperparameters of AL For tuning the hyperparameters of the active learning method, we grid search the decay weight β in 0, 0.25, 0.5, 0.75 and the weight balance rate α in 0.25, 0.5, 0.75, 1. The optimal hyperparameters are 0.75 and 0.75 for all language pairs of GEOQUERY and 0.75 and 0.25 for multilingual NLMAP.

In the Diff Lang setting, we assume we can access the data in a language pair other than the experimental one. For selecting English utterances to be translated into German, Thai, and Greek, we tune the hyperparameters on the data of En-Th, En-EL, and En-De pairs, respectively.

In the Diff Data setting, we assume we can access the data in the same language pair as our experimental one but in a different domain with a different type of LF. For selecting English utterances in GEOQUERY for translation, we tune the hyperparameters on the bilingual NLMAP. For selecting utterances in NLMAP, we tune the hyperparameters on the GEOQUERY in the language pair, En-De.

A.2 Parser Accuracies on English Test Sets

As in Fig. 4, training the parser on the data in the target language does not significantly influence the parser's performance on the English test sets. Therefore, in Sec. 4, we only report the experimental results on the test sets in the target languages.

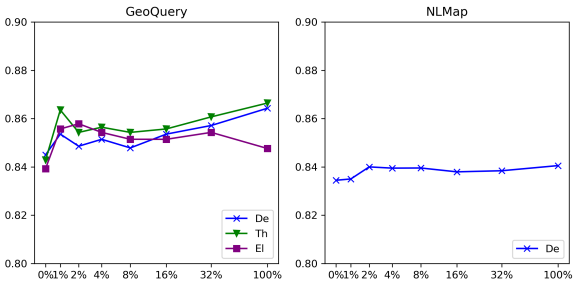

Figure 4: The parsing accuracies at different iterations on the English test sets of GEOQUERY and NLMAP after selecting data in German (De), Thai (Th) and Greek (El) using LFS-LC-D, respectively.