ZeLa: Advancing Zero-Shot Multilingual Semantic Parsing with Large Language Models and Chain-of-Thought Strategies

Dinh-Truong Do, Minh-Phuong Nguyen, Le-Minh Nguyen
Japan Advanced Institute of Science and Technology, Japan
{truongdo,phuongnm,nguyenml}@jaist.ac.jp

Abstract

In recent years, there have been significant advancements in semantic parsing tasks, thanks to the introduction of pre-trained language models. However, a substantial gap persists between English and other languages due to the scarcity of annotated data. One promising strategy to bridge this gap involves augmenting multilingual datasets using labeled English data and subsequently leveraging this augmented dataset for training semantic parsers (known as zero-shot multilingual semantic parsing). In our study, we propose a novel framework to effectively perform zero-shot multilingual semantic parsing under the support of large language models (LLMs). Given data annotated pairs (`sentence, semantic representation`) in English, our proposed framework automatically augments data in other languages via multilingual chain-of-thought (CoT) prompting techniques that progressively construct the semantic form in these languages. By breaking down the entire semantic representation into sub-semantic fragments, our CoT prompting technique simplifies the intricate semantic structure at each step, thereby facilitating the LLMs in generating accurate outputs more efficiently. Notably, this entire augmentation process is achieved without the need for any demonstration samples in the target languages (zero-shot learning). In our experiments, we demonstrate the effectiveness of our method by evaluating it on two well-known multilingual semantic parsing datasets: MTOP and MASSIVE.

Keywords: Zero-shot, Multilingual Semantic Parsing, Large Language Model

1. Introduction

Enabling cross-lingual technologies is crucial because of their high applicability in global communication, breaking down language barriers, and fostering collaboration among diverse communities worldwide. Recent advancements in natural language processing (NLP), especially the development of advanced multilingual language models have attracted significant research interest (Conneau et al., 2020; Xue et al., 2021; Muennighoff et al., 2023). Semantic parsing is a fundamental task within the field of NLP, with broad applications in both business and everyday life (Do et al., 2023; Mansimov and Zhang, 2022). For instance, it plays a significant role in the development of Virtual Assistants (Fischer et al., 2021). The primary goal of semantic parsing is to convert user input into a semantic form that computers can process as structured data (e.g., X_{eng}, Y_{eng} in Figure 1). However, creating annotated data for multilingual semantic parsing is time-consuming and requires a high degree of human expertise.

To address the problem of the lack of annotated data in the multilingual semantic parsing task, previous research in this area can be broadly categorized into two main approaches. The first approach involves training an aligner using the English reference dataset, which is then used to predict semantic forms in other languages (Nicosia et al., 2021; Gritta et al., 2022). The second approach skips the aligner training step and relies

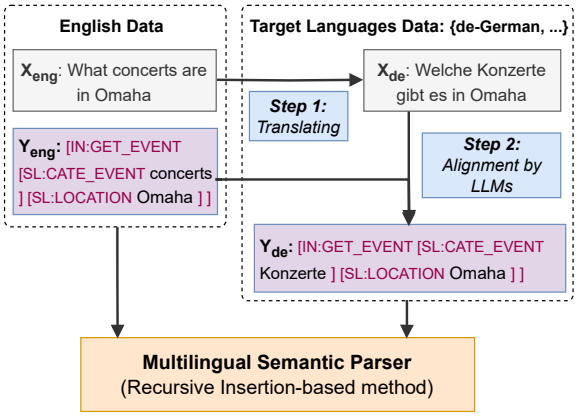

Figure 1: Our framework utilizes LLMs to perform zero-shot multilingual semantic parsing.

on Large Language Models (LLMs) incorporating human support (in the selection of a few demonstrations) to align the reference form with the desired target semantic form (Awasthi et al., 2023; Winata et al., 2021), utilizing in-context learning techniques (Garg et al., 2022).

In this study, we introduce ZeLa, a framework designed to enhance the performance of zero-shot multilingual semantic parsing (Figure 1). We achieve this by harnessing the inherent cross-lingual generalization capabilities of LLMs for data augmenting on new languages and incorporating the current state-of-the-art (SOTA) semantic parsing method, grammar-based RINE (Do et al., 2023). Our approach consists of two main phases:

LREC-COLING 2024, pages 17783–17794

LLM-based augmentation and multilingual semantic parsing. In the first phase, **LLM-based augmentation**, we initiate by utilizing off-the-shelf translating tools to convert English utterances into the desired target languages. Subsequently, we employ a range of CoT-prompting techniques to guide LLMs in the progressive parsing of complex logical structures in new languages to achieve a multilingual corpus for the training process. For example, given a pair in English (a sentence and its semantic form), the sentence is translated into a new language like German as "`Welche Konzerte gibt es in Omaha`". With the LLM model, the translated sentence is then used to generate a "silver" German semantic representation, "`[IN:GET_EVENT [SL:CATE_EVENT Konzerte ] [SL:LOCATION Omaha ] ]`". Compared to prior approaches that rely on a few examples of the new languages for demonstrations (Awasthi et al., 2023; Winata et al., 2021), our method utilizes only the English utterance-English semantic form pair to guide the LLMs that incrementally predict the semantic form in the target language (zero-shot setting). In the second phase, **multilingual semantic parsing**, we employ two approaches. The first approach follows the standard seq2seq method, consistent with previous works that treat the output semantic form as a sequence of text (Nicosia et al., 2021; Awasthi et al., 2023). The second approach, the grammar-based RINE model (Do et al., 2023; Mansimov and Zhang, 2022), views the semantic form as a sequence of recursive steps. Our experimental results, conducted on two well-established multilingual semantic parsing datasets, MTOP (Li et al., 2021) and MASSIVE (FitzGerald et al., 2023), demonstrate the effectiveness of our framework. In the realm of zero-shot multilingual semantic parsing, the ZeLa framework achieves an Exact Match (EM) score on the MTOP dataset that surpasses a state-of-the-art (SOTA) method by 1.43 points. Additionally, it shows similarly promising results on the MASSIVE dataset, showcasing its potential for advancing multilingual semantic parsing tasks. Our paper makes the following contributions:

- Introducing effective CoT techniques for augmenting multilingual semantic parsing datasets from the given English annotated pair data.
- Applying the SOTA hierarchical semantic parsing model, grammar-based RINE, for addressing the multilingual semantic parsing.
- Conducting comprehensive experiments that demonstrate the effectiveness of the ZeLa framework on two datasets, namely, MTOP and MASSIVE.

2. Related Works

Zeroshot Multilingual Semantic Parsing. In the field of multilingual semantic parsing, MTOP (Li et al., 2021) and MASSIVE (FitzGerald et al., 2023) have emerged as benchmarks to evaluate models. These benchmarks include a challenging multilingual zero-shot setting, where models are exclusively given English data and must generate predictions in target languages. This setting is significant because developing semantic parsing datasets for languages other than English can be quite challenging. Translation-based approach is a common approach in zero-shot multilingual semantic parsing (Hartrumpf et al., 2008; Liang et al., 2020; Fang et al., 2021). Translation can be performed at prediction time, wherein the user query is translated into English and then processed by an English semantic parser to obtain the logical representation (Artetxe et al., 2020; Uhrig et al., 2021). Alternatively, this approach can be used during model training by translating English utterances into other languages and employing these translated versions as augmented training data (Sherborne et al., 2020; Moradshahi et al., 2020; Nicosia et al., 2021; Awasthi et al., 2023). Previous research has shown that incorporating translation during training yields better results compared to translating at inference time (Yang et al., 2022). This strategy is also adopted in building our semantic parsing model in this study.

LLMs on Semantic Parsing. An LLM refers to a language model characterized by its substantial number of parameters, which can vary from several billion parameters, *e.g.* T5 (Raffel et al., 2020) with 11 billion parameters, to several hundred billion parameters, *e.g.* PALM (Chowdhery et al., 2022) with 540 billion parameters. These models are primarily built using the Transformer architecture (Vaswani et al., 2017) and undergo initial training on massive volumes of unlabeled text. After this pre-training phase, they can be fine-tuned to adapt their learned representations for human instructions or specific tasks in the NLP field, including semantic parsing (Touvron et al., 2023). Various approaches have been proposed to tackle the challenges of semantic parsing by harnessing the capabilities of in-context learning (ICL) in LLMs. Shin and Van Durme (2022) demonstrated that when performing ICL, LLMs pre-trained on programming code, like Codex (Chen et al., 2021), outperform LLMs like GPT-3 (Brown et al., 2020a), which are primarily trained in natural language linguistics, particularly in tasks related to semantic parsing. An et al. (2023) pointed out that the success of ICL in semantic parsing depends on the choice of demonstration set. They identified three

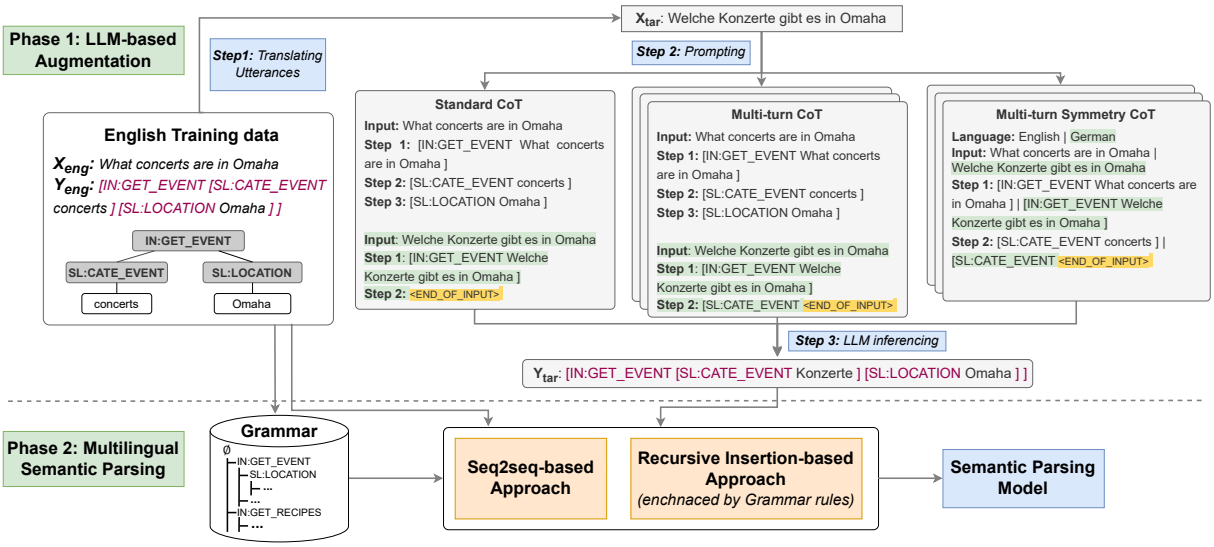

Figure 2: System architecture of our ZeLa framework.

critical factors when selecting demonstration samples: diversity, similarity, and complexity. Similarity involves merging basic structures within explicit expressions, diversity evaluates the recurrence of patterns across different contexts, and complexity measures the richness of information contained in each example. Additionally, another method introduced by Levy et al. (2023) focuses on creating comprehensive demonstrations within semantic parsing. Their approach aims to collect demonstrations that cover sub-logical structures necessary for predicting new inputs.

Conceptually, the most similar to our work, Awasthi et al. (2023) demonstrated the effectiveness of using LLMs to translate English datasets into multiple languages through ICL. To achieve this, they manually selected a set of demonstrations in the target languages and used these demonstrations to construct input data for LLMs. The outputs from LLMs were then utilized as augmented data to train the multilingual semantic parser. However, in contrast, our approach is entirely focused on a zero-shot multilingual setting, eliminating the need for any demonstration samples in the target language.

3. Methodology

Our study is centered on zero-shot multilingual semantic parsing. We begin with an English dataset containing pairs of data, each represented as `(utterance, semantic-form)`, denoted as $\mathcal{D}_{eng} = \{(x^i_{eng}, y^i_{eng})\}$. Our proposed framework, ZeLa, consists of two main phases: *LLM-based Augmentation* and *Multilingual Semantic Parsing* (Figure 2). First phase, we employ an augmentation technique to transform $\mathcal{D}_{eng}$ into augmented data, denoted as $\mathcal{D}_{tar} = \{(x^i_{tar}, y^i_{tar})\}_{tar \in \{de, fr, \cdots\}}$. In this context, x^i_{tar} represents the translated utterance derived from x^i_{eng}, and y^i_{tar} corresponds to the logical form in the target language. It is important to note that these logical forms in both English (y^i_{eng}) and the target language (y^i_{tar}) share the same semantic schema. For instance, as seen in Figure 2, where both English and German logical forms share the schema `[IN:GET_EVENT [SL:CATE_EVENT ] [SL:LOCATION ] ]`. The only difference between them lies in the specific span within each slot. Finally, we obtained a new augmented dataset by combining data pairs on all languages $D_{eng} \cup \mathcal{D}_{tar|tar \in \{de, fr, \cdots\}}$. In the second phase, we train a multilingual semantic parser with augmented data from the first phase. In the following sections, we will provide a detailed explanation of our methodology for achieving this objective.

3.1. LLM-based Data Augmentation

We found that despite the effectiveness of LLMs in understanding natural sentences, they struggle to generalize complex semantic structures. One reason for this challenge lies in the fact that LLMs are primarily trained on large-scale natural language data, making them proficient at tasks involving natural sentences (e.g., question answering or machine translation) rather than semantic parsing tasks. To address this challenge, we break down the entire semantic representation into sub-semantic fragments to simplify the complex semantic structure at each step (Figure 3). We define a semantic fragment as a non-terminal node in a hierarchical semantic representation, encompassing semantic tokens and the corresponding text span of that node (*e.g.,* `[SL:LOCATION` *Om-*

aha]). This approach makes each semantic fragment more natural than the original comprehensive semantic structure. By providing examples of list semantic fragments in the English language, LLMs based on that to generalize similar semantic fragments in a new language. In addition, we

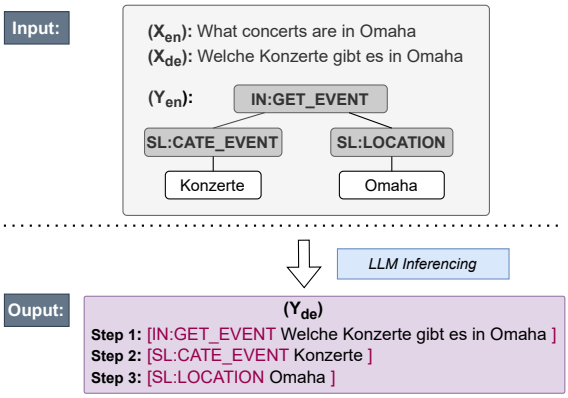

Figure 3: Overview of data augmentation process (en → de) supported by LLMs.

propose three strategies using the CoT prompting technique to instruct LLMs to generate a semantic form in the target language step-by-step (y^i_{tar}) from the given triple (x^i_{eng}, y^i_{eng}, x^i_{tar}): Standard CoT, Multi-turns CoT, Multi-turn symmetry CoT.

Standard CoT In this strategy, the LLMs generate all the semantic fragments of the target language in one turn (Figure 4). Notably, the semantic schema between different languages is exactly the same, therefore, we use the root node ([IN:GET_EVENT *Welche Konzerte ...*]) to initialize the CoT prompting template as a trigger encouraging the LLMs to continue generating the next semantic fragments. Finally, we utilize these generated semantic fragments to reconstruct the semantic representation in the target language.

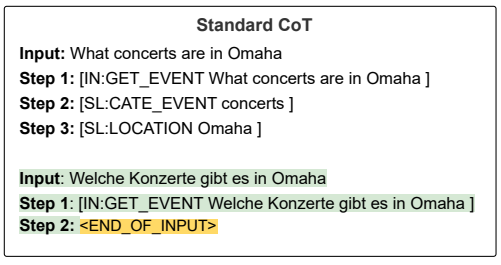

Figure 4: Standard CoT prompting example.

Multi-turn CoT As we mentioned, the semantic schema is shared between languages; however, in the Standard CoT strategy, the semantic form generated by LLMs is not guaranteed this characteristic. Therefore, in this strategy, we only use LLMs to incrementally generate the span related to each semantic frame decoded in English. For example, in Figure 5, in the first turn, the LLMs are required to predict *"Konzerte"* is text span of SL:CATE_EVENT, and continuously predict span *"Omaha"* belong to SL:LOCATION in the second turn (last turn).

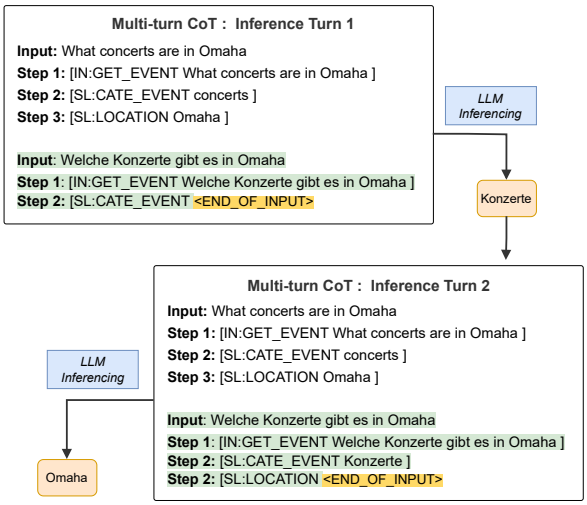

Figure 5: Multi-turn CoT prompting.

Multi-turn symmetry CoT Compared with the multi-turn CoT prompting strategy, this strategy augmented the characteristic of the alignment of semantic fragments between English and the target language. To this end, we pair each semantic fragment of both languages together (Figure 6). This alignment aids in transferring information and enhances the coordination and coherence between steps in different languages, making the parsing process more effective.

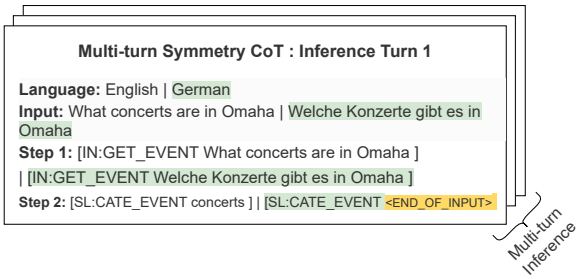

Figure 6: Symmetry multilingual CoT prompting.

Filtering outputs In order to address the potential presence of noisy generated text, we have implemented a systematic filtering procedure composed of several sequential steps. The objective of this process is to curate our augmented data to

maintain a high standard of quality. The steps involved are as follows:

- **Span Alignment**: First, we discard samples whose spans do not match the translated utterance. This ensures the generated text remains coherent and relevant to the intended content.
- **Label and Intent Validation**: Next, we check for samples that contain label slots or intents that are not recognized. We use a reference label set extracted from English data to validate this. Samples with unrecognized labels or intents are removed from consideration.
- **Semantic Parsing Tree Compatibility**: Additionally, we assess whether the samples can be converted into accurate semantic parsing trees. Any samples that cannot undergo this conversion successfully are also excluded.

After this rigorous filtering process, we are left with a refined and augmented dataset in the target language. This dataset is now well-prepared and suitable for training the semantic parsers, ensuring that they learn from high-quality data.

3.2. Multilingual Semantic Parsing

3.2.1. Seq2seq-based approach

This approach employs a standard seq2seq method to perform multilingual semantic parsing, following previous research (Awasthi et al., 2023; Nicosia et al., 2021). It uses an encoder-decoder multilingual pre-trained model as its backbone (Xue et al., 2021) with augmented training data. The loss function is defined as the negative log-likelihood of the true tokens in the output sequence, and it is defined as follows:

$$Loss = -\sum_{t=1}^{T} \log(p\ (\ y_t \mid y_{<t}, x)) \qquad (1)$$

Here, T denotes the length of the output sequence, y_t represents the true token at time step t, $y_{<t}$ encompasses the sequence of tokens leading up to time step $t-1$, x corresponds to the input sequence within the augmented training data, and $p\ (\ y_t \mid y_{<t}, x)$ denotes the model's predicted probability of token y_t at time step t, taking into consideration the sequence of tokens up to time step $t-1$ and the input sequence x.

3.2.2. Recursive Insertion-based approach

In this approach, we utilize the recursive-insertion based approach (Mansimov and Zhang, 2022), enhanced by grammar constraints. It employs an encoder-only multilingual pre-trained language model as a backbone (Conneau et al., 2020). The parsing process is represented as an incremental generation of sub-parsed trees, with the output of the previous step serving as the input for the current step. Grammar constraints are extracted from English data and are used to guide the parsing process (*e.g.* Table 1). The loss function combines node label prediction, start position prediction, and end position prediction with a grammar-based penalty to filter out unpromising node label predictions.

Step	Linearized representation of full logical form
$\mathcal{P}_0$	What concerts are in Omaha
$\mathcal{P}_1$	[IN:GET_EVENT What concerts are in Omaha]
$\mathcal{P}_2$	[IN:GET_EVENT What [SL:CATE_EVENT concerts] are in Omaha]
$\mathcal{P}_3$	[IN:GET_EVENT What [SL:CATE_EVENT concerts] are in [SL:LOCATION Omaha]]

Table 1: Example chain of incremental trees in parsing process using of recursive insertion-based approach.

Following previous work (Do et al., 2023), we extract grammar from the English semantic parsing tree (tree representation of logical form) to cause the unpromising label prediction to be ignored. Specifically, we extract parent-child grammar rules from the semantic parsed tree of English data $\mathcal{G} = \{A \rightarrow B \mid A, B$ are non-terminal nodes $\}$, for example `IN:GET_EVENT` $\rightarrow$ `SL:LOCATION`. It is important to note that the sample in both English and the target language shares an identical schema, differing only in the arrangement of slots within each intent (Awasthi et al., 2023). Consequently, the grammar extracted from English data possesses a general applicability that extends to target languages. After obtaining grammar, we are ready to train a multilingual semantic parsing model for all languages by using the grammar-based RINE model (Do et al., 2023).

4. Experiment

In this section, we describe the details of the experiment setups and the main results [1].

4.1. Datasets and Evaluation Metric

To assess the effectiveness of our methods, we conducted experiments using two well-known multilingual semantic parsing datasets: MTOP (Li et al., 2021) and MASSIVE (FitzGerald et al., 2023). Each dataset comprises English utterances and their corresponding translations into other languages.

[1]The source code of this work is released at https://github.com/truongdo619/ZeLa

MTOP (Li et al., 2021) This dataset is a parallel multilingual semantic parsing dataset that encompasses utterances and their decoupled compositional representations in six languages: English, German, French, Hindi, Thai, and Spanish. The term "parallel" signifies that the dataset was generated beginning with English utterances and their annotations. Subsequently, these utterances and their corresponding semantic representations were translated into the other five languages through a meticulous process involving post-processing, post-editing, and the filtering of uncertain utterances. MTOP covers 11 different domains, encompassing 117 intent types and 78 slot types. On average, each language in the dataset contains approximately 12.3K training utterances, 1.5K development utterances, and 2.7K test utterances.

MASSIVE (FitzGerald et al., 2023) This dataset is a diverse multilingual semantic parsing dataset that spans 51 different languages. Unlike MTOP, it solely contains tree representations without nested utterances. This dataset encompasses 18 domains, 60 intent types, and 50 slot types. On average, each language in the dataset includes about 11.5K training utterances, 2K development utterances, and 3K test utterances. In our experiments, we focused on six languages: English, German, Spanish, French, Hindi, and Thai.

Evaluation Metric In line with prior research (Awasthi et al., 2023), we employ the agnostic exact match (EM) as our primary evaluation metric. This metric compares two logical forms while disregarding the order of slots within an intent, thus ensuring a correct evaluation. We assess our methods, as well as other methods, based on this exact match score.

4.2. Experimental Setting

LLM-based Augmentation In this phase, we utilized three versions of the Llama 2 models (Touvron et al., 2023) with 7B, 13B, and 70B parameters. To ensure stable results when working with LLMs, we set the temperature value to 0 as in previous works (Levy et al., 2023; Zhuo et al., 2023).

Seq2seq-based Multilingual Semantic Parsing The pre-trained mT5-Large checkpoint (1.2B parameters) (Xue et al., 2021) was utilized for initialization of our seq2seq semantic parser. Fine-tuning was performed using a combination of English gold data and augmented data from other languages. We employed the Adam optimizer with a learning rate of 1e-5, a warm-up period of 1000 steps, and a batch size of 32. Training extended over 10,000 steps and took an average of 25 hours on a single A100 80GB GPU. We selected the best checkpoint based on performance on the development set and used it for predictions on the test set.

Recursive-based Multilingual Semantic Parsing We utilized the XLM-Roberta model (355M parameters) as an encoder. The training process involved the use of the Adam optimizer with a number of warm-up steps set to 1000. The learning rate was chosen from the options {**1e-05**, 5e-05, 1e-06}[2], and the training spanned 50 epochs (1000 steps). Grammar rules were extracted from the training set of the English gold data.

Baseline We reproduce the TAF method (Nicosia et al., 2021; Awasthi et al., 2023) as a strong baseline in our study. For hyperparameters, we maintain the same values as specified in the original papers.

4.3. Main Results

We compared our proposed methods with methods from previous research, including zero-shot setting, few-shot setting, and the TAF method (Awasthi et al., 2023; Nicosia et al., 2021), on two datasets: MTOP and MASSIVE. Specifically, for the MTOP dataset, we considered several methods: (1) Seq2seq Zero-shot: Trained exclusively on English data and utilized semantic parsing model based on the seq2seq approach (Awasthi et al., 2023). (2) Seq2seq Few-shot: Trained with additional human-selected samples combined with English data (Awasthi et al., 2023). (3) Seq2seq TAF: Employed English data mixed with corresponding augmented data using the TAF method (Awasthi et al., 2023). (4) Seq2seq Zero-shot: Our reproduction of the zero-shot setting using the seq2seq approach. (5) Seq2seq TAF: Our reproduction of the TAF method. (6) RINE-based Zero-shot: Our zero-shot setting uses English data only with the recursive insertion-based method (RINE) enhanced by grammar constraints. (7) Seq2seq ZeLa: Our proposed method utilizes the seq2seq approach. (8) RINE-based ZeLa: Our proposed method utilizes the recursive insertion-based method enhanced by grammar constraints.

MTOP In the MTOP part of Table 2, our best method outperformed previous works by achieving a 1.4 EM score improvement compared to the state-of-the-art TAF method (Awasthi et al., 2023; Nicosia et al., 2021), as shown by the results in rows (3) and (8). This improvement was particularly significant for low-resource languages like Hindi and Thai, as seen in rows (7)

[2]The values in the best performance are **bold**.

Method	MTOP						MASSIVE					
	de	es	fr	hi	th	Avg	de	es	fr	hi	th	Avg
(1) Seq2seq Zero-shot (Awasthi et al., 2023)	54.4	57.8	62.8	42.3	42.1	51.9	54.3	53.4	54.6	40.1	49.4	50.4
(2) Seq2seq Few-shot (Awasthi et al., 2023)	62.8	69.5	65.9	55.3	53.9	61.5	54.3	58.1	58.0	54.4	60.0	57.0
(3) Seq2seq TAF (Awasthi et al., 2023)	75.0	74.9	78.0	63.0	60.8	70.3	67.5	64.6	65.3	61.6	63.5	64.5
(Our methods)												
(4) Seq2seq Zero-shot	54.0	58.9	58.9	44.1	38.3	50.8	54.4	51.8	54.5	44.9	51.3	51.4
(5) Seq2seq TAF	73.2	**75.2**	78.5	61.9	62.6	70.3	63.3	62.5	62.1	58.5	**67.1**	62.8
(6) RINE-based Zero-shot	63.5	68.1	70.3	54.4	43.9	60.0	62.2	56.4	58.0	55.0	57.6	57.8
(7) Seq2seq ZeLa	73.9	71.9	76.2	**71.0**	**62.4**	71.1	68.0	**65.0**	**66.3**	**62.7**	62.0	64.8
(8) RINE-based ZeLa	**75.4**	74.0	**78.7**	70.3	60.3	**71.7**	**68.2**	64.0	65.6	62.3	65.6	**65.1**

Table 2: Performance comparison using Exact Match on MTOP and MASSIVE test sets.

and (8). Among the methods using seq2seq semantic parsers, namely rows (1), (2), (3), and (7), the ZeLa seq2seq-based approach achieved the highest results. This demonstrated the effectiveness of our augmentation method using LLMs enhanced by our multilingual CoT prompting strategies. Furthermore, when using the recursive insertion-based method enhanced by grammar rules in rows (6) and (8), performance further improved compared to the seq2seq-based method. This highlighted the value of incorporating grammar information and breaking down the parsing process into substeps.

MASSIVE In the case of the MASSIVE dataset, as shown in Table 2, we evaluated the performance of our method alongside previous approaches. Our method achieved superior performance, surpassing the TAF method (Awasthi et al., 2023; Nicosia et al., 2021) by a margin of 0.6 EM. This result underscores the generalizability of our approach beyond the MTOP dataset, demonstrating its effectiveness across different datasets. Furthermore, the performance trends observed in the MASSIVE dataset closely mirror those seen in the MTOP dataset. Specifically, our augmentation method exhibited superior performance when paired with the seq2seq semantic parser, outperforming the TAF method, as shown in rows (3) and (5). Additionally, the adoption of a recursive-insertion-based approach in row (6), enhanced by grammar rules, led to further performance improvements compared to the seq2seq method, mirroring the trends observed in the MTOP dataset.

In addition, we performed a bootstrap t-test (Efron and Tibshirani, 1994) with the null hypothesis that our proposed ZeLa framework (row 8) and the baseline TAF method (row 6) have the same expected values. This analysis involved three random seeds for each experimental setting. The obtained p-values in both MTOP and MASSIVE datasets were less than 0.05, providing strong evidence that our proposed framework significantly surpasses the baseline in performance.

Method	de	es	fr	hi	th	Avg
Standard CoT	71.3	70.2	75.5	59.0	56.6	66.5
Multi-turn CoT	72.1	70.4	75.6	61.4	56.7	67.3
Multi-turn symmetry CoT	**72.3**	**71.3**	**77.7**	**66.7**	**59.8**	**69.6**

Table 3: **Impact of CoT prompting strategies**: Multi-turn symmetry CoT results more effective augmented datasets yielding higher EM accuracy.

5. Analysis

To gain deeper insights into our framework, we conducted several analyses. All the experiments in this section are performed on the MTOP dev set.

5.1. Role of CoT Prompting Strategies

In Table 3, we present the results of three proposed multilingual CoT strategies: standard, multi-turn, and multi-turn symmetry. The multi-turn symmetry CoT strategy achieved the most effective results. We attribute this superior performance to the symmetry strategy's capacity to improve the alignment and coherence between steps at the same level in both English and the target languages.

5.2. Role of LLM Size

Table 4 presents the results obtained with different parameter sizes of LLama 2 (Touvron et al., 2023). We can see that larger LLM models consistently deliver better performance, with the 70B model achieving the best results, followed by the 13B model and, finally, the 7B model. However, it is important to note that the analysis also highlights the effectiveness of smaller models. For example, even though LLama 7B is only one-tenth the size of LLama 2 70B, it still manages to achieve 93% of the performance of the larger model. Similarly, LLama 13B, which is five times smaller than LLama 2 70B, attains 95% of its performance. This suggests that one can choose the LLM size based on specific computational constraints without sacrificing significant performance.

Method	de	es	fr	hi	th	Avg
Llama-7B	68.7	69.5	73.6	59.9	53.6	65.0
Llama-13B	71.5	71.3	75.8	59.3	54.4	66.5
Llama-70B	**72.3**	**71.3**	**77.7**	**66.7**	**59.8**	**69.6**

Table 4: **Impact of LLM size**: EM performance of semantic parsers trained on translated datasets improve with increasing the size of LLMs.

Method	de	es	fr	hi	th	Avg
MBART	71.0	71.3	73.7	64.4	55.8	67.2
Google	**72.3**	71.3	**77.7**	**66.7**	59.8	69.6
Oracle	72.0	**73.0**	77.6	65.5	**60.9**	**69.8**

Table 5: **Impact of Translation**: Results using Google translation approximate human translation (Oracle).

5.3. Impact of Machine Translation

In this analysis (Table 5), we assess the impact of different machine translation tools on the final performance of our method. We consider translated utterances from the off-the-shelf MBart-based model (Tang et al., 2020), Google translation[3], and the oracle setting using gold translated utterances from the dataset. We can see that there is a big improvement when using Google translation compared with the mBART by 2.4 EM score. However, the results of Google translation and gold translation are nearly identical, despite the latter involving human annotators. This shows the effectiveness of using Google translation in multilingual semantic parsing tasks compared to other machine translation methods, aligning with previous works (Shi et al., 2022; Li et al., 2014).

5.4. Error Analysis

In this section, we conduct an error analysis to assess the predictions made by our ZeLa framework compared to the gold data. We categorize errors into five main types: Wrong Intent, Wrong Slot Label, Wrong Slot Span, Extra Slot, and Missing Slot. Table 5.4 presents examples of these five error categories and their respective percentages in the total errors. The analysis revealed that the most frequent error category was "Wrong Span Prediction," where the schema was correctly predicted, but the span within each slot was incorrect.

Furthermore, we assessed our framework's performance in reducing the quantity of the above five error types compared to the TAF baseline (Figure 7). The analysis demonstrated substantial improvements, with the most significant enhancements observed in the "wrong slot label" category. We attribute this improvement to our augmentation approach, which employs symmetry CoT prompting. This approach only requires LLMs to predict label spans while already knowing the schema, instead of generating both the schema and the label spans as in the TAF method. This underscores the effectiveness of the proposed framework.

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

Error	Example
Wrong Slot Span (43.3%)	Utterance: Wer arbeitet bei Long John Silver 's ? Prediction: [IN:GET_CONTACT [SL:EMPLOYER Long John Silver]] Gold: [IN:GET_CONTACT [SL:EMPLOYER Long John Silver 's]]
Extra Slot (16.9%)	Utterance: Schick eine Videonachricht an den Smoothie - Chat Prediction: [IN:SEND_MESSAGE [SL:TYPE_CONTENT Videonachricht] [SL:GROUP Smoothie]] Gold: [IN:SEND_MESSAGE [SL:GROUP Smoothie]]
Wrong Intent (15.0%)	Utterance: Starte meinen Timer neu Prediction: [IN:RESTART_TIMER [SL:METHOD_TIMER Timer]] Gold: [IN:RESUME_TIMER [SL:METHOD_TIMER Timer]]
Missing Slot (14.7%)	Utterance: Spiele einen bestimmten Rap - Künstler Prediction: [IN:PLAY_MUSIC [SL:MUSIC_GENRE Rap]] Gold: [IN:PLAY_MUSIC [SL:MUSIC_GENRE Rap] [SL:MUSIC_TYPE Künstler]]
Wrong Slot Label (10.1%)	Utterance: Welche Neuigkeiten gibt es in der Musikbranche ? Prediction: [IN:GET_STORIES_NEWS [SL:NEWS_TYPE Neuigkeiten] [SL:NEWS_TOPIC Musikbranche]] Gold: [IN:GET_STORIES_NEWS [SL:NEWS_TYPE Neuigkeiten] [SL:NEWS_CATEGORY Musikbranche]]

Table 6: Examples of error categories on the dev set of MTOP dataset.

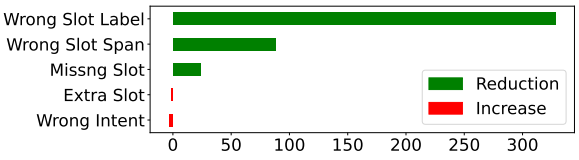

Figure 7: Enhancement observed across five error types in terms of quantity when comparing ZeLa with the baseline.

6. Conclusion

In conclusion, this paper presents the ZeLa framework, a versatile tool designed to significantly enhance zero-shot multilingual semantic parsing tasks by leveraging the inherent cross-lingual capabilities of large language models. Through a combination of an innovative LLM-based multilingual augmentation approach, and advanced multilingual semantic parsing techniques, our framework demonstrates outstanding performance on well-established multilingual semantic parsing datasets like MTOP and MASSIVE. ZeLa's performance surpasses state-of-the-art methods, achieving a notable 1.4-point increase in the exact match score on the MTOP

dataset. Moreover, our approach's adaptability extends its potential applications to a wide range of other natural language processing tasks, underscoring the promise of cross-lingual advancements in bolstering the resilience and effectiveness of natural language processing systems.

Acknowledgments

This work is supported partly by AOARD grant FA23862214039. The views and conclusions contained herein are those of the authors only and should not be interpreted as representing those of the U.S. Government.

7. Bibliographical References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11027–11052, Toronto, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi, and Partha Talukdar. 2023. Bootstrapping multilingual semantic parsers using large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2455–2467, Dubrovnik, Croatia. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Truong Do, Phuong Nguyen, and Minh Nguyen. 2023. StructSP: Efficient fine-tuning of task-oriented dialog system by using structure-aware boosting and grammar constraints. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10206–10220, Toronto, Canada. Association for Computational Linguistics.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. Filter: An enhanced fusion method for cross-lingual language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12776–12784.

Michael H Fischer, Giovanni Campagna, Euirim Choi, and Monica S Lam. 2021. Diy assistant: a multi-modal end-user programmable virtual assistant. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pages 312–327.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pages 30583–30598. Curran Associates, Inc.

Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci. 2022. CrossAligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4048–4061, Dublin, Ireland. Association for Computational Linguistics.

Sven Hartrumpf, Ingo Glöckner, and Johannes Leveling. 2008. Efficient question answering with question decomposition and multiple answer streams. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 421–428. Springer.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Haiying Li, Arthur C Graesser, and Zhiqiang Cai. 2014. Comparison of google translation with human translation. In *the twenty-seventh international flairs conference*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. 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, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Elman Mansimov and Yi Zhang. 2022. Semantic parsing in task-oriented dialog with recursive insertion-based encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11067–11075.

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, Online. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Massimo Nicosia, Zhongdi Qu, and Yasemin Altun. 2021. Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. 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, Online. Association for Computational Linguistics.

Peng Shi, Linfeng Song, Lifeng Jin, Haitao Mi, He Bai, Jimmy Lin, and Dong Yu. 2022. Cross-lingual text-to-SQL semantic parsing with representation mixup. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5296–5306, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Richard Shin and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5417–5425, Seattle, United States. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sarah Uhrig, Yoalli Garcia, Juri Opitz, and Anette Frank. 2021. Translate, then parse! a strong baseline for cross-lingual AMR parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 58–64, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Diyi Yang, Ankur Parikh, and Colin Raffel. 2022. Learning with limited text data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 28–31, Dublin, Ireland. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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, Dubrovnik, Croatia. Association for Computational Linguistics.