Efficiently Aligned Cross-Lingual Transfer Learning for Conversational Tasks using Prompt-Tuning

Anonymous ACL submission

Abstract

Cross-lingual transfer of language models trained on high-resource languages like English has been widely studied for many NLP tasks, but focus on conversational tasks has been rather limited. This is partly due to the high cost of obtaining non-English conversational data, which results in limited coverage. In this work, we introduce XSGD[1] for cross-lingual alignment pretraining, a parallel and large-scale multilingual conversation dataset that we created by translating the English-only Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2020) into 105 other languages. XSGD contains about 330k utterances per language. To facilitate aligned cross-lingual representations, we develop an efficient prompt-tuning-based method for learning alignment prompts. We also investigate two different classifiers: NLI-based and vanilla classifiers, and test cross-lingual capability enabled by the aligned prompts. We evaluate our model's cross-lingual generalization capabilities on two conversation tasks: slot-filling and intent classification. Our results demonstrate strong and efficient modeling ability of NLI-based classifiers and the large cross-lingual transfer improvements achieved by our aligned prompts, particularly in few-shot settings. We also conduct studies on large language models (LLMs) such as text-davinci-003 and ChatGPT in both zero- and few-shot settings. While LLMs exhibit impressive performance in English, their cross-lingual capabilities in other languages, particularly low-resource ones, are limited.

1 Introduction

It has long been known that NLP research and applications are concentrated on high-resource languages such as English, French, and Japanese. This limitation introduces bias and prevents people in minority language groups from accessing recent NLP technologies.

Driven by advances in large-scale training, there has been an increase in the number of approaches that attempt to learn general-purpose multilingual representations, which aim to capture shared knowledge across languages. Jointly trained multilingual language models such as XLM-R (Conneau et al., 2020) and mBART (Liu et al., 2020), coupled with supervised fine-tuning in the source (English) language, have been quite successful in transferring linguistic and task knowledge from one language to another without using any task labels in the target language, a.k.a. *zero-shot transfer*. Despite their effectiveness, studies (Wu and Dredze, 2019; Pires et al., 2019; K et al., 2020) have also highlighted key factors for successful transfer which include structural similarity between languages and the tasks under consideration. When it comes to conversational tasks, studies on cross-lingual zero-shot transfer have been limited to only few domains and languages.

To investigate the cross-lingual transfer ability on conversational tasks, we create the XSGD dataset by translating data from the English-only Schema-Guided Dialogue or SGD (Rastogi et al., 2020), which is currently the largest multi-domain dialogue corpora. While previous work such as Multi^{2}WOZ (Hung et al., 2022) has also tried to expand monolingual datasets into multiple languages, it is primarily a translation of development and test dialogues from the English-only MultiWOZ dataset (Budzianowski et al., 2018; Zang et al., 2020) into Arabic, Chinese, German, and Russian. In contrast, XSGD comprises 106 languages (including English), with roughly 330k utterances and 10 domains per language, as compared to the 7 domains and 29.5k utterances per language in Multi^{2}WOZ.

Recently, several studies (Li and Liang, 2021; Lester et al., 2021; Hambardzumyan et al., 2021)

[1] https://console.cloud.google.com/storage/browser/multilingual-sgd-data-research

have shown the potential of prompt tuning. In particular, Tu et al. (2022) observed that prompt tuning can achieve much better cross-lingual transfer than model fine-tuning across multiple XTREME tasks (Hu et al., 2020) using significantly fewer parameters. In this work, we propose an efficient prompt-tuning-based method that utilizes soft prompts to obtain stronger cross-lingually aligned representations on the XSGD dataset. The aligned prompts enable models to learn cross-lingual representations that can improve cross-lingual retrieval. Additionally, we compare the performance of vanilla and NLI-based formulations on intent classification task. The latter utilizes label descriptions or label names in conjunction with utterances for entailment prediction. We find that it exhibits stronger few-shot cross-lingual generalization capability for English-only tuning. Finally, our experimental results on intent classification and slot filling demonstrate consistent performance improvements with our learned aligned prompts, especially in few-shot settings.

Our contributions are summarized as follows:

- We have constructed a large parallel multilingual conversation corpus comprising 106 languages. We are releasing this dataset to facilitate and foster further research on multilingual conversation tasks.
- We have also introduced an efficient prompt-tuning-based approach for aligning sentence representations across multiple languages.
- We explored two different task formulations in the context of cross-lingual settings. We found that the NLI-based formulation demonstrated much stronger cross-lingual ability than the vanilla one, especially in few-shot settings.
- Our experiments showed that the aligned prompt we proposed is effective for cross-lingual transfer, particularly in the few-shot setting, where we observe significant gains. Our study also showns the benefits of our approach, even when compared to large language models (LLMs) such as text-davinci-003 and ChatGPT.

2 Background

2.1 Multilingual Models

Pre-trained multilingual language models, such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mBART (Liu et al., 2020) have demonstrated remarkable zero-shot cross-lingual transfer ability across a range of NLP tasks (Pires et al., 2019; Wu and Dredze, 2019). Moreover, some prior work, such as Artetxe and Schwenk (2019); Luo et al. (2021); Zhang et al. (2019), has leveraged parallel data to further enhance the cross-lingual transfer ability of these models through fine-tuning the entire architecture. Our work mainly explore a similar direction for conversation tasks, but with a more efficient approach where only a small portion of parameters are fine-tuned.

2.2 Cross-lingual Benchmarks

To evaluate zero-shot cross-lingual transfer ability, it is a standard practice to fine-tune the models exclusively on English tasks and then evaluate them on non-English test sets. XTREME (Hu et al., 2020) is a widely used benchmark in this regard, comprising four categories of tasks: sentence classification, structure prediction, question answering, and retrieval. For conversation tasks, the emerging benchmark is MASSIVE (FitzGerald et al., 2022), which includes around 1 million utterances across a range of languages[2].

3 XSGD Dataset

Prior work has focused on enhancing pre-trained language models (PLMs) for either deeper understanding of conversational contexts or improved cross-lingual generalization. For example, Wu et al. (2020) and Vulić et al. (2021) have explored adapting general-purpose English PLMs (Devlin et al., 2019; Liu et al., 2019) by applying conversation-specific training objectives on large-scale English conversational corpus.

One of the main challenges to achieve cross-lingual conversational capability is the lack of paired multi-lingual conversational corpus. In this work, we take the initiative on this challenge and create a multi-lingual dataset XSGD on top of the SGD dataset (Rastogi et al., 2020). To this end, we leverage Google Translate API [3] and translate the original SGD dataset into 105 languages. A complete list of the 105 languages can be found in Appendix A. We follow the same train, development, and test splits as in the original SGD dataset.

[2]Although this dataset does not contain any dialogue as our created dataset XSGD, it is of higher quality. As a result, we will be using it as a benchmark for downstream tasks.

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

Human Evaluation Our parallel dataset is the largest multilingual TOD corpus (330k per language), however, it inherits noise from the translation API. It is prohibitively expensive to do full-scale manual quality control because of its scale across 106 languages[4].

Languages	Human Evaluation	
	Fluency	Meaning
Indonesian	99%	98%
Swahili	100%	100%
Khmer	94%	99%
Urdu	97%	100%
Hawaiian	95%	99%
Yoruba	98%	100%

Table 1: Data quality results with Human evaluation.

We conduct human evaluation on 100 randomly sampled examples with workers from Amazon Mechanical Turk (AMT) on 6 low-resource languages (Indonesian, Swahili, Urdu, Khmer, Hawaiian, Yoruba) with different scripts[5]. Each sample is a translation pair that are randomly select consecutive turns within each dialogue. For quality control purpose, we set up a quiz to test Turkers's language skills. Each assignment is evaluated by three different Turkers. Turkers who passed the quiz are asked to evaluate the translation pairs based on 2 individual qualities (meaning and fluency): whether adequately expresses the meaning of English text, and whether the translated text is fluent. We provide our evaluation template of Hawaiian language in Figure 4 of Appendix. As shown in Table 1, we notice the high quality of our dataset. Surprisingly, at least 98% have the same meaning of English text.[6].

In next section, we show an efficient transfer learning method to use this large scale dataset for alignment pretraining. Then we further tune the aligned model on clean data with gold-labels so that noise will hopefully have a minor effect on our final model. Our evaluation dataset is also a high quality multilingual dataset.

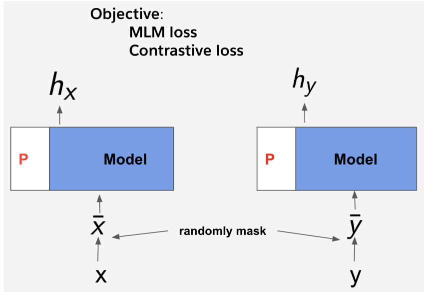

Figure 1: Framework for learning aligned prompts on multilingual conversational corpus. We denote P as the aligned prompts, which are tuned on the dialogue translation pairs, $\langle x, y \rangle$. The backbone model parameters are frozen. These aligned prompts are used for conversation downstream tasks.

4 Method

Recently, several studies (Li and Liang, 2021; Lester et al., 2021; Hambardzumyan et al., 2021) have shown that prompt tuning looks promising on many NLU tasks. More recently, Tu et al. (2022) observe that prompt tuning can achieve significantly better cross-lingual transfer than fine-tuning across several XTREME tasks (Hu et al., 2020), despite only tuning 0.1% to 0.3% of the parameters compared with whole model fine-tuning.

4.1 Aligned Prompts on Conversation Domain

In the zero-shot cross-lingual setting, models are fine-tuned solely on English and then evaluated on other languages. However, their performance on non-English languages, especially low-resource ones, tend to deteriorate (Hu et al., 2020; FitzGerald et al., 2022) . To address this issue, we propose a prompt-tuning-based method that utilizes translation data to learn aligned prompts, which can lead to improved cross-lingual transfer performance, especially when task data in English is limited.

Sequence Pairs Our dialogue corpus consists of dialogues with approximately 20 turns each. To reduce the sequence length of each dialogue during training, we randomly select consecutive turns within each dialogue in each epoch and concatenate them into a sequence. We repeat this process for the corresponding turns in the target language. We use this way to construct translation pairs dynamically during training, and then use the resulting translation pairs $\langle x_i, y_i \rangle$ from two different languages to learn aligned representations for an improved cross-lingual generalization capability[7].

[4]It is an interesting direction to explore how to improve the quality of this public dataset via an economically efficient way in the future, for example, Majewska et al. (2023).

[5]Two languages (Hawaiian, Yoruba) are not even supported by backbone model XLM-R

[6]We hypothesize the conversation domain is easier to get high translation quality.

[7]In our experiment, x is always English.

Masked Language Modeling (MLM) Loss This is a popular learning objective to learn deep bidirectional representations. MLM is defined based on the reconstruction loss of a certain percentage of randomly masked input tokens given the rest of the context. We leverage this loss to adapt backbone models to the conversation domain. We conduct token masking dynamically during batch training. Formally, the MLM loss is defined as:

$$\mathrm{L}_{mlm} =$$

$$-\frac{1}{M}\left(\sum_{x_m \in MX} \log prob(x_m) + \sum_{y_m \in MY} \log prob(y_m)\right)$$

where M is the total number of masked tokens in $\langle x, y \rangle$ and MX and MY are the masked tokens in x_i and y_i, respectively. $prob(x_m)$ and $prob(y_m)$ denote the probabilities of generating x_m and y_m from their corresponding masked tokens, respectively.

Contrastive Loss We leverage contrastive learning to enhance the representations. And it would not be possible without our parallel data XSGD, which unlocks the possibility of learning stronger cross-lingual representations via alignment objective formulated via contrastive loss. Figure 1 illustrates the process. In a mini-batch of translation pairs, for $\langle x, y \rangle$, the positive sample for masked x is the masked translation y. The negative samples are all the other translations $\hat{y}$ in the same mini-batch.

We first draw a batch of translation pairs. For each translation pair, we dynamically masked each sequence. The contrastive loss is

$$\mathrm{L}_{contra} = -\frac{1}{N}\left(\sum_{\langle h_x, h_y \rangle \in H} \log \frac{\exp(sim(h_x, h_y)/\tau)}{\sum_{y'} \exp(sim(h_x, h_{y'})/\tau)}\right)$$

where H is the translation representations of the batch, τ is the temperature term, N is the mini batch size, y' is from mini batch. h_x and h_y are the CLS token representations of masked sequence x and y respectively, sim is the similarity function. cosine similarity is used in our experiments. We set $\tau = 0.05$ in our experiments.

Total Loss The overall learning objective is the sum of $\mathrm{L}_{\mathrm{mlm}}$ and $\mathrm{L}_{\mathrm{contra}}$.

5 Experimental Setup

5.1 Datasets

SGD We use the Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2020) for intent classification. There are about 16K dialogues and 20 domains. For each domain, there are a different number of intents, services and dialogues. Each service provides a schema listing the supported intents along with their natural language descriptions. For example, service "payment" have two intents "MakePayment" and "RequestPayment". The description of an intent called "MakePaymen" is "Send money to your contact". Zero-shot evaluation is used, because lots of intents in the dev and test are unseen in the training set. For training, we only sample 5-shots per service as our training set and evaluate on the whole dev set. For cross-lingual evaluation, we use the translated utterance from XSGD[8].

MASSIVE We use MASSIVE (FitzGerald et al., 2022) as another dataset for evaluation[9]. There are 52 languages and about 1 million utterances in this dataset. For each language, there are about 11k train utterances, about 2k dev utterances, about 3K test utterances. We use this for evaluation on two conversation understanding tasks: intent classification and slot filling. There are 60 intents and 55 slot types. Accuracy and F1 score are the metrics for intent classification and slot filling, respectively.

5.2 Task Classifiers

Intent Classifiers We use [CLS] representation from the encoder as the sentence representation. Two different intent classifiers (NLI-based classifier and vanilla classifier) are considered in our experiments. Figure 2 shows more details.

Vanilla classifier uses the utterance representation to predict intent label. The learning and inference is done as a multi-label classifier.

NLI-based text classification has been investigated by (Qu et al., 2021), (Zhang et al., 2020) and (Yin et al., 2019) and proved to show superior performance in few-shot setting. In NLI-based text classification scenario, utterance and intent description or intent name are combined to make a prediction. During training, positive samples are formed by concatenating utterance and its intent description. Negative samples are constructed in the mini batch by sampling a negative intent description. To balance the training process, we keep the positive to negative ratio 1:1 for each batch. Cross-entropy loss is used during training. For inference, we select the label with largest entailment

[8]According to human evaluation results, we think it is reasonable to use them in some preliminary experiments.

[9]We use the version MASSIVE 1.1, which can be downloaded at https://github.com/alexa/massive.

Example (utterance, intent) = (“What can I do for you?” “I want to rent a movie.”, “RentMovie”)

Intent description “RentMovie” : “Find movies to watch by genre and, optionally, director or actors”

- **Option1** Vanilla classifier: utterance ⟹ label

  (“What can I do for you?” “I want to rent a movie.”, 10)

- **Option 2** NLI-based classifier: (utterance, intent description) ⟹1: entailment; 0:non-entailment)

  (“What can I do for you?” “I want to rent a movie.”, “Find movies to watch by genre and, optionally, director or actors”, 1)

Figure 2: Two different classifiers (NLI-based classifier and vanilla classifier) are proposed for intent classification task. For NLI-based classifier training, negative samples are constructed in the mini batch. English intent description are also used for the evaluation on the other languages. See more details in 5.2.

score. The prediction is correct if and only if the predicted label is correct and the largest entailment score is larger than 0.5 [10].

Slot Classifier Slot filling is treated as a token level classification task. We report F1 score for this task.

5.3 Training

For the backbone model, we use XLM-R (Conneau et al., 2020) in the most of experiments, which is a pretrained multilingual masked language model with 560M parameters on 2.5B of filtered data containing 100 languages. We also use XLM-RoBERTa-XL with 3.5B parameters in some settings. More details can be seen in Appendix B.

6 Aligned Prompts Results

In section 4, we propose a method that learns aligned prompts on conversation pair data in order to improve cross-lingual transfer ability. In this section, we show some aligned prompts results.

Retrieval Results To justify what are the learn for these aligned prompts, we perform similarity search on Tatoeba. With aligned prompts, we use the CLS token representation as the sentence representation, and do nearest-neighbor search. Figure 3 displays the Tatoeba test results for several languages. Notably, our results demonstrate that aligned prompts can achieve significantly higher retrieval accuracy, even when the prompt length is only 1. Furthermore, performance can be further improved with additional prompts; however, it is important to note that using too many prompts can actually hurt performance. In our subsequent experiments, the prompt length was set to 16, unless otherwise specified.

[10]The 0.5 threshold is for out-of-scope (OOS) prediction, which is required in the SGD dataset. The MASSIVE dataset doesn’t have OOS, so the threshold can be disregarded.

	non-conversation	conversation
5-shots	51.7 (1.1)	**55.2** (1.3)
15-shots	63.0 (0.5)	**66.5** (0.5)
all-shots	76.1 (0.6)	**77.7** (0.5)

Table 2: Cross-lingual transfer (Training only on English annotation data, and evaluate on all languages) performance (with standard deviation) on intent classification when using aligned prompts from two different domains: conversation and non-conversation. All results are averaged over all languages of 5 runs.

Conversation Pairs vs. Non-Conversation Pairs Previous works have utilized parallel corpora from non-conversational domains, such as OPUS (Tiedemann, 2012). To evaluate the effectiveness of XSGD, we randomly selected a parallel dataset from OPUS of a similar size and learned aligned prompts using the same method. Table 2 presents the results of intent classification on a conversation downstream task, demonstrating that the performance of aligned prompts on XSGD significantly outperforms that of the non-conversational domain dataset across different settings (5-, 15-, all-shots).

7 Downstream Tasks Results

In this section, we perform experiments on a conversation benchmark MASSIVE and report the performance results on all languages. We try the following three tuning methods.

Fine-tuning (FT): In this setting, all available parameters are tunable.

Prompt Tuning (PT): For prompt tuning, the backbone model is fixed, only a small number of parameters (prompts) and task classifiers parameters are updated. We use continuous prompts and layer prompts (Li and Liang, 2021; Liu et al., 2022).

Aligned Prompt Tuning (APT): With the parallel translation data, we can learn aligned prompt for

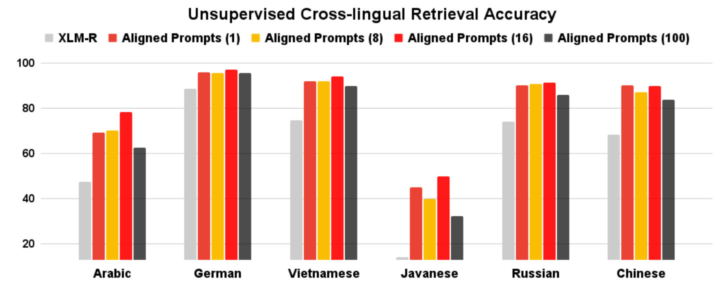

Figure 3: Unsupervised cross-lingual retrieval results (accuracy) for several linguistically diverse selected languages. The backbone model for these aligned prompts are XLM-R models. The length of prompts is 1, 8, 16, 100 respectively. XLM-R results are token from Hu et al. (2020).

aligned cross-lingual representation in Section 4. These prompts can be used for a warm-up start for these downstream task with prompt learning.

	en	zh-TW	zh-CN	ja	ko	AVG
NLI-based Classifier						
5-shots	47.8	29.1	31.3	25.7	38.3	**36.1**
15-shots	70.8	51.8	53.1	43.5	61.8	**58.3**
all	89.9	65.0	69.4	54.3	83.7	77.3
Vanilla Classifier						
5-shots	9.4	3.6	4.4	4.2	6.6	5.6
15-shots	10.2	13.7	13.7	9.2	11.5	9.9
all	90.6	69.6	71.1	53.7	84.0	**78.8**

Table 3: Averaged accuracy (%) of the NLI-based classifier and the vanilla classifier on the MASSIVE intent classification task when fine-tuning on English only and evaluating on all 52 languages.

7.1 Intent Classification

	en	zh-TW	zh-CN	ja	ko	AVG
5-shots						
FT	9.4	3.6	4.4	4.2	6.6	5.9 (3.3)
PT	51.3	17.0	16.8	15.3	30.8	24.9 (11.5)
APT	**65.2**	**49.3**	**52.1**	**38.5**	**59.3**	**55.2** (1.3)
15-shots						
FT	10.2	13.7	13.7	9.2	11.5	28.7 (17.4)
PT	75.8	50.2	56.5	43.6	63.7	58.2 (2.3)
APT	**78.0**	**59.1**	**62.9**	**47.7**	**71.7**	**66.5** (0.5)
all						
FT	**90.6**	**69.6**	**71.1**	53.7	84.0	**78.8** (0.5)
PT	89.7	63.9	68.2	**55.6**	82.1	76.8 (0.1)
APT	90.1	67.7	70.5	54.5	**84.4**	77.7 (0.5)

Table 4: Accuracy (%) of vanilla classifier on MASSIVE intent classification task when training on English only and evaluate on all 52 languages.

Fine Tuning Table 3 shows the performance of the fine-tuned XLM-R model on English. Both of the intent classifiers achieve higher performance with more data. In few-shot experiments, the NLI-based classifier outperforms the vanilla classifier by a significant margin. The average performance on all 52 languages reaches 58.3% accuracy with only 15 samples per intent. However, the vanilla classifier works better with the full data.

	en	zh-TW	zh-CN	ja	ko	AVG
5-shots						
FT	47.8	29.1	31.3	25.7	38.3	24.2 (6.8)
PT	59.9	38.7	40.0	30.0	49.4	38.1 (16.5)
APT	**69.8**	**51.1**	**52.4**	**45.4**	**64.8**	**59.8** (1.6)
15-shots						
FT	70.8	51.8	53.1	43.5	61.8	46.0 (11.9)
PT	75.8	54.8	57.8	43.5	68.7	60.3 (2.6)
APT	**89.7**	**58.5**	**62.8**	**51.8**	**75.0**	**67.5** (1.1)
all						
FT	89.9	65.0	**69.4**	**54.3**	83.7	76.8 (0.6)
PT	89.7	56.4	56.4	36.0	83.9	75.6 (0.4)
APT	**90.2**	**66.1**	68.4	52.0	**85.2**	**78.9** (0.2)

Table 5: Accuracy (%) of NLI-based classifier on MASSIVE intent classification task when training on English only and evaluate on all 52 languages.

Vanilla Classifier In Table 4, we observe poor performance on few-shot settings for vanilla classifiers on intent tasks. However, significant gains are achieved with our method (from 5.9% to 24.9% on 5-shots and from 28.7% to 58.2% on 15-shots). We also observe that aligned prompts can further improve performance, with the best results obtained in few-shot settings. Additionally, the variances in task performance across all languages with aligned prompts are significantly smaller than fine-tuning and prompt tuning only. Although prompt tuning achieves higher accuracy on few-shot settings than fine-tuning, there is still a small gap, even with aligned prompts and full data training.

NLI-based Classifier An advantage of using NLI-based classifiers is their ability to evaluate unseen intent labels if their descriptions are known. Additionally, we demonstrate strong performance on the SGD dataset. In Table 5, we present the results of fine-tuning with prompt tuning and aligned prompts for the MASSIVE dataset. With aligned prompts, we achieve strong accuracy results of 59.8% on 5-shots and 67.7% on 15-shots. Moreover, the English result on 15-shots with aligned prompts is comparable to the result obtained from full data training. These findings suggest that NLI-based classifiers with aligned prompts can efficiently learn with few samples. Aligned prompts consistently outperform other methods in this setting, indicating strong modeling ability and cross-lingual transfer ability.

LLMs Results We conducted experiments using both ChatGPT and the latest GPT-3.5 model (text-davinci-003 as of May, 2023) from OpenAI. We sampled 100 examples for each language and used the prompts provided in the Appendix. In the few-shot setting, the in-context examples were taken from the English partition. The intent classification results are presented in Table 6. The text-davinci-003 model showed significant improvements as more in-context examples were included, however, the ChatGPT model only demonstrated improvement in English. The cross-lingual ability of ChatGPT was found to be even worse, which led us to hypothesize that the data used to train ChatGPT is predominantly in English. Based upon these results, we can draw a conclusion that cross-lingual is still challenging in the era of LLMs, and smaller models still have an advantage over LLMs for the ability to quickly adapt into new domains through fine-tuning or prompt-tuning.

	en	AVG
text-davinci-003		
zero-shot	59.0	40.8
1-shot	71.0	51.2
5-shot	83.0	**54.6**
ChatGPT		
zero-shot	63.0	**54.6**
1-shot	76.0	51.2
5-shots	**87.0**	51.3

Table 6: Accuracy (%) of ChatGPT and text-davinci-003 on MASSIVE intent classification task.

Takeaway Upon analyzing the results presented in Tables 4 and 5, we can observe significant improvements with aligned prompts as compared to prompting tuning alone. For instance, the improvements for vanilla classifiers are 30.3%, 8.3%, and 0.9% for 5-shots, 15-shots, and full data training, respectively. Similarly, for NLI-based classifiers, the gains are 11.7%, 7.2%, and 3.3% for the same settings. We note that there is a clear trend where the gain of cross-lingual transfer ability decreases as more English training data is used. Furthermore, NLI-based classifiers exhibit superior cross-lingual transfer ability, particularly in the few-shot setting.

7.2 Slot Filling

Table 7 shows the evaluation results for slot filling using the XLM-R backbone model. Our models were trained solely on English data, but we report the results for all languages. However, the fine-tuned models' results for Chinese and Japanese are significantly worse than those for English. In fact, the gaps are much larger than those in a similar setting for the intent classification task. This observation suggests that slot filling is considerably more challenging than intent classification.

The performance differences between fine-tuning and prompt tuning for all languages averaged across are 6.4%, -3.4%, and -6.2%, respectively. These results indicate that fine-tuning is more effective for improving slot filling performance than prompt tuning. However, this also suggests that there is still room for improvement for the current prompt-based methods.

With aligned prompts, we achieve consistent improvements over 5 runs, with gains of 4.5%, 1.3%, and 0.1% in the averaged F1 score. These results are consistently better, but the improvements are smaller as the training dataset size increases.

	en	AVG
5-shots		
FT	41.0	27.8 (3.3)
PT	59.5	34.2 (1.2)
APT	**62.6**	**38.7** (0.9)
15-shots		
FT	70.7	**49.0** (1.1)
PT	70.9	45.6 (0.9)
APT	**72.4**	46.9 (1.2)
all		
FT	**83.9**	**61.6** (1.0)
PT	83.3	55.4 (0.1)
APT	83.5	55.5 (0.5)

Table 7: Slot filling F1 (%) results on MASSIVE benchmark when training on English only and evaluate on all 52 languages.

XLM-R-XL and OpenAI API Results To test the limits of the prompt tuning method, we conducted experiments using prompt tuning and aligned prompts. Initially, we learned the aligned prompts on parallel XSGD data with a similar setting, where the prompt length is 16 and the backbone model is XLM-R-XL.

Table 7 and Table 8 displays the results of prompt tuning and aligned prompts on these settings. There are significant performance gains, particularly for aligned prompts. When scaling up the backbone model size from XLM-R to XLM-R-XL, the improvements with aligned prompts are 5.2% and 5.0% for 15-shots and full English data, respectively. Meanwhile, the improvements with prompt tuning are only 1.0% and 0.5%. This finding indicates that aligned prompts provide better modeling ability when increasing the backbone model size.

For the experiments with OpenAI models, we adapted prompts from Qin et al. (2023). More details about the prompts and results are available in the Appendix. Overall, LLMs exhibit poor performance in the slot filling task, with an average F1 score ranging from 3% to 6% across all languages.

	en	zh-TW	zh-CN	ja	AVG
15 shots					
PT	71.7	9.2	10.1	5.1	46.6 (1.9)
APT	**73.3**	**20.5**	**22.1**	**13.2**	**52.1** (0.5)
all					
PT	**83.1**	14.3	14.9	9.4	55.9 (0.7)
APT	82.8	**22.9**	**23.6**	**11.7**	**60.5** (0.7)

Table 8: Averaged Slot filling F1 (%) results with 5 runs on MASSIVE benchmark when training on English only and evaluate on all 52 languages. The prompt lengths is 16. XLM-RoBERTa-XL is used as the backbone model.

Discussion We observe gains in cross-lingual ability with aligned prompts. However, there is still room for future improvements. The gains achieved with current aligned prompts methods are smaller than those achieved in few-shot settings. Also, the prompt tuning method on complex tasks, such as slot filling, still lags behind the fine-tuning method. These observations suggest that further research is needed to explore how to design more sophisticated and efficient methods for cross-lingual transfer.

8 Related Work

Methods for Cross-lingual Transfer In recent years, many cross-lingual methods have been developed for non-conversational tasks using parallel data. However, continued pretraining on parallel data has been found to improve retrieval performance by making the pre-training task more similar to the downstream setting, but does not lead to a significant improvement in performance on other tasks (Luo et al., 2021; Chi et al., 2021; Zhang et al., 2019). These methods often require updating all model parameters or using larger scale monolingual corpora that cover all languages, which can make them difficult to use with large language models. In this work, we used a prompt-tuning-based method that only tunes few prompts and achieved significant gains in few-shot settings. We believe that more sophisticated work in this direction can be done in the future.

Resources for Multilingual Conversation One of the fundamental objectives of artificial intelligence is to enable machines to communicate with humans. To achieve this, annotated conversation corpora are crucial. Conversation datasets have evolved from single-domain ones such as ATIS (Price, 1990) to more complex and diverse ones such as MultiWOZ (Budzianowski et al., 2018) and SGD (Rastogi et al., 2020). In recent years, several multilingual conversation datasets have been proposed to develop multilingual conversational models. However, most existing conversation systems are predominantly built for English or a few other major languages. For example, Schuster et al. (2019) introduced an annotation corpus of 57k utterances in English (43k), Spanish (8.6k), and Thai (5k) across three domains. Multi^{2}WOZ dataset (Hung et al., 2022) is much larger annotation corpus with five languages (including English) and 29.5k utterances per language. Due to high cost for collecting multilingual conversation data, Ding et al. (2022) introduces a novel data curation method for creating GlobalWoZ with 20 languages. In this work, we have created a new parallel multilingual dataset called XSGD by translating the English-only Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2020) into 106 different languages. Although this dataset may contain some noise due to the translation process, we think it is a valuable resource for researchers interested in exploring multilingual conversational tasks.

9 Conclusion

In this paper, we present XSGD, a large-scale parallel multilingual conversation corpus that can be used for aligned cross-lingual transfer. Additionally, we propose a prompt-tuning method to learn alignment prompts, which can further improve the efficiency of the cross-lingual transfer. We evaluate our approach on intent classification and slot-filling tasks, and our experiments demonstrate its effectiveness. We also study popular LLMs and find that their performance on non-English languages remain to be improved.

Limitations

Although the translated data can be a little noisy, in our work, we did not mainly use the data directly on downstream tasks. Instead, we propose an efficient transfer learning method to use this large scale dataset for alignment pretraining. Then we further tune the aligned model on clean data with gold-labels so that noise will hopefully have a minor effect on our final model. Our evaluation dataset is also a high quality multilingual TOD dataset. So the proposed method and conclusion are still solid.

When conducting experiments with the OpenAI API, the large number of intent types (60) and slot types (55) posed a challenge in designing effective prompts. To address this, we conducted surveys and explored various prompt templates based on the works of Bang et al. (2023); Qin et al. (2023); Lai et al. (2023), among others. However, it is possible that we may have overlooked some potential prompt templates. There is room for improving the performance of text-davinci-003 and ChatGPT in future iterations.

We acknowledge that there are other parameter-efficient tuning techniques (Houlsby et al., 2019; Hu et al., 2022; Ben Zaken et al., 2022) and other LLMs, such as BLOOM (Scao et al., 2022) and LLamA (Touvron et al., 2023). It is however non-trivial to compare against different parameter efficient methods on various different LLMs, which requires a significant amount of GPU hours and can warrant a paper by itself. Our contribution includes the massive XSGD multilingual data and an effective prompt-tuning based alignment method. We leave the exploration of other methods as future work.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. 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 3418–3430, Online. Association for Computational Linguistics.

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.

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

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657, Dublin, Ireland. Association for Computational Linguistics.

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. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. 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 4921–4933, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Ponzetto, and Goran Glavaš. 2022. Multi2WOZ: A robust multilingual dataset and conversational pre-training for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703, Seattle, United States. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual {bert}: An empirical study. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *ArXiv*, abs/2304.05613.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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. *CoRR*, abs/1907.11692.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. 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 3980–3994, Online. Association for Computational Linguistics.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

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

P. J. Price. 1990. Evaluation of spoken language systems: the ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *ArXiv*, abs/2302.06476.

Jin Qu, Kazuma Hashimoto, Wenhao Liu, Caiming Xiong, and Yingbo Zhou. 2021. Few-shot intent classification by gauging entailment relationship between utterance and semantic label. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 8–15, Online. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 8689–8696.

Teven Le Scao, Angela Fan, and al Christopher Akiki etc. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1168, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082, Online. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

A Languages Except English on XSGD

List of 105 language ISO-639 code (`https://cloud.google.com/translate/docs/languages`) translated through Google Translate API (English is not included): af, am, ar, az, be, bg, bn, bs, ca, ceb, co, cs, cy, da, de, el, eo, es, et, eu, fa, fi, fr, fy, ga, gd, gl, gu, ha, haw, he, hi, hmn, hr, ht, hu, hy, id, ig, is, it, ja, ka, kk, km, kn, ko, ku, ky, la, lb, lo, lt, lv, mg, mi, mk, ml, mn, mr, ms, mt, my, ne, nl, no, ny, or, pa, pl, pt, ro, ru, rw, si, sk, sl, sm, sn, so, sq, sr, st, su, sv, sw, ta, te, tg, th, tk, tl, tr, tt, ug, uk, ur, uz, vi, xh, yi, yo, zh-CN, zh-TW, zu

B Licenses of Datasets

- SGD (Rastogi et al., 2020): Attribution-ShareAlike 4.0 International Public License.
- Massive (FitzGerald et al., 2022): Apache License.
- XSGD created by us: Attribution-ShareAlike 4.0 International.

C More Training Details

For the aligned prompts learning, we use Adam optimizer (Kingma and Ba, 2015) with warm up rate 0.1 and learning rate e−3. The number of epoch is 10. The mini-batch size are 64 and 32 for XLM-R and XLM-RoBERTa-XL, respectively.

On the conversation downstream tasks, we tune the learning rate in $\{0.1, 5e-2, 2e-2, 1e-2, 5e-3, 2e-3, 1e-3\}$. For experiments on XSGD, we do fine-tuning for 3 epochs and prompt-tuning for 30 epochs. For Massive benchmark, we fine tuning on intent classification and slot filling task for 30 epochs. For prompt tuning, the max number of epoch is 1000. We do early stopping based on performance on the English dev set. 1 A100 GPU with 40G memory is used for experiments. And most experiments are done in one day.

D Ablation Study on Learning Objectives

An ablation study was conducted to analyze the learning losses for three different settings: prompt tuning (PT), aligned prompts (APT), and APT (with MLM only). The results on XSGD are shown in Figure 9, while the results on MASSIVE intent classification can be seen in Figure 10.

	en	hi	ms	vi	gd	tg	AVG
Prompt Tuning							
l = 16	97.2	94.3	94.2	94.6	86.4	74.7	90.0
Aligned Prompts							
	97.7	**95.5**	**95.7**	**95.2**	**89.7**	**75.3**	**91.4**
Aligned Prompts (w/ MLM only)							
	96.8	93.3	93.1	92.7	88.5	75.0	89.7

Table 9: Intent classification accuracy (%) on XSGD. Here we select some languages, which are in different language family or low-resourced.

E Prompt Templates and Results

Prompt templates in experimental settings. **[schema]** and **[utt]** are the intent set and the raw utterance text respectively. And utt1, label1, utt2, label2 are in-context examples.

	en	AVG
5-shots		
PT	51.3	24.9 (11.5)
APT	**65.2**	**55.2** (1.3)
APT (w/ MLM only)	61.9	30.9 (7.1)
15-shots		
PT	75.8	58.2 (2.3)
APT	**78.0**	**66.5** (0.5)
APT (w/ MLM only)	78.2	61.2 (1.8)

Table 10: Accuracy (%) of vanilla classifier on MASSIVE intent classification task when training on English only and evaluate on all 52 languages.

Intent Classification Task

Zero-shot Setting

```
Please tell me the
intent of the following
utterance:[utt] given the
intent set [schema]
```

Few-shots Setting

```
Given the intent set
[schema], please tell
me the intent of the
following utterances.

utt1
label1
utt2
label2
...
utt
```

Slot Filling Task

```
Please identify slots s
from the given text.  The
text from utt with slot
annotations is formatted
as [label :  entity] .

Text:[utt]
Slot:
```

F Amazon Mechanical Turk Template

G XSGD

Table 13 shows the intent classification results when training on English-only data and evaluating on all languages. We find that prompt tuning

Please identify slots app_name, currency_name, radio_name, email_folder, relation, sport_type, media_type, music_genre, drink_type, ingredient, time_zone, game_name, weather_descriptor, coffee_type, podcast_name, general_frequency, transport_type, time, playlist_name, transport_descriptor, movie_name, cooking_type, place_name, device_type, email_address, change_amount, timeofday, audiobook_name, joke_type, game_type, transport_agency, event_name, song_name, artist_name, order_type, person, player_setting, house_place, business_name, food_type, music_album, meal_type, definition_word, podcast_descriptor, transport_name, audiobook_author, date, movie_type, music_descriptor, list_name, news_topic, color_type, Other, personal_info, business_type, alarm_type from the given text. The text from utt with slot annotations is formatted as [label : entity].

Text: weck mich diese woche um fünf uhr morgens auf
Slot:
app_name : weck, currency_name : None, radio_name : None, email_folder : None, relation : None, sport_type : None, media_type : None, music_genre : None, drink_type : None, ingredient : None, time_zone : None, game_name : None, weather_descriptor : None, coffee_type : None, podcast_name : None, general_frequency : None, transport_type : None, time : fünf uhr morgens, playlist_name : None, transport_descriptor : None, movie_name : None, cooking_type : None, place_name : None, device_type : None, email_address : None, change_amount : None, timeofday : morgens, audiobook_name : None, joke_type : None, game_type : None, transport_agency : None, event_name : None, song_name : None, artist_name : None, order_type : None, person : None, player_setting : None, house_place : None, business_name : None, food_type : None, music_album : None, meal_

Table 11: One example input and output pair for slot filling. The utterance and OpenAI API response are colored in green and blue, respectively.

has better cross-lingual transfer ability and aligned prompts further improve the performance.

Figure 5 in the Appendix presents performance comparison of the three different methods (FT: fine-tuning; PT: prompt tuning; APT: aligned prompt tuning). The figure indicates that prompt tuning outperforms fine-tuning, while aligned prompt tuning achieves the best performance. However, the models still struggle with some low-resource languages, especially those that are not supported by the backbone model XLM-R (e.g., haw (Hawaiian), yo (Yoruba), tk (Turkmen), sn (Shona)).

Languages	Intent Classification				Slot Filling	
	text-davinci-003 zero-shot	ChatGPT zero-shot	text-davinci-003 5-shots	ChatGPT 5-shots	text-davinci-003 zero-shot	ChatGPT zero-shot
	Acc.	Acc.	Acc.	Acc.	F1	F1
Afrikaans	52	62	64	49	10.3	5.4
Amharic	5	14	13	8	0.0	0.0
Arabic	45	62	66	57	8.5	5.5
Azerbaijani	33	48	61	40	5.3	1.9
Bengali	32	56	45	46	3.0	1.9
Catalan	45	64	55	52	6.6	6.1
Welsh	21	31	34	21	2.9	2.0
Danish	62	70	72	65	12.7	5.3
German	55	76	76	72	13.6	5.4
Greek	45	66	67	75	7.9	3.7
English	59	63	83	87	23.8	1.6
Spanish	52	65	67	58	10.7	10.4
Persian	39	70	66	65	5.4	1.9
Finnish	45	62	62	49	5.3	3.5
French	54	78	77	73	12.9	8.8
Hebrew	42	64	60	55	1.6	0.0
Hindi	35	63	60	63	7.1	1.9
Hungarian	55	64	66	53	3.6	2.0
Armenian	11	26	21	22	0.0	5.5
Indonesian	55	60	70	63	11.1	1.9
Icelandic	46	57	49	40	4.7	3.6
Italian	60	66	67	63	6.0	5.3
Japanese	53	70	66	66	1.8	0.0
Javanese	19	15	25	21	1.6	0.0
Georgian	13	22	21	28	0.0	0.0
Khmer	15	22	34	18	4.3	2.0
Kannada	17	41	26	50	3.4	0.0
Korean	55	72	74	75	3.2	4.0
Latvian	41	49	52	41	1.7	7.2
Malayalam	17	40	27	40	1.6	5.6
Mongolian	14	24	30	25	0.0	0.0
Malay	51	49	66	55	11.7	1.9
Burmese	0	8	13	10	0.0	0.0
Norwegian	51	66	67	63	14.3	6.8
Dutch	63	71	71	64	12.8	5.8
Polish	60	64	71	68	13.2	1.8
Portuguese	53	62	65	60	14.5	10.5
Romanian	54	63	65	55	3.3	12.3
Russian	56	72	64	71	5.6	5.4
Slovenian	56	61	59	57	7.6	3.9
Albanian	39	41	47	35	6.2	2.0
Swedish	59	75	66	69	9.8	3.5
Swahili	21	47	27	34	0.0	3.6
Tamil	17	29	37	32	0.0	0.0
Telugu	22	33	32	31	0.0	0.0
Thai	50	62	69	69	3.5	4.0
Tagalog	49	58	59	51	10.1	6.2
Turkish	46	65	67	57	9.8	1.9
Urdu	18	52	30	46	3.5	2.0
Vietnamese	45	65	65	64	10.9	3.6
Simplified Chinese	60	75	74	64	0.0	0.0
Traditional Chinese	57	70	71	71	0.0	0.0

Table 12: The performance results of the OpenAI API using our prompts are presented. 100 examples are sampled for each language. For the slot filling task, the prompt used is adapted from Qin et al. (2023). It should be noted that due to the large number of slot types (55), the slot results are not satisfactory.

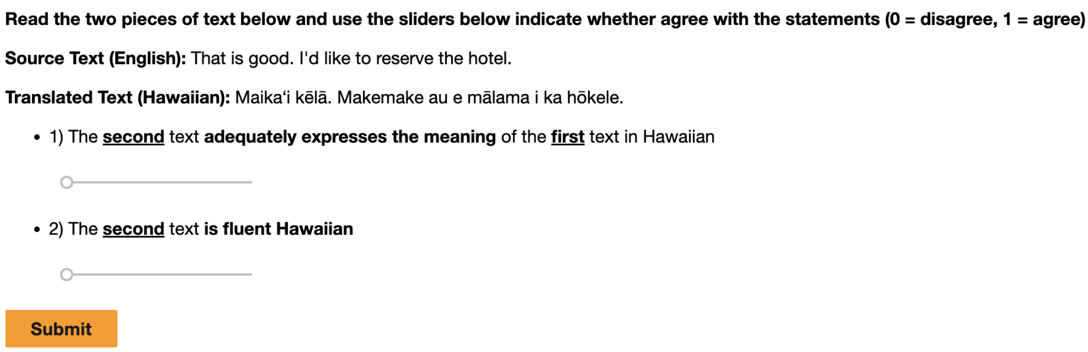

Figure 4: Human evaluation template for our dataset.

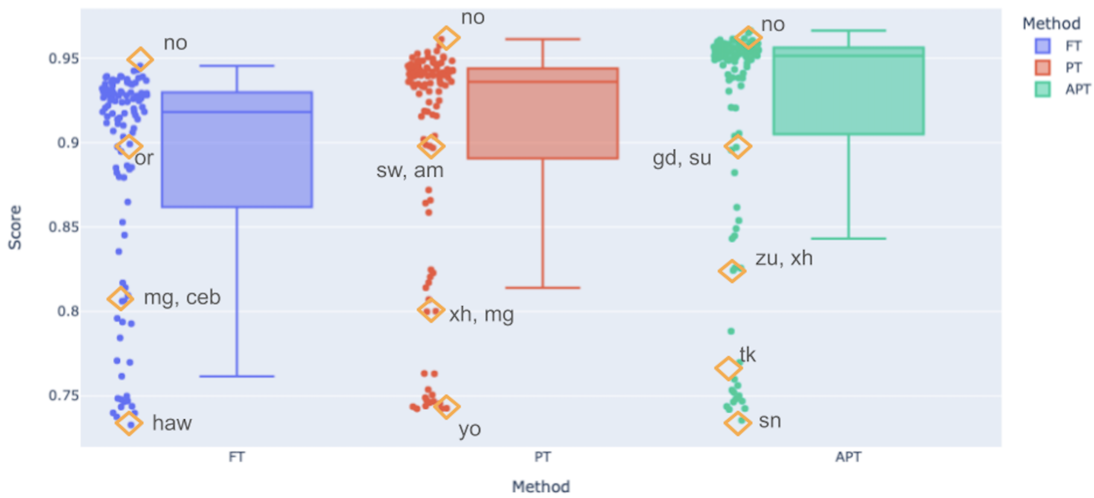

Figure 5: Intent classification performance of different models (FT: fine-tuning; PT: prompt tuning; APT: aligned prompt tuning) over all languages on XSGD. The scores represent the accuracy of each language. We can see the models are still struggled with languages that are not supported by the backbone model XLM-R.

	en	hi	ms	vi	gd	tg	AVG
Fine Tuning							
	95.7	92.8	93.2	93.9	84.5	75.0	88.6
Prompt Tuning							
l = 4	93.6	90.8	90.7	90.5	83.7	74.5	87.5
l = 8	96.2	94.4	93.8	94.7	85.8	74.3	89.8
l = 16	97.2	94.3	94.2	94.6	86.4	74.7	90.0
Aligned Prompts							
	97.7	**95.5**	**95.7**	**95.2**	**89.7**	**75.3**	**91.4**

Table 13: Intent classification accuracy (%) on XSGD. Here we select some languages, which are in different language family or low-resourced. The monolingual training corpus size of "gd" for backbone model XLM-R is small (∼0.1 GB). "tg" (Tajik) is also not supported by the backbone model.