GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems

Anonymous ACL submission

Abstract

Over the last few years, there has been a move towards data curation for multilingual task-oriented dialogue (ToD) systems that can serve people speaking different languages. However, existing multilingual ToD datasets either have a limited coverage of languages due to the high cost of data curation, or ignore the fact that dialogue entities barely exist in countries speaking these languages. To tackle these limitations, we introduce a novel data curation method that generates **GlobalWoZ** — a large-scale multilingual ToD dataset globalized from an English ToD dataset for three unexplored use cases of multilingual ToD systems. Our method is based on translating dialogue templates and filling them with local entities in the target-language countries. Besides, we extend the coverage of target languages to 20 languages. We will release our dataset and a set of strong baselines to encourage research on multilingual ToD systems for real use cases.

1 Introduction

One of the fundamental objectives in pursuit of artificial intelligence is to enable machines with the ability to intelligently communicate with human in natural languages, with one of the widely-heralded applications being the task-oriented dialogue (ToD) systems (Gupta et al., 2006; Bohus and Rudnicky, 2009). Recently, ToD systems have been successfully deployed to assist users with accomplishing certain domain-specific tasks such as hotel booking, alarm setting or weather query (Eric et al., 2017; Wu et al., 2019; Lin et al., 2020; Zhang et al., 2020), thanks to the joint advent of neural networks and availability of domain-specific data. However, most existing ToD systems are predominately built for English, limiting their service for *all* of the world's citizens. The reason of this limitation lies in the stark lack of high-quality multilingual ToD datasets due to the high expense and challenges of human annotation (Razumovskaia et al., 2021).

One solution to this is annotating conversations in other languages from scratch, e.g., CrossWoZ (Zhu et al., 2020) and BiToD (Lin et al., 2021). However, these methods involve expensive human efforts for dialogue collection in the other languages, resulting in a limited language coverage. The other major line of work focused on translating an existing English ToD dataset into target languages by professional human translators (Upadhyay et al., 2018; Schuster et al., 2019; van der Goot et al., 2021; Li et al., 2021). Despite the increasing language coverage, these methods simply translated English named entities (e.g., location, restaurant name) into the target languages, while ignored the fact that these entities barely exist in countries speaking these languages. This hinders a trained ToD system from supporting the real use cases where a user looks for local entities in a target-language country. For example in Figure 1, a user may look for the British Museum when traveling to London (A.), while look for the Oriental Pearl Tower when traveling to Shanghai (B.).

In addition, prior studies (Cheng and Butler, 1989; Kim, 2006) have shown that code-switching phenomena frequently occurs in a dialogue when a speaker cannot express an entity immediately and has to alternate between two languages to convey information more accurately. Such phenomena could be ubiquitous during the cross-lingual and cross-country task-oriented conversations. One of the reasons for code-switching is that there are no exact translations for many local entities in the other languages. Even though we have the translations, they are rarely used by local people. For example in Figure 1 (C.), after obtaining the recommendation from a ToD system, a Chinese speaker traveling to London would rather use the English entity "British Museum" than its Chinese translation to search online or ask local people. To verify this code-switching phenomena, we have also conducted a case study (§6.1) which shows that

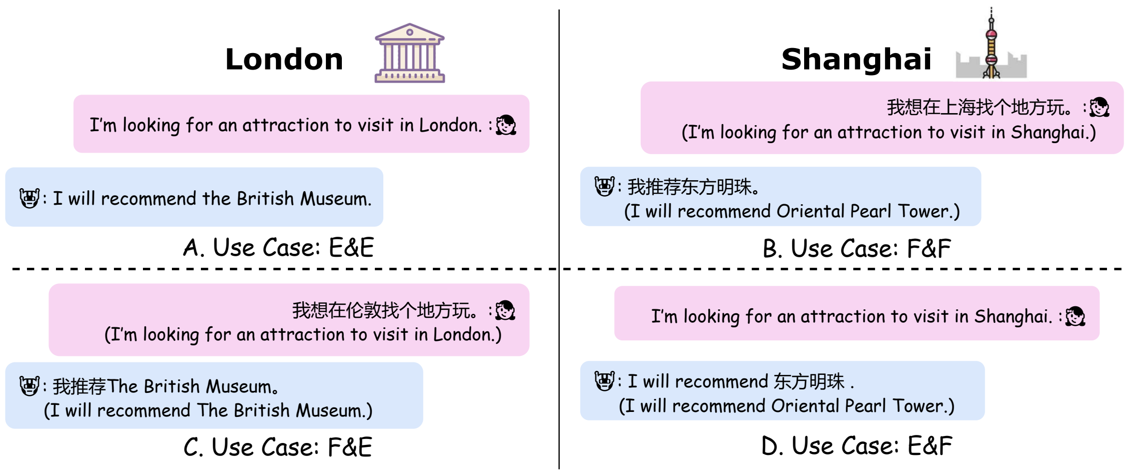

Figure 1: Examples of four use cases for multilingual ToD systems.

searching the information about translated entities online yields a much higher failure rate than searching them in their original languages. Motivated by these observations, we define *three unexplored use cases*[1] of multilingual ToD where a foreign-language speaker uses ToD in the foreign-language country (**F&F**) or an English country (**F&E**), and an English speaker uses ToD in a foreign-language country (**E&F**). These use cases are different from the traditional **E&E** use case where an English speaker uses ToD in an English-speaking country.

To bridge the aforementioned gap between existing data curation methods and the real use cases, we propose a novel data curation method that *globalizes* an existing multi-domain ToD dataset beyond English for the three unexplored use cases. Specifically, building on top of MultiWoZ (Budzianowski et al., 2018) — an English ToD dataset for dialogue state tracking (DST), we create GlobalWoZ, a new multilingual ToD dataset in three new target-languages via machine translation and crawled ontologies in the target-language countries.

Our method only requires minor human efforts to post-edit a few hundred machine-translated dialogue templates in the target languages for evaluation. Besides, as cross-lingual transfer via pre-trained multilingual models (Devlin et al., 2019; Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021) has proven effective in many cross-lingual tasks, we further investigate another question: *How do these multilingual models trained on the English ToD dataset transfer knowledge to our globalized dataset?* To answer this question, we prepare a few baselines by evaluating popular ToD systems on our created test datasets in a *zero-shot* cross-lingual transfer setting as well as a *few-shot* setting.

Our contributions include the following:

- To the best of our knowledge, we provide the first step towards analyzing three unexplored use cases for multilingual ToD systems.
- We propose a cost-effective method that creates a new multilingual ToD dataset from an existing English dataset. Our dataset consists of high-quality test sets which are first translated by machines and then post-edited by professional translators in three target languages (Chinese, Spanish and Indonesian). We also leverage machine translation to extend the language coverage of test data to another 17 target languages.
- Our experiments show that current multilingual systems and translate-train methods fail in zero-shot cross-lingual transfer on the dialogue state tracking task. To tackle this problem, we propose several data augmentation methods to train strong baseline models in both zero-shot and few-shot cross-lingual transfer settings.

2 Data Curation Methodology

In order to globalize an existing English ToD dataset for the three aforementioned use cases, we propose an approach consisting of four steps as shown in Figure 2: (1) we first extract dialogue templates from the English ToD dataset by replacing English-specific entities with a set of general-purpose placeholders (§2.1); (2) we then translate the templates to a target language for both training and test data, with one key distinction that we only post-edit the test data by professional translators to ensure the data quality for evaluation (§2.2); (3) next, we collect ontologies (Kiefer et al., 2021) containing the definitions of dialogue acts, local

[1]See comparisons of these use cases in Appendix A

entities and their attributes in the target-language countries (§2.3); (4) finally, we tailor the translated templates by automatically substituting the placeholders with entities in the extracted ontologies to construct data for the three use cases (§2.4).

2.1 Automatic Template Creation

We start with MultiWoZ 2.2 (Zang et al., 2020) – a high-quality multi-domain English ToD dataset with more accurate human annotations compared to its predecessors MultiWoZ 2.0 (Budzianowski et al., 2018) and MultiWoz 2.1 (Eric et al., 2020). For the sake of reducing human efforts for collecting ToD context in the target languages, we re-use the ToD context written by human in MultiWoZ as the dialogue templates. Specifically as shown in Figure 2, we replace the English entities in MultiWoz by a set of general-purpose placeholders such as `[attraction-name0]` and `[attraction-postcode1]`, where each placeholder contains the entity's domain, attribute and ID. To do so, we first build a dictionary with entity-placeholder pairs by parsing the annotations of all dialogues. For example, from a dialogue text —*"I recommend Whale of a time and the post code is cb238el."*, we obtain two entity-placeholder pairs from its human annotations, i.e., (*Whale of a time*, `[attraction-name0]`) and (*cb238el*, `[attraction-postcode1]`). Next, we identify entities in the dialogue by their word index from the human annotations, replace them with their placeholders in the dictionary, and finally obtain dialogue templates with placeholders. Notably, we skip the entities with their attributes of `[choice]` and `[ref]` that represent the number of choices and booking reference number, as these attributes could be used globally.

2.2 Labeled Sequence Translation

Following Liu et al. (2021) that translates sentences with placeholders, we use a machine translation system[2] to translate dialogue templates with our designed placeholders. As we observe, a placeholder containing an entity domain, attribute and ID (e.g., `attraction-name0`) is useful to provide contextually meaningful information to the translation system, thus usually resulting in a high-quality translation with the placeholder unchanged [3]. This also enables us to easily locate the placeholders in the translation output and replace them with new entities in the target language.

To build a high-quality test set for evaluation, we further hire professional translators to post-edit a few hundred machine-translated templates, which produces natural and coherent sentences in the target languages. With the goal of selecting representative test templates for post-editing, we first calculate the frequency of all the 4-gram combinations in the MultiWoZ data, and then score each dialogue in the test set by the sum of the frequency of all the 4-gram combinations in the dialogue divided by the dialogue's word length. We use this scoring function to estimate the representiveness of a dialogue in the original dataset. Finally, we select the top 500 high-scoring dialogues in the test set for post-editing.[4] We also use the same procedure to create a small high-quality training set for few-shot cross-lingual transfer setting.

2.3 Collection of Local Ontology

Meanwhile, we crawl the attribute information of local entities in three cities from public websites (e.g., tripadvisor.com, booking.com) to create three ontologies for the three corresponding target languages respectively. As shown in Table 7 in Appendix D, we select Barcelona for Spanish (an Indo-European language), Shanghai for Mandarin (a Sino-Tibetan language) and Jakarta for Indonesian (an Austronesian language), which cover a set of typologically different language families.

Given a translated dialogue template, we can easily sample a random set of entities for a domain of interest from a crawled ontology and assign the entities to the template's placeholders to obtain a new dialogue in the target language. Repeating this procedure on each dialogue template, we can easily build a high-quality labeled dataset in the target language. Table 8 in Appendix E shows the statistics of our collected entities in the target languages compared with the English data. The number of our collected entities are either larger than or equal to those in the English data except for the "train" domain; we collected the information about only 100 "trains" for each languages due to the complexity in collecting relevant information.

2.4 Template Filling for Three Use Cases

After the above steps, we assign entities in a target language to the translated templates in the same

[2] We use Google Translate (`https://cloud.google.com/translate`), an off-the-shelf MT system.

[3] Appendix B has an example of label sequence translation.

[4] Appendix C shows the English test data distribution.

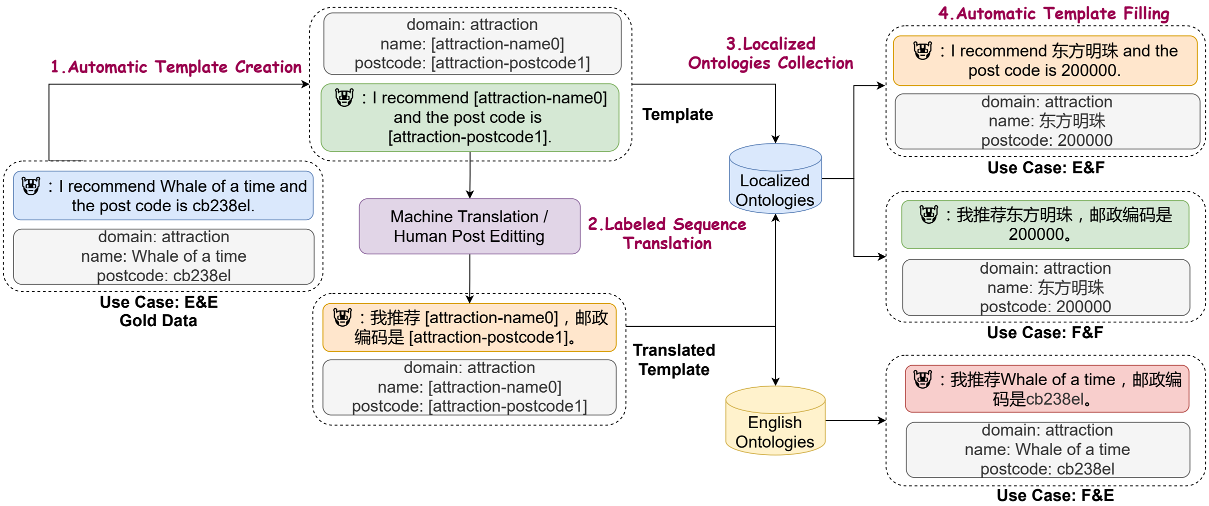

Figure 2: Illustration of our proposed pipeline.

target language for the F&F case, while assigning target-language entities to the English (source-language) templates for the F&E case. As for the E&F case, we keep the original English context by skipping the translation step and replace the placeholders with local entities in the target language (see Figure 2 for examples).

To sum up, our proposed method has three key properties: (1) our method is *cost-effective* as we only require a limited amount of post-editing efforts for a test set when compared to the expensive crowd-sourced efforts from the other studies; (2) we can easily sample entities from an ontology to create *large-scale machine-translated data* as a way of data augmentation for training; (3) our method is *flexible* to update entities in a ToD system whenever an update of ontology is available, e.g., extension of new entities. We refer the readers to Table 9 for the data statistics of GlobalWoZ and Figure 8 for dialogue examples in the appendix.

3 Task & Settings

3.1 Dialogue State Tracking

Our experiments focus on the dialogue state tracking (DST), one of the fundamental components in a ToD system that predicts the goals of a user query in multi-turn conversations. We follow the setup in MultiWoZ (Budzianowski et al., 2018) to evaluate ToD systems for DST by the joint goal accuracy which measures the percentage of correctly predicting all goals in a multi-turn conversation.

3.2 Experimental Settings

Zero-Shot Cross-lingual Transfer: Unlike prior studies that annotate a full set of high-quality training data for a target language, we investigate the *zero-shot cross-lingual transfer* setting where we have access to only a high-quality human-annotated English ToD data (referred to as gold standard data hereafter). In addition, we assume that we have access to a machine translation system that translates from English to the target language. We investigate this setting to evaluate how a multilingual ToD system transfers knowledge from a high-resource source language to a low-resource target language.

Few-Shot Cross-lingual Transfer: We also investigate few-shot cross-lingual transfer, a more practical setting where we are given a small budget to annotate ToD data for training. Specifically, we include a small set (100 dialogues) of high-quality training data post-edited by professional translators (§2.2) in a target language, and evaluate the efficiency of a multilingual ToD on learning from a few target-language training examples.

4 Proposed Baselines

We prepare a base model for GlobalWoZ in the zero-shot and few-shot cross-lingual transfer settings. We select Transformer-DST (Zeng and Nie, 2020) as our base model as it is one of the state-of-the-art models on both MultiWoZ 2.0 and MultiWoZ 2.1[5]. In our paper, we replace its BERT encoder with an mBERT encoder (Devlin et al., 2019) for our base model and propose a series of training methods for GlobalWoZ. As detailed below, we propose several data augmentation baselines that create different training and validation data for

[5]According to the leaderboards of Multi-domain Dialogue State Tracking on MultiWoZ 2.0 and MultiWoZ 2.1 on paperwithcode.com as of 11/15/2021.

training a base model. Note that all the proposed baselines are model agnostic and the base model can be easily substituted with other popular models (Heck et al., 2020; Lin et al., 2020). For each baseline, we first train a base model on its training data for 20 epochs and use its validation set to select the best model during training. Finally we evaluate the best model of each baseline on the same test set from GlobalWoZ. We will release GlobalWoZ and our pre-trained models to encourage faster adaptation to future research. We refer the readers to Table 10 and Table 11 in Appendix H while reading the subsequent methods for a better understanding.

4.1 Pure Zero-Shot (E&E)

We train a base model on the gold standard English data (E&E) and directly apply the learned model to the test data of the three use cases in GlobalWoZ. With this method, we simulate the condition of having labeled data only in the source language for training, and evaluate how the model transfers knowledge from English to the three use cases. We use **Zero-Shot (E&E)** to denote this method.

4.2 Translate-Train

We use our data curation method (§2) to translate the templates by an MT system but replace the placeholders in the translated templates with machine-translated entities to create a set of pseudo-labeled training data. Next, we train a base model on the translated training data without local entities, and evaluate the model on the three use cases. We denote this method as **Translate-Train**.

4.3 Single-Use-Case Training

By skipping the human post-editing step in our data curation method (§2), we leverage a machine translation system to automatically create a large set of pseudo-labeled training data with local entities for the three use cases. In the F&F case, we translate the English templates by the MT system and replace the placeholders in the translated templates with foreign-language entities to create a training dataset. In the F&E case, we replace the placeholders in the translated templates with the original English entities to create a code-switched training dataset. In the E&F case, we use the original English templates and replace the placeholders in the English templates with foreign-language entities to create a code-switch training dataset. With this data augmentation method, we can train a base model on each pseudo-labeled training dataset created for each use case. We denote this method as **SUC** (Single-Use-Case).

4.4 Bi-/Multi-lingual Bi-Use-Case Training

We investigate the performance of combining the existing English data and the pseudo-labeled training data created for one of the three use cases (i.e., F&F, F&E, E&F), one at a time, to do bi-use-case training. In the bilingual training, we only combine the gold English data (E&E) with the pseudo-labeled training data in one target language in one use case for joint training. We denote this method as **BBUC** (Bilingual Bi-Use-Case). In the multilingual training, we combine gold English data (E&E) and pseudo-labeled training data in all languages in one use case for joint training. We denote this method as **MBUC** (Multilingual Bi-Use-Case).

4.5 Multilingual Multi-Use-Case Training

We also propose to combine the existing English data (E&E) and all the pseudo-labeled training data in all target languages for all the use cases (F&F, F&E, E&F). We then train a single model on this combined multilingual training dataset and evaluate the model on test data in all target languages for all three use cases . We denote this method as **MMUC** (Multilingual Multi-Use-Case).

5 Experiment Results

In this section, we show the results of all methods in the zero-shot (§5.1) and few-shot (§5.2) settings.

5.1 Zero-shot Cross-lingual Transfer

5.1.1 Use Case F&F, F&E and E&F

Table 1 reports the joint goal accuracy of all proposed methods on the three different sets of test data in the F&F, F&E, and E&F use cases[6]. Both Zero-Shot (E&E) and Translate-Train struggle, achieving average accuracy of less than 10 in all use cases. Despite its poor performance, Zero-Shot (E&E) works much better in F&E than F&F, while its results in F&F and E&F are comparable, indicating that a zero-shot model trained in E&E can transfer knowledge about local English entities more effectively than knowledge about English context in downstream use cases. Besides, we also find that Zero-Shot (E&E) performs better on the Spanish or Indonesian context than the Chinese

[6] Appendix I reports the results in the E&E use case.

Case	Methods	zh	es	id	avg
F&F	Zero-Shot (E&E)	1.22	1.38	1.26	1.28
	Translate-Train	2.61	2.59	5.74	3.65
	SUC (F&F)	36.97	24.66	25.26	28.96
	BBUC (E&E + F&F)	37.32	25.52	26.39	29.74
	MBUC (E&E + F&F)	**38.01**	**26.03**	**28.22**	**30.76**
F&E	Zero-Shot (E&E)	6.92	11.34	9.09	9.12
	Translate-Train	2.28	4.97	4.67	3.97
	SUC (F&E)	56.28	41.94	47.93	48.71
	BBUC (E&E + F&E)	59.87	48.20	54.79	54.29
	MBUC (E&E + F&E)	**60.37**	**53.56**	**54.93**	**56.28**
E&F	Zero-Shot (E&E)	1.69	1.81	1.82	1.77
	Translate-Train	1.39	1.76	1.86	1.67
	SUC (E&F)	38.56	28.00	43.82	36.79
	BBUC (E&E + E&F)	39.87	27.29	45.48	37.54
	MBUC (E&E + E&F)	**40.20**	**29.22**	**47.06**	**38.83**

Table 1: Zero-shot cross-lingual accuracy on DST over three target languages in three use cases.

context in F&E. One possible reason is that English is closer to the other Latin-script languages (Spanish and Indonesian) than Chinese.

Our proposed data augmentation methods (SUC, BBUC, MBUC) perform much better than non-adapted methods (Zero-Shot (E&E) and Translate-Train) that do not leverage any local entities for training. In particular, it is worth noting that even though Translate-Train and SUC both do training on foreign-language entities in F&F and E&F, there is a huge gap between these two methods, since Translate-Train has only access to the machine-translated entities rather than the real local entities used by SUC. This huge performance gaps not only show that Translate-Train is not an effective method in practical use cases but also prove that having access to local entities is a key to building a multilingual ToD system for practical usage.

Comparing our data augmentation methods SUC and BBUC, we find that the base model can benefit from training on additional English data (E&E), especially yielding a clear improvement of up to 5.58 average accuracy points in F&E. Moreover, when we increase the number of languages in the bi-use-case data augmentations (i.e., MBUC), we observe an improvement of around 1 average accuracy points in all three use cases w.r.t. BBUC. These observations encourage a potential future direction that explores better data augmentation methods to create high-quality pseudo-training data.

5.1.2 One Model for All

Notice that we can train a single model by MMUC for all use cases rather than training separate models, one for each use case. In Figure 3, we compare

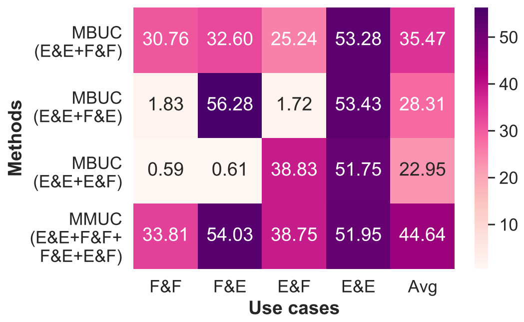

Figure 3: Performance of MMUC vs MBUC on the test data of the four use cases, F&F, F&E, E&F and E&E.

MMUC and MBUC (rows) on the test data in the four use cases (columns). Although MMUC may not achieve the best results in each use case, it achieves the best average result over the four use cases, indicating the potential of using one model to simultaneously handle all the four use cases.

5.2 Few-shot Cross-lingual Transfer

In few-shot experiments, we use the same scoring function based on frequency of all 4-gram combinations (§2.2) to select 100 additional dialogues from train set for human-post editing, and create high-quality training data for each of the three use cases. To avoid overfitting on this small few-shot dataset, we combine the few-shot data with the existing English data for training a base model (Few-Shot+Zero-Shot (E&E)). Next, we also investigate a model trained with additional synthetic data created by our proposed SUC. In Figure 4, we find that our proposed SUC without additional few-shot data has already outperformed the model trained with few-shot data and English data (Few-shot + Zero-Shot (E&E)), indicating that the model benefit more from a large amount of pseudo-labeled data than a small set of human-labeled data. If we combine the data created by SUC with the few-shot data or with both few-shot and English data to train the model, we observe improvements over SUC, especially with a clear gain of 8.06 accuracy points in F&E. We refer the readers to Table 13 in the appendix for detailed scores in all target languages.

6 Discussion

6.1 Motivation for Code-Switched Use Cases

One key research question is to validate whether code-switched use cases with local entities (i.e., F&E, E&F) are practically more useful for information seeking. To answer this question, we compare the failure rate of using local entities and machine-

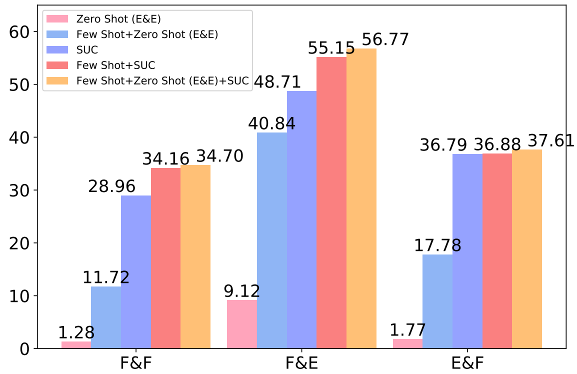

Figure 4: Few-shot cross-lingual average joint accuracy on DST over three target languages in three use cases.

Translate	Search	En→Zh	En→Es	En→Id	Zh→En	Es→En	Id→En
✓	✓	35	42	36	62	30	31
✓	✗	61	34	51	18	18	15
✗	✓	0	24	13	11	50	54
✗	✗	4	0	0	8	2	0
Failure Case (MTed Entities)		65	58	64	37	70	69
Failure Rate (MTed Entities)		65%	58%	64%	37%	70%	69%
Failure Rate (Original Entities)		3%	3%	3%	0%	1%	0%

Table 2: The search and translation results of 100 translated entities on Google. En→Zh refers to the translation of English entities to Mandarin and Zh→En refers to the translation of Mandarin entities to English.

translated entities in information search, which is a proxy to the efficiency of using these two types of entities in conversations. We first randomly select 100 entities (33 attractions, 33 hotels and 34 restaurants) of Cambridge, Shanghai, Barcelona and Jakarta. We translate the English entities into Mandarin, Spanish and Indonesian and the foreign-language entities into English via Google Translate. We then manually search the translated entities on Google to check whether we can find the right information of the original entities. Notice that the failure of the above verification partially come from the translation error made by Google Translate, or the search failure due to the fact that this entity does not have a bilingual version at all. In Table 2, we observe a high failure rate of around 60% for almost all translated directions (except Zh→En) due to translation and search failures, significantly exceeding the low failure rate of searching original entities online. Besides, even if we can find the right information of the translated entities, local people may not recognize or use the translated entities for communication, thus this results in inefficient communication with local people.

6.2 Overestimate of Translate-Train

In previous translation-based work, a multilingual ToD system is usually built based on the translation

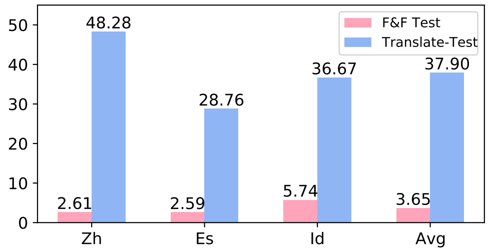

Figure 5: Joint accuracy of Translate-Train for DST on the F&F Test vs Translate-Test data.

Train Set	E&E (en)	F&F (zh)	F&F (es)	F&F (id)	avg
Local Context Only	5.46	1.77	2.37	2.40	3.20
Local Entities Only	6.39	0.36	2.41	2.75	3.05
Local Context & Entities	**52.78**	**36.97**	**24.66**	**25.26**	**38.13**

Table 3: Comparison of training with local context or/and local entities on the joint accuracy for DST in E&E (en) and F&F (zh, es, id).

of English training data (Translate-Train), and is evaluated on translated test data without any local entities (Translate-Test). To verify whether this procedure is reliable to build a multilingual ToD system, we also create a test dataset with translated entities instead of local entities in the target languages. As shown in Figure 5, we find the Translate-Train model performs well on the test data with translated entities, but performs badly on the test data with real local entities. To the best of our knowledge, we provide the first analysis to identify this performance gap between the translated test data and data with real local entities in a more realistic use case. Our work sheds light on the development of a globalized multilingual ToD system in practical use cases.

6.3 Local Context vs. Local Entities

We compare the impact of training a model on data with either local contexts or local entities when the model is evaluated on monolingual test data in F&F and E&E. Specifically, when the train set has access to local context only, all the entities in the train set are replaced by entities in non-target languages. Similarly, when the train set has access to local entities only, the contexts in the train set are replaced by context in the non-target languages. Table 3 shows that both local contexts and local entities are essential to building ToD systems in the target language. A further analysis in Table 14 and Table 15 in the appendix shows that training with local entities is more important if the entities and contexts are written in the same type of language script (e.g. Latin script).

Method	zh	es	id	ar	da	de	el	fr	he	it	ja	ko	nl	no	pt	ru	sv	th	tr	vi	avg
F&F	1.22	1.38	1.26	1.49	1.52	1.52	1.51	2.04	1.47	1.55	1.48	1.51	1.55	1.51	1.53	1.52	1.41	1.57	1.22	1.41	1.48
F&E	6.92	11.34	9.09	6.80	10.97	10.15	6.74	15.87	7.81	9.40	3.17	4.92	11.79	11.46	10.12	8.97	10.31	10.89	5.98	7.92	9.03
E&F	1.69	1.81	1.82	1.94	1.98	1.96	2.01	2.82	1.99	1.98	1.92	1.92	1.94	1.97	1.95	1.99	1.89	1.86	2.00	1.99	1.97

Table 4: Results of Zero-Shot (E&E) on test data of F&F, F&E and E&F in 20 languages. Test data of F&F and F&E in the three languages highlight in pink color are built with MTPE data and the rest are built with MT data.

Use Case	F2F		F2E	
Methods	**MT Test**	**MTPE Test**	**MT Test**	**MTPE Test**
Zero-Shot (E&E)	1.29	1.28	9.64	9.12
Translate-Train	3.71	3.65	4.17	3.97
SUC	35.78	28.96	56.15	48.71
BBUC	36.31	29.74	57.84	54.29
MBUC	**37.89**	**30.76**	**58.76**	**56.28**
Spearman's correlation	**1.0**		**1.0**	

Table 5: Comparison of average joint accuracy on DST reported on MT test data and MTPE test data for use case F&F and F&E

6.4 Scaling up to 20 Languages

With our proposed data curation method, it is possible to extend the dataset to cover more languages without spending extra costs if we skip the human post-editing step. Before doing so, one key question is whether the evaluation on the translated data without human post-editing is reliable as a proxy of the model performance. Thus, we conduct the experiments by evaluating the model performance of all baselines (§4) on two sets of test data built with local entities: (1) **MT** test data where translated template is created by machine translation only (§2.2); (2) **MTPE** test data where translated template is first translated by machines and post-edited later by professional translators. As shown in Table 5, the overall reported results on MT test data are higher than those reported on MTPE test data, which is expected because the distribution of the MT test data is more similar to the MT training data. Although there are some differences on individual languages, the conclusions derived from the evaluations on the MT test data remain the same as those derived from the evaluation on the MTPE test data. We also calculate the Spearman rank correlation coefficient between the average results reported on MTPE test data and MT test data in Table 5, which shows a statistically high correlation between the system performance on the MT test data and MTPE test data[7]. Therefore, we show that the MT test data can be used as a proxy to estimate the model performance on the real test data for more languages. Thus we build MT test data for another 17 languages that are supported by Google Translate, Trip Advisor and Booking.com at the same time, as stated in Table 7 and Table 8 in the appendix. Table 4 shows the results of Zero-Shot (E&E) on the test data of F&F, F&E and E&F in 20 languages.

[7]Table 16 in the appendix shows detailed scores.

7 Related Work

Over the last few years, the success of ToD systems is largely driven by the joint advent of neural network models (Eric et al., 2017; Wu et al., 2019; Lin et al., 2020) and collections of large-scale annotation corpora. These corpora cover a wide range of topics from a single domain (e.g., ATIS (Hemphill et al., 1990), DSTC 2 (Henderson et al., 2014), Frames (El Asri et al., 2017), KVRET (Eric et al., 2017), WoZ 2.0 (Wen et al., 2017), M2M (Schatzmann et al., 2007)) to multiple domains (e.g., MultiWoZ (Budzianowski et al., 2018), SGD (Rastogi et al., 2020)). Most notably among these collections, MultiWoZ is a large-scale multi-domain dataset that focuses on transitions between different domains or scenarios in real conversations (Budzianowski et al., 2018). Due to the high cost of collecting task-oriented dialogues, only a few monolingual or bilingual non-English ToD datasets are available (Zhu et al., 2020; Quan et al., 2020; Lin et al., 2021). While there is an increasing interest in data curation for multilingual ToD systems, a vast majority of existing multilingual ToD datasets do not consider the real use cases when using a ToD system to search for local entities in a country. We fill this gap in this paper to provide the first analysis on three previously unexplored use cases.

8 Conclusions

In this paper, we provide an analysis on three unexplored use cases for multilingual task-oriented dialogue systems. We propose a new data curation method that leverages a machine translation system and local entities in target languages to create a new multilingual TOD dataset, GlobalWoZ. We propose a series of strong baseline methods and conduct extensive experiments on GlobalWoZ to encourage research for multilingual ToD systems. Besides, we extend the coverage of languages on multilingual ToD to 20 languages, marking the one step further towards building a globalized multilingual ToD system for all of the world's citizen.

References

Dan Bohus and Alexander I. Rudnicky. 2009. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23:332–361.

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 (EMNLP)*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Li-Rong Cheng and Katharine Butler. 1989. Code-switching: a natural phenomenon vs language 'deficiency'. *World Englishes*, 8(3):293–309.

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 (ACL)*, 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 (NAACL), Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 422–428, Marseille, France. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Narendra Kumar Gupta, Gökhan Tür, Dilek Z. Hakkani-Tür, Srinivas Bangalore, Giuseppe Riccardi, and Mazin Gilbert. 2006. The at&t spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:213–222.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Bernd Kiefer, Anna Welker, and Christophe Biwer. 2021. Vonda: A framework for ontology-based dialogue management. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 93–105. Springer Singapore.

Eunhee Kim. 2006. Reasons and motivations for code-mixing and code-switching. *Issues in EFL*, 4(1):43–61.

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 (EACL)*, pages 2950–2962, Online. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv preprint arXiv:2106.02787*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Volume 1: Long Papers*, pages 5834–5846, Online. 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 (TACL)*, 8:726–742.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940, 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 (AAAI)*, volume 34, pages 8689–8696.

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Anna Korhonen, and Ivan Vulić. 2021. Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. *arXiv preprint arXiv:2104.08570*.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL); Companion Volume, Short Papers*, pages 149–152.

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 (NAACL), Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 2479–2497, Online. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 808–819, Florence, Italy. 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 (NAACL)*, pages 483–498, Online. 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.

Yan Zeng and Jian-Yun Nie. 2020. Jointly optimizing state operation prediction and value generation for dialogue state tracking. *arXiv preprint arXiv:2010.14061*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, pages 1–17.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics (TACL)*, 8:281–295.

Appendix

A Comparison of Four Use Cases

Use Case	Source ToD	Speaker (ToD Context)	Country (ToD Ontology)
F&F	English	Foreign Lang.	Foreign Lang.
F&E		Foregin Lang.	English
E&F		English	Foreign Lang.
E&E		English	English

Table 6: Four use cases of multilingual ToD systems: A foreign language or English speaker travels to a country of a foreign language or English.

B Examples of Labeled Sequence Translation

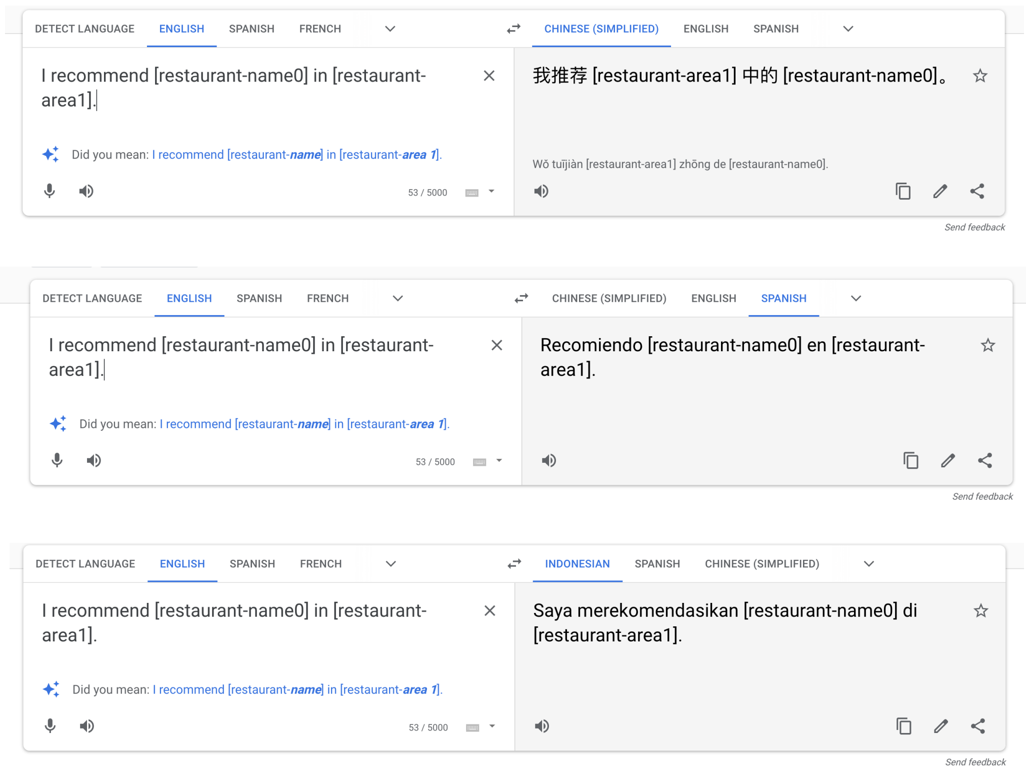

Figure 6: An instance of labeled sequence translation with google translate, from English to three target languages, Mandarin, Spanish and Indonesian.

C Test Set Distribution

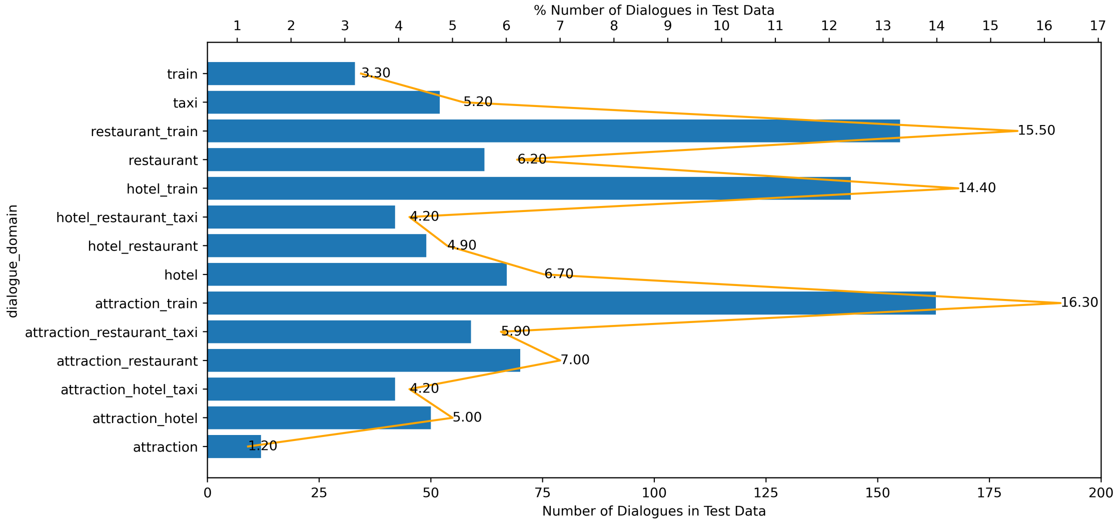

Figure 7: Gold English Test Set Distribution by Domains. We follow this distribution to select the top 500 high-scoring dialogues in the test set for post-editing.

D Selected Languages

Language	ISO639-1code	Language Family	# Wikipedia articles (in millions)	High / Middle/ Low Resource	Writing Script	Selected City
English	en	IE: Germanic	6.35	High	Latin	Cambridge
Swedish	sv	IE: Germanic	2.95	High	Latin	Stockholm
German	de	IE: Germanic	2.61	High	Latin	Berlin
French	fr	IE: Romance	2.35	High	Latin	Paris
Dutch	nl	IE: Germanic	2.06	High	Latin	Amsterdam
Russian	ru	IE: Slavic	1.74	High	Cyrillic	Moscow
Italian	it	IE: Romance	1.71	High	Latin	Rome
Spanish	es	IE: Romance	1.71	High	Latin	Barcelona
Japanese	ja	Japonic	1.28	High	Ideograms	Tokyo
Vietnamese	vi	Austro-Asiatic	1.27	High	Latin	Ho Chi Minh City
Mandarin	zh	Sino-Tibetan	1.22	High	Chinese ideograms	Shanghai
Arabic	ar	Afro-Asiatic	1.13	High	Arabic	Cairo
Portuguese	pt	IE: Romance	1.07	High	Latin	Lisbon
Indonesian	id	Austronesian	0.59	Middle	Latin	Jakarta
Norwegian	no	IE: Germanic	0.56	Middle	Latin	Oslo
Korean	ko	Koreanic	0.55	Middle	Hangul	Seoul
Turkish	tr	Turkic	0.42	Middle	Latin	İstanbul
Hebrew	he	Afro-Asiatic	0.30	Low	Hebrew	Tel Aviv
Danish	da	IE: Germanic	0.27	Low	Latin	Copenhagen
Greek	el	IE: Greek	0.20	Low	Greek	Athens
Thai	th	Kra-Dai	0.14	Low	Brahmic	Bangkok

Table 7: Statistics about languages in the cross-lingual benchmark. The selected 21 languages (including English) belong to 8 language families and 1 isolate, with Indo-European (IE) having the most members. We categorize the languages with more than 1 million, more than 400 thousand but less than 1 million, less than 400 thousand Wikipedia articles as high resource languages, middle resource languages and low resource languages. For each language, we select one city for each language to collect localized ontology.

E Statistics of Entities in the Collected Ontology

Languages	rest.	hotel	attr.	train	taxi
en	110	33	79	2828	222
zh	3000	496	1000	100	4496
es	3000	426	1000	100	4426
id	3000	999	792	100	4791
ar	2989	680	1000	100	4669
da	2343	165	1000	100	3508
de	2988	659	1000	100	4647
el	2600	1000	1000	100	4600
fr	3000	1000	1000	100	5000
he	1558	258	1000	100	2258
it	3000	800	1000	100	2800
ja	2967	864	1000	100	4831
ko	2990	532	1000	100	4522
nl	2990	537	1000	100	4527
no	1293	95	757	100	2145
pt	2993	951	1000	100	4944
ru	2985	531	1000	100	4516
sv	3000	214	891	100	4105
th	2995	1000	1000	100	4995
tr	2986	533	1000	100	4519
vi	2991	773	1000	100	4764

Table 8: Statistics of entities in the collected ontology in different languages. We count the number of entities in the database of each domain. Noticed that in the Taxi database of MultiWoZ, it only list down the taxi colors, taxi types and taxi phones. The taxi destination and departure refer to the entities in the restaurant, hotel and attraction domains. Thus, we use the sum of the number of entities in Restaurant, Hotel and Attraction domains as a proxy of the total number of entities in taxi domain. Besides, we follow MultiWoZ to collect one hospital and one police station for each city.

F Statistics of GlobalWoZ

Use Case	F&F				F&E				E&F			
Languages	Train & Dev	Method	Test	Method	Train & Dev	Method	Test	Method	Train & Dev	Method	Test	Method
zh	9438	MT	1000	MTPE	9438	MT	1000	MTPE	9438	Human	1000	Human
es	9438	MT	1000	MTPE	9438	MT	1000	MTPE	9438	Human	1000	Human
id	9438	MT	1000	MTPE	9438	MT	1000	MTPE	9438	Human	1000	Human
ar	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
da	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
de	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
el	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
fr	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
he	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
it	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
ja	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
ko	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
nl	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
no	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
pt	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
ru	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
sv	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
th	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
tr	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human
vi	9438	MT	1000	MT	9438	MT	1000	MT	9438	Human	1000	Human

Table 9: Statistics of created dataset, GlobalWoZ for each use case in each target language. For E&F, as the context is the original Engish data, we consider it is created by human. For test data of zh, es and id, we replace the entities twice to boostrap the test data to 1000 dialogues. We are currently preparing the post editing of the other 500 dialogues in test data. Meanwhile, we are leveraging machine translation to prepare the train data for the 17 languages and will release it with baselines in the next version soon.

G Dialogue Examples

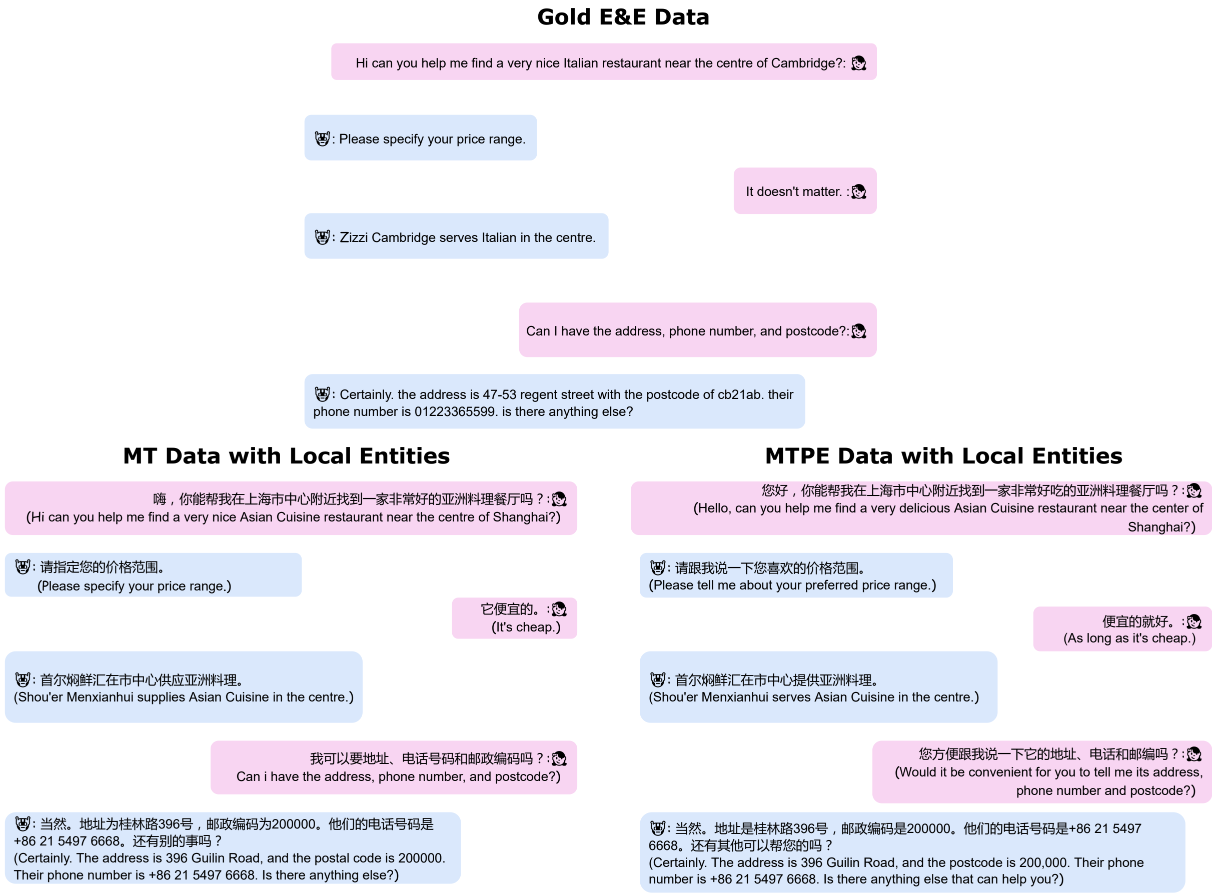

Figure 8: Examples of some utterances in original E&E data, MT data and MTPE data,

H Summary of Proposed Baselines

Methods	En Context	En Entities	Local Context	Local Entities	Translated Entites
Zero-Shot (E&E)	✔	✔			
Translate-Train			✔		✔
SUC (F&F)			✔	✔	
SUC (F&E)		✔	✔		
SUC (E&F)	✔			✔	

Table 10: Accessibility of different types of context and entities for each method.

Methods	E&E	F&F	F&E	E&F
Zero-Shot (E&E)	✔			
Translate-Train				
SUC (F&F)		✔		
SUC (F&E)			✔	
SUC (E&F)				✔
BBUC (E&E + F&F)	✔	✔		
BBUC (E&E + F&E)	✔		✔	
BBUC (E&E + E&F)	✔			✔
MBUC (E&E + F&F)	✔	✔		
MBUC (E&E + F&E)	✔		✔	
MBUC (E&E + E&F)	✔			✔
MMUC (E&E + F&F + F&E + E&F)	✔	✔	✔	✔

Table 11: Accessibility of data in each use case for each method. Noticed that Translate-Train doesn't have access to the data of the four use cases. Translate-Train has access to a set of pseudo-labeled training data created by replacing the placeholders in the translated template with machine-translated entities instead of local entities.

I Use Case E&E

We also compare the performance of all methods on the original E&E test data. As **Zero-Shot (E&E)** is trained on monolingual English training data, it gets a high accuracy of 52.78 on the English test data. In contrast, **Translate-Train** and **SUC** (F&F) perform poorly on the English test data, because both of them have no access to any English data. Comparing to **SUC** (F&F), **SUC** (F&E) and **SUC** (E&F) achieve higher accuracy scores as they either have access to English context or English entities. When we perform bilingual and multilingual joint training (i.e., **BBUC** and **MBUC**), the base model has a performance increase except **MBUC** (E&E + E&F). This shows that bilingual and multilingual joint training may be used to improve the performance on source language. Further research can be done in this line.

Methods	En
Zero-Shot (E&E)	52.78
Translate-Train	2.27
SUC (F&F)	1.09
SUC (F&E)	6.39
SUC (E&F)	5.46
BBUC (E&E + F&F)	52.87
BBUC (E&E + F&E)	**53.69**
BBUC (E&E + E&F)	53.05
MBUC (E&E + F&F)	53.28
MBUC (E&E + F&E)	53.43
MBUC (E&E + E&F)	51.75

Table 12: Joint accuracy on DST in three target languages on the English test data.

J Breakdown of Few Shot Results

Zero Shot (E&E)				
Use Case	Zh	Es	Id	Avg
F2F	1.22	1.38	1.26	1.28
F2E	6.92	11.34	9.09	9.12
E2F	1.69	1.81	1.82	1.77
Few Shot + Zero Shot (E&E)				
Use Case	Zh	Es	Id	Avg
F2F	15.93	7.13	12.09	11.72
F2E	39.88	39.38	43.26	40.84
E2F	20.61	14.17	18.55	17.78
SUC				
Use Case	Zh	Es	Id	Avg
F2F	36.97	24.66	25.26	28.96
F2E	56.28	41.94	47.93	48.71
E2F	38.56	28.00	43.82	36.79
Few Shot + SUC				
Use Case	Zh	Es	Id	Avg
F2F	37.81	25.15	39.51	34.16
F2E	58.39	53.03	54.02	55.15
E2F	38.75	27.66	44.23	36.88
Few Shot + Zero Shot (E&E) + SUC				
Use Case	Zh	Es	Id	Avg
F2F	37.52	26.44	40.15	34.70
F2E	59.21	54.93	56.17	56.77
E2F	39.51	27.84	45.48	37.61

Table 13: A breakdown of few-shot cross-lingual average joint accuracy on DST over three target languages in three use cases.

K Breakdown of the Results of Local Context vs Local Entities by Languages

E&E (en)				
Context vs Entities	Zh	Es	Id	Avg
En_Context	5.37	5.33	5.67	5.46
En_Entites	3.49	7.78	7.90	6.39
F&F (zh)				
Context vs Entities	En	Es	Id	Avg
Zh_Context	1.74	1.77	1.80	1.77
Zh_Entites	0.27	0.73	0.10	0.36
F&F (es)				
Context vs Entities	En	Zh	Id	Avg
Es_Context	1.73	2.01	3.37	2.37
Es_Entites	3.92	0.44	2.86	2.41
F&F (id)				
Context vs Entities	En	Zh	Es	Avg
Id_Context	2.07	2.18	2.94	2.40
Id_Entites	3.92	0.84	3.48	2.75

Table 14: A breakdown of comparison of the impact of local context and local entities on joint accuracy for DST in each language. The cases where context and entities are in different script types are highlighted in lavender color.

Train Set	**different script type**	**same script type**
Local Context Only	**2.48**	3.52
Local Entities Only	0.98	**4.98**

Table 15: Comparison of the impact of script type on Local Context Only vs Local Entities Only. It shows that training with local entities is more important if the entities and contexts are written in the same type of language script (e.g. Latin script), otherwise training with local contexts is more important.

L Breakdown of MT Test Data vs MTPE Test Data by Languages

Languages	Zh		Es		Id	
F2F	MT	MTPE	MT	MTPE	MT	MTPE
Zero-Shot (E&E)	1.19	1.22	1.40	1.38	1.28	1.26
Translate-Train	2.50	2.61	2.81	2.59	5.81	5.74
SUC	37.79	36.97	26.95	24.66	42.59	25.26
BBUC	38.62	37.32	27.34	25.52	42.96	26.39
MBUC	39.11	38.01	29.17	26.03	45.39	28.22
Spearman's correlation	1.00		1.00		1.00	
F2E	MT	MTPE	MT	MTPE	MT	MTPE
Zero-Shot (E&E)	7.61	6.92	11.67	11.34	9.64	9.09
Translate-Train	2.25	2.28	5.25	4.97	5.03	4.67
SUC	57.10	56.28	55.70	41.94	55.64	47.93
BBUC	59.05	59.87	57.68	48.20	56.80	54.79
MBUC	60.48	60.37	57.04	53.56	58.23	54.93
Spearman's correlation	1.00		0.90		1.00	

Table 16: Spearman rank correlation coefficient between the results on MTPE test data and MT test data for each language.