AllWOZ: Towards Multilingual Task-Oriented Dialog Systems for All

Anonymous ACL submission

Abstract

A commonly observed problem of the state-of-the-art natural language technologies, such as Amazon Alexa and Apple Siri, is that their services do not extend to most developing countries' citizens due to language barriers. Such populations suffer due to the lack of available resources in their languages to build NLP products. This paper presents AllWOZ, a multilingual multi-domain task-oriented customer service dialog dataset covering eight languages: English, Mandarin, Korean, Vietnamese, Hindi, French, Portuguese, and Thai. Furthermore, we create a benchmark for our multilingual dataset by applying mT5 (Xue et al., 2021) in a meta-learning setting (Finn et al., 2017).

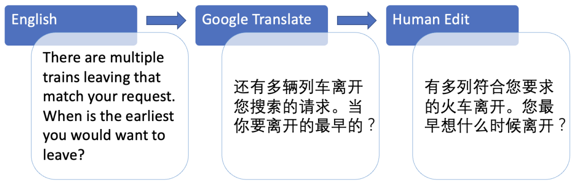

Figure 1: Three stages of our data collection: Data selection, machine translation and human correction.

1 Introduction

Task-oriented dialog systems are crucial for business solutions. While task-oriented dialog systems have made tremendous success in English, there is still a pressing urgency to build systems that can serve 6,900 other languages all over the world to enable universal technology access (Ruder et al., 2019; Aharoni et al., 2019; Arivazhagan et al., 2019). 94% of the world's population do not have English as their first language, and 75% do not speak English at all. Most developing countries' citizens cannot benefit from state-of-the-art language technologies due to language barriers.

Building dialog systems for most languages is challenging due to a lack of data. Automatic translation is a powerful tool to generate more resources. However, the state-of-the-art translations still suffer from low fluency and coherence. Moreover, they have difficulties dealing with the mentioned entities in the dialog, which is essential in serving the functional purposes of task-oriented dialog systems.

To facilitate the development of multilingual task-oriented dialog systems, we create a new dataset AllWOZ based on MultiWOZ (Budzianowski et al., 2018; Zang et al., 2020). AllWOZ is a multilingual multi-domain task-oriented dialog dataset with intent and state annotation. It has eight languages across various language families: English, Mandarin, Korean, Vietnamese, Hindi, French, Portuguese and Thai. We will extend the dataset to more than 20 languages in our future work.

Many languages have similarities in syntax and vocabulary, and multilingual learning approaches that leverage the shared structure of the input space have proven to be effective in alleviating data sparsity. In this work, we applied a meta-learning training schema for multilingual adaptation to take advantage of shared language structures.

Our contributions are as follows: (1) We collect a new dataset, AllWOZ, for multilingual task-oriented dialog systems. (2) Extensive experiments show that few-shot learning could improve the model performance on our dataset. We would make our dataset and models public.

2 Related Work

Early work in this direction focused on individual tasks, such as grammar induction (Ruder et al., 2019; Snyder et al., 2009), part-of-speech (POS) tagging (Täckström et al., 2013), parsing (McDonald et al., 2011), and text classification (Klementiev et al., 2012). General-purpose multilingual representation learning has gained increasing attention during the past few years. Approaches that are applicable to multiple tasks have been researched on both word-level (Mikolov et al., 2013; Faruqui and Dyer, 2014; Artetxe et al., 2017) and sentence-level (Devlin et al., 2019; Lample and Conneau, 2019). However, previous work processed text within a short context window due to the lack of datasets with long texts. There is little publicly available dialog resource that contains a diverse set of languages. A multilingual multi-domain natural language understanding (NLU) dataset with Thai, English, and Spanish (Schuster et al., 2019). (Mrksic et al., 2017) annotated only two additional languages in WOZ 2.0 and (Liu et al., 2019) proposed a mixed-language training method for cross-lingual NLU and dialog state tracking (DST) tasks.

In terms of algorithms, Schuster et al. (2019) found that in low resource settings, multilingual contextual word representations produce better results than using cross-lingual static embeddings. This suggests that simply using pre-trained multilingual embedding, such as MASS (Song et al., 2019) and mBART (Liu et al., 2020), which trained on auto-encoding objectives are not ideal for solving the dialog task. This prompts us to propose new algorithms that not only utilizes pretrained multilingual embedding, but also considers dialogue context information.

3 Multilingual Dialog Collection

To build a multilingual task-oriented dialog system, we collect a new dataset, AllWOZ, consisting of paired dialogs between different languages based on the MultiWOZ dataset. We first carefully sample dialogs from MultiWOZ and then translate those dialogs into different languages with Google Translation Tool. To ensure the quality of the translation, we recruit native speakers for each language to correct the translation results.

3.1 Data Selection

MultiWOZ (Budzianowski et al., 2018) is the most popular task-oriented dialog dataset, covering seven domains and containing 10K+ dialogs. Many works devote effort to correcting and improving the dataset (Eric et al., 2020; Qian et al., 2021; Han et al., 2021; Ye et al., 2021). We conduct translation jobs on the MultiWOZ 2.2 (Zang et al., 2020) since it is the most widely-accepted version. As mentioned in Sec. 1, most languages lack dialog training data, so our goal is to build dialog models under few-shot settings. Therefore, we sample 100

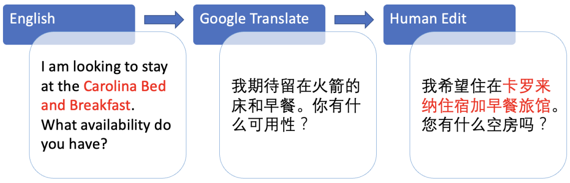

Figure 2: Data generated from machine translation are noisy when there are entities in the sentence.

dialogs (1476 turns) from the test set in total. In order to maintain the prior knowledge of each domain, we keep the same domain distribution of the whole test set during sampling. For example, as shown in Table 1, there are 38 out of 100 sampled dialogs involved in the attraction domain, and 399 out of 1000 dialogs involved in the attraction domain in the test set. Those two ratios are very close. Similarly, for all five domains, the ratio of the dialog number counted for sampled dialogs (left side of slash in Table 1) over the number for the whole test set (right side of slash) keeps consistent. The same case happens when it comes to the turn number. As for each domain, we expect the sampled dialogs to cover as much information as possible. So, during the sampling, we record the dialog state annotations of chosen dialogs and skip the dialog with similar annotations. As shown in the last row of Table 1, the sampled dialog covers all possible slot types.

3.2 Machine Translation

In order to reduce the human workload of translation, we first utilize the Google Translation Tool to automatically translate both dialog utterances and dialog state annotations. Fig. 1 shows an example of the translation flow. Machine-translated utterances are usually of low quality, mainly because some entity tokens like “Carolina Bed and Breakfast” is hard for machine to translate.

3.3 Human Correction

To build a high-quality dataset, we recruit native speakers to correct the errors in the machine-translated utterances. Our dataset currently covers eight languages: English, Mandarin, Korean, Vietnamese, Hindi, French, Portuguese and Thai. For each language except English, we recruit a bilingual speaker to edit the machine-translated utterances based on the original English dialogs. In addition to dialog utterances, we also require the translators to edit the machine translations of the

Num. of	Attraction	Hotel	Restaurant	Taxi	Train
Dialog	38/399	40/395	39/445	13/198	55/489
Turn	213/2433	254/2588	224/2867	38/640	331/2946
Slot Type	3/3	10/10	7/7	4/4	6/6

Table 1: The statistics for the sampled 100 dialogs/whole test set in terms of dialog number, turn number and the number of different slot types over all five domains in test set. The sampled dialogs are evenly distributed across the five domains.

dialog states (e.g. "Carolina Bed and Breakfast" in Fig. 1), because tasks like dialog state tracking and end-to-end dialog generation require those states. However, some entity tokens in dialog states have polysemy and the translation of the dialog states does not match the semantic meaning in the dialog utterance. For example, the token "moderate" refers to price by default in the dialog utterance. However, as an isolated token in dialog states, it is translated as "mild". To ensure that all the translations of dialog states are natural and coherent, we ask the translator to translate all dialog states first. Then, they should translate dialog utterances based on the dialog states' translations, in order to avoid inconsistency between utterances and slots. If any state looks not coherent or natural in an utterance, translators are required to edit the translation of dialog states and translate all related utterances again.

4 Experiments

In this section, we introduce how we divide the dataset for train, validation and evaluation, as well as the experiment setting.

4.1 Data Partition

For each language, the translated 100 dialogs are divided into two partitions, each with 40/10/50 dialogs. We first randomly sample a target language, then all the other seven languages are considered source languages. The experiments aim to explore whether the parallelism among source languages can help learn the target language under few-shot settings. For the source languages, we use the 40 dialogs as training data and 10 dialogs as validation data. And for the target language, since we focus on the few-shot learning, we utilize the partition of 10 dialogs as training data and 40 dialogs for validation. The remaining partition of 50 dialogs is used as the test set. In order to achieve trustful results, we run each experiment for two times. Each time we randomly re-sample the data partition and in the table we report the average score, along with the standard deviation.

4.2 Benchmark Models

Inspired by the success of pre-trained multilingual model (Song et al., 2019; Liu et al., 2020; Lin et al., 2020), we choose mT5 (Xue et al., 2020) as our backbone model, a multilingual pre-trained encoder-decoder language model. It is pre-trained on mC4 (Raffel et al., 2019), which covers 101 languages in total, including the 8 languages that we propose to translate. The experiment are conducted under the following settings:

- **Vanilla Training** The vanilla method is directly fine-tuning mT5 model with mixed dialogs from each source languages and then test on the target language.
- **Vanilla + English Pretrain** Inspired by the success of pre-trained language models, we first pre-train the model on the full-size MultiWOZ dataset (English), then conduct fine-tuning with the parallel dialogs of different source languages.
- **DAML** In order to explore the relation between the parallel dialogs, we adopt DAML (Qian and Yu, 2019) to train our model.

4.3 Metrics

Following (Budzianowski et al., 2018), we adopt Inform Rate, Success Rate, and BLEU (Papineni et al., 2002) score as our main evaluation metrics. **Inform Rate** represents the accuracy of successfully providing the correct entity (e.g., the name of a restaurant that satisfies all user's constraints in the restaurant domain). **Success Rate** measures how well the system answers all the requested information. **BLEU** score is adopted to evaluate the quality of the generated response, compared with the ground truth response. We also use **Slot Accuracy** to evaluate the quality of dialog state tracking.

Language	Model	BLEU	Inform	Success	Slot Accuracy
English	Vanilla	15.34$_{\pm 1.90}$	43.00$_{\pm 24.00}$	6.00$_{\pm 2.83}$	71.07$_{\pm 1.75}$
	+English Pretrain	20.39$_{\pm 3.92}$	57.00$_{\pm 12.73}$	11.00$_{\pm 12.73}$	75.27$_{\pm 8.19}$
	DAML	18.32$_{\pm 0.58}$	54.00$_{\pm 2.83}$	7.00$_{\pm 4.24}$	73.29$_{\pm 0.10}$
French	Vanilla	16.93$_{\pm 1.51}$	26.00$_{\pm 2.83}$	5.00$_{\pm 1.41}$	70.48$_{\pm 2.86}$
	+English Pretrain	19.94$_{\pm 3.12}$	23.00$_{\pm 1.41}$	4.00$_{\pm 0.00}$	73.74$_{\pm 8.63}$
	DAML	18.07$_{\pm 0.47}$	19.00$_{\pm 0.85}$	6.00$_{\pm 0.00}$	65.80$_{\pm 7.89}$
Vietnamese	Vanilla	16.95$_{\pm 1.09}$	23.00$_{\pm 1.41}$	6.00$_{\pm 2.83}$	68.46$_{\pm 1.72}$
	+English Pretrain	19.63$_{\pm 1.15}$	25.00$_{\pm 1.41}$	6.00$_{\pm 2.83}$	76.36$_{\pm 4.76}$
	DAML	18.89$_{\pm 1.74}$	23.00$_{\pm 1.41}$	5.00$_{\pm 1.41}$	70.61$_{\pm 8.36}$
Portuguese	Vanilla	13.35$_{\pm 4.51}$	23.00$_{\pm 1.41}$	5.00$_{\pm 1.41}$	75.13$_{\pm 5.16}$
	+English Pretrain	14.13$_{\pm 0.16}$	24.00$_{\pm 0.00}$	4.00$_{\pm 0.00}$	69.68$_{\pm 1.75}$
	DAML	16.66$_{\pm 1.94}$	24.00$_{\pm 0.00}$	6.00$_{\pm 2.83}$	72.82$_{\pm 2.68}$
Korean	Vanilla	7.29$_{\pm 0.09}$	24.00$_{\pm 0.00}$	4.00$_{\pm 0.00}$	69.93$_{\pm 1.42}$
	+English Pretrain	9.59$_{\pm 1.17}$	25.00$_{\pm 1.41}$	4.00$_{\pm 0.00}$	78.04$_{\pm 0.77}$
	DAML	8.15$_{\pm 2.15}$	23.09$_{\pm 1.41}$	4.00$_{\pm 0.00}$	74.42$_{\pm 1.92}$
Mandarin	Vanilla	3.42$_{\pm 1.56}$	26.00$_{\pm 2.83}$	5.00$_{\pm 1.41}$	74.53$_{\pm 4.98}$
	+English Pretrain	7.13$_{\pm 2.10}$	24.00$_{\pm 0.00}$	6.00$_{\pm 2.83}$	79.40$_{\pm 0.25}$
	DAML	3.93$_{\pm 0.29}$	24.00$_{\pm 0.00}$	4.00$_{\pm 0.00}$	69.83$_{\pm 1.81}$
Hindi	Vanilla	15.24$_{\pm 0.70}$	26.00$_{\pm 2.83}$	5.00$_{\pm 1.41}$	71.62$_{\pm 2.36}$
	+English Pretrain	16.37$_{\pm 0.26}$	23.00$_{\pm 1.41}$	4.00$_{\pm 0.00}$	75.79$_{\pm 1.56}$
	DAML	14.55$_{\pm 5.44}$	23.00$_{\pm 1.41}$	6.00$_{\pm 2.83}$	72.52$_{\pm 1.44}$
Thai	Vanilla	12.36$_{\pm 0.82}$	25.00$_{\pm 1.41}$	6.00$_{\pm 0.00}$	76.84$_{\pm 2.36}$
	+English Pretrain	10.76$_{\pm 1.43}$	22.00$_{\pm 2.83}$	4.00$_{\pm 0.00}$	65.17$_{\pm 0.66}$
	DAML	10.39$_{\pm 6.71}$	25.00$_{\pm 1.41}$	7.00$_{\pm 4.24}$	77.15$_{\pm 3.30}$
Average	Vanilla	12.61$_{\pm 1.50}$	27.00$_{\pm 3.89}$	5.25$_{\pm 1.41}$	72.26$_{\pm 1.81}$
	+English Pretrain	**14.75**$_{\pm 1.63}$	**27.88**$_{\pm 1.24}$	5.38$_{\pm 2.30}$	**74.18**$_{\pm 2.77}$
	DAML	13.74$_{\pm 2.46}$	27.63$_{\pm 0.53}$	**5.63**$_{\pm 1.94}$	72.06$_{\pm 1.73}$

Table 2: Performances of three benchmark models in terms of BLEU score, Inform Rate, Success Rate and Slot Accuracy for each language.

4.4 Results

The results of all three benchmark models for each language are included in the Table 2. From the table, we observe that pre-training on English MultiWOZ corpus improves all the metrics. With the English pre-training, the model does not only perform better on the dialog state tracking task, but also better on the language generation task. The improvement for non-English language indicates that dialog knowledge from the English pre-training data can be adapted to a new language through paralleled dialog data. Therefore, the similar structures that different languages share help the model to generalize to new languages based on the embedded information about English data.

The DAML approach, without introducing extra English corpus, improves the average BLEU score and inform rate. It also has the best success rate among all approaches. By forcing the model to learn the similar structures that different languages share, the DAML approach works well in the few-shot setting and outperforms the model pre-trained on English corpus for both Thai and Portuguese. The performance of DAML over the "Vanilla" setting also shows that parallel corpus brings significant advantages when the pre-trained multilingual models are used for downstream tasks in a language that we do not have a lot of available data.

5 Conclusion

We created a new multilingual dialog data with eight languages focusing on customer service tasks. We find that our model, which uses meta-learning to learn the shared structures between languages, performs significantly better than normal training in a few-shot setting and could achieve comparable results when there is enough training data.

One limitation of our work is that we only have 8 languages so far. In future work, we plan to expand the dataset to 30 languages. In addition, we will study how to perform zero-shot generation on all languages, and how to improve performance on both tasks and generations in the zero-shot setting.

References

Roee Aharoni, M. Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *ArXiv*, abs/1903.00089.

N. Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, M. Johnson, M. Krikun, M. Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Y. Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, I. Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *EMNLP*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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*, pages 422–428, Marseille, France. European Language Resources Association.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Chelsea Finn, P. Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. In *NLPCC*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

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

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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 1297–1303, Hong Kong, China. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

N. Mrksic, Ivan Vulic, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gasic, A. Korhonen, and S. Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

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

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation inconsistency and entity bias in MultiWOZ. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 326–337, Singapore and Online. Association for Computational Linguistics.

Kun Qian and Z. Yu. 2019. Domain adaptive dialog generation via meta learning. In *ACL*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65:569–631.

Sebastian Schuster, S. Gupta, Rushin Shah, and M. Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. *ArXiv*, abs/1810.13327.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 73–81, Suntec, Singapore. Association for Computational Linguistics.

K. Song, X. Tan, Tao Qin, Jianfeng Lu, and T. Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, A. Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *ArXiv*, abs/2010.11934.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *ArXiv*, abs/2104.00773.

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. *arXiv preprint arXiv:2007.12720*.

A Licenses for Relevant Artifacts

- MultiWOZ: Apache License 2.0
- mT5: Apache License 2.0
- DAML: CC BY-NC-SA 4.0

B Experimental Details

B.1 Hyper-Parameters

For the meta-learning approach: The meta-learning rate is 3×10^{-4}. The fine-tune learning rate is 5×10^{-4}. The batch size is 1. The maximum epoch number for the meta-learning phase is 5. The maximum epoch number for fine-tuning phase is 80. For other approaches: The fine-tune learning rate is 8×10^{-4}. The batch size is 1. The maximum epoch number for the pre-training phase is 3. The maximum epoch number for fine-tuning phase is 10.

B.2 Computational Budget

All experiments are run on NVIDIA RTX A6000. The total running time is around 80 hours.

C Human Annotation

C.1 Guidelines

Here is the full text of instructions given to participants: First, you should translate the slots in MultiWoz into the target language that you are working on. Then, you should correct the automatic translations into fluent utterances. All mentioned slots should be present in the corrected utterance and consistent with the slot translation. The purpose of the data is to build multilingual dialog systems.

C.2 Recruitment Details

We recruit 7 native speakers to correct the corresponding machine translations in 7 languages. The total payment is $50 per person.