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).

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 translation. We recruit native speakers for each language except English, we recruit a bilingual speaker to edit the machine-translated utterances. Our dataset currently covers 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 dialogs (1476 turns) from the test set. 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.

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

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

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.
<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>BLEU</th>
<th>Inform</th>
<th>Success</th>
<th>Slot Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Vanilla</td>
<td>15.34±1.90</td>
<td>43.00±2.40</td>
<td>6.00±2.83</td>
<td>71.07±1.75</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>20.39±3.92</td>
<td>57.00±1.23</td>
<td>11.00±1.23</td>
<td>75.27±2.19</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>18.32±0.58</td>
<td>54.00±2.83</td>
<td>7.00±2.42</td>
<td>73.29±2.10</td>
</tr>
<tr>
<td>French</td>
<td>Vanilla</td>
<td>16.93±1.51</td>
<td>26.00±2.83</td>
<td>5.00±2.41</td>
<td>70.48±2.86</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>19.94±3.12</td>
<td>23.00±1.41</td>
<td>4.00±2.00</td>
<td>73.74±2.83</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>18.07±0.74</td>
<td>19.00±0.85</td>
<td>6.00±2.00</td>
<td>65.80±3.89</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>Vanilla</td>
<td>16.95±1.09</td>
<td>23.00±1.41</td>
<td>6.00±2.83</td>
<td>68.46±1.72</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>19.63±1.15</td>
<td>25.00±1.41</td>
<td>6.00±2.83</td>
<td>76.36±2.76</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>18.89±1.74</td>
<td>23.00±1.41</td>
<td>5.00±2.41</td>
<td>70.61±2.36</td>
</tr>
<tr>
<td>Portuguese</td>
<td>Vanilla</td>
<td>13.35±4.51</td>
<td>23.00±1.41</td>
<td>5.00±2.41</td>
<td>75.13±5.16</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>14.13±2.16</td>
<td>24.00±0.00</td>
<td>4.00±2.00</td>
<td>69.68±1.75</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>16.66±1.94</td>
<td>24.00±0.00</td>
<td>6.00±2.83</td>
<td>72.82±2.68</td>
</tr>
<tr>
<td>Korean</td>
<td>Vanilla</td>
<td>7.29±0.09</td>
<td>24.00±0.00</td>
<td>4.00±2.00</td>
<td>69.93±1.42</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>9.59±1.17</td>
<td>25.00±1.41</td>
<td>4.00±2.00</td>
<td>78.04±0.77</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>8.15±2.15</td>
<td>23.09±1.41</td>
<td>4.00±2.00</td>
<td>74.42±1.92</td>
</tr>
<tr>
<td>Mandarin</td>
<td>Vanilla</td>
<td>3.42±1.56</td>
<td>26.00±2.83</td>
<td>5.00±2.41</td>
<td>74.53±4.98</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>7.13±2.10</td>
<td>24.00±0.00</td>
<td>6.00±2.83</td>
<td>79.40±2.35</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>3.93±0.29</td>
<td>24.00±0.00</td>
<td>4.00±2.00</td>
<td>69.83±1.81</td>
</tr>
<tr>
<td>Hindi</td>
<td>Vanilla</td>
<td>15.24±0.70</td>
<td>26.00±2.83</td>
<td>5.00±2.41</td>
<td>71.62±2.36</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>16.37±2.26</td>
<td>23.00±1.41</td>
<td>4.00±2.00</td>
<td>75.79±1.56</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>14.55±5.44</td>
<td>23.00±1.41</td>
<td>6.00±2.83</td>
<td>72.52±1.44</td>
</tr>
<tr>
<td>Thai</td>
<td>Vanilla</td>
<td>12.36±0.82</td>
<td>25.00±1.41</td>
<td>6.00±2.83</td>
<td>76.84±2.36</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>10.76±1.43</td>
<td>22.00±2.83</td>
<td>4.00±2.00</td>
<td>65.17±0.66</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>10.39±2.71</td>
<td>25.00±1.41</td>
<td>7.00±2.42</td>
<td>77.15±3.30</td>
</tr>
<tr>
<td>Average</td>
<td>Vanilla</td>
<td>12.61±1.50</td>
<td>27.00±3.89</td>
<td>5.25±2.41</td>
<td>72.26±1.81</td>
</tr>
<tr>
<td></td>
<td>+English Pretrain</td>
<td>14.75±2.63</td>
<td>27.88±1.24</td>
<td>5.38±2.30</td>
<td>74.18±2.77</td>
</tr>
<tr>
<td></td>
<td>DAML</td>
<td>13.74±2.46</td>
<td>27.63±0.53</td>
<td>5.63±1.94</td>
<td>72.06±1.73</td>
</tr>
</tbody>
</table>

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 languages indicates that dialog knowledge from the English pre-training data can be adapted to a new language through parallelized 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


Ting Han, Ximing Liu, Ryuichi Takenobu, Yixin Lian, Chongxuan Huang, Daizhen Wang, 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 NLPPC.


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


Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou,


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 $3 \times 10^{-4}$. The fine-tune learning rate is $5 \times 10^{-4}$. 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 $8 \times 10^{-4}$. 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.