Multi^{2}WOZ: A Robust Multilingual Dataset and Conversational Pretraining for Task-Oriented Dialog

Anonymous ACL submission

Abstract

Research on (multi-domain) task-oriented dialog (TOD) has predominantly focused on the *English* language, primarily due to the shortage of robust TOD datasets in other languages, preventing the systematic investigation of cross-lingual transfer for this crucial NLP application area. In this work, we introduce MULTI^{2}WOZ, a new multilingual multi-domain TOD dataset, derived from the well-established English dataset MULTIWOZ, that spans four typologically diverse languages: Chinese, German, Arabic, and Russian. In contrast to concurrent efforts (Ding et al., 2021; Zuo et al., 2021), MULTI^{2}WOZ contains gold-standard dialogs in target languages that are directly comparable with development and test portions of the English dataset, enabling reliable and comparative estimates of cross-lingual transfer performance for TOD. We then introduce a new framework for *multilingual conversational specialization* of pretrained language models (PrLMs) that aims to facilitate cross-lingual transfer for arbitrary downstream TOD tasks. Using such conversational PrLMs specialized for concrete target languages, we systematically benchmark a number of zero-shot and few-shot cross-lingual transfer approaches on two standard TOD tasks: Dialog State Tracking and Response Retrieval. Our experiments show that, in most setups, the best performance entails the combination of (i) conversational specialization in the target language and (ii) few-shot transfer for the concrete TOD task. Most importantly, we show that our conversational specialization in the target language allows for an exceptionally *sample-efficient few-shot transfer* for downstream TOD tasks.

1 Introduction

Task-oriented dialog (TOD) is arguably one of the most popular NLP application areas (Yan et al., 2017; Henderson et al., 2019, *inter alia*), with more importance recently given to more realistic, multi-domain conversations (Budzianowski et al., 2018; Ramadan et al., 2018) (in which users may handle more than one task during the conversation, e.g., booking a *taxi* and making a reservation at a *restaurant*). Unlike for many other NLP tasks (Hu et al., 2020; Liang et al., 2020; Ponti et al., 2020, *inter alia*), the progress towards *multilingual multi-domain* TOD has been hindered by the lack of sufficiently large and high-quality datasets in languages other than English (Budzianowski et al., 2018; Zang et al., 2020) and more recently, Chinese (Zhu et al., 2020). Creating TOD datasets for new languages from scratch or via translation of English datasets is significantly more expensive and time-consuming than for most other NLP tasks. The absence of multilingual datasets that are comparable (i.e., aligned) across languages prevents a reliable estimate of effectiveness of cross-lingual transfer techniques in multi-domain TOD (Razumovskaia et al., 2021).

In order to address these gaps, in this work we introduce MULTI^{2}WOZ, a reliable and large multilingual evaluation benchmark for multi-domain task-oriented dialog, derived by translating the monolingual English-only MultiWOZ data (Budzianowski et al., 2018; Eric et al., 2020) to four linguistically diverse major world languages, each with a different script: Arabic (AR), Chinese (ZH), German (DE), and Russian (RU).

Compared to the products of concurrent efforts that derive multilingual datasets from English MultiWOZ (Ding et al., 2021; Zuo et al., 2021), our MULTI^{2}WOZ is: (1) much *larger* – we translate all dialogs from development and test portions of the English MultiWOZ (in total 2,000 dialogs containing the total of 29.5K utterances); (2) much more *reliable* – complete dialogs, utterances as well as slot-values – have been manually translated (without resorting to error-prone heuristics), and the quality of translations has been validated through quality control steps; and (3) *parallel* – the same set of dialogs has been translated to all target

languages, enabling direct comparison of performance of multilingual models and cross-lingual transfer approaches across languages.

We then use MULTI^{2}WOZ to benchmark a range of state-of-the-art zero-shot and few-shot methods for cross-lingual transfer in two standard TOD tasks: Dialog State Tracking (DST) and Response Retrieval (RR). As the second main contribution of our work, we propose a general framework for improving performance and sample-efficiency of cross-lingual transfer for TOD tasks. We first leverage the parallel conversational OpenSubtitles corpus (Lison and Tiedemann, 2016) to carry out a conversational specialization of a PrLM for a given target language, irrespective of the downstream TOD task of interest. We then show that this intermediate conversational specialization in the target language (i) consistently improves the DST and RR performance in both zero-shot and few-shot transfer, and (ii) drastically improves sample efficiency of few-shot transfer.

2 Multi^{2}WOZ

In this section we describe the construction of the MULTI^{2}WOZ dataset, providing also details on inter-translator reliability. We then discuss two concurrent efforts in creating multilingual TOD datasets from MultiWOZ and their properties, and emphasize the aspects that make MULTI^{2}WOZ a more reliable and useful benchmark for evaluating cross-lingual transfer for TOD.

2.1 Dataset Creation

Language Selection. We translate all 2,000 dialogs from the development and test portions of the English MultiWOZ 2.1 (Eric et al., 2020) dataset to Arabic, Chinese, German, and Russian. Languages were selected based on the following criteria: (1) linguistic diversity (DE and RU belong to different Indo-European subfamilies – Germanic and Slavic, respectively; ZH is a Sino-Tibetan language and AR Semitic), (2) diversity of scripts (DE and RU use Latin and Cyrillic scripts, respectively, both *alphabet* scripts; AR script represents the *Abjad* script type, whereas the ZH Hanzi script belongs to *logographic* scripts), (3) number of native speakers (all four are in the top 20 most-spoken world languages), and (4) access to native and fluent speakers of those languages (also proficient in English).

Two-Step Translation. Following the well-established practice, we carried out a two-phase translation of the English data: (1) an *automatic translation* of the dialogs – utterances as well as the annotated slot values – followed by (2) the *manual post-editing*. We first automatically translated all utterances and slot values from the development and test dialogs from the MultiWOZ 2.1 (Eric et al., 2020) (1,000 dialogs in each portion; 14,748 and 14,744 utterances, respectively) to our four target languages, using Google Translate.[1] We then hired two native speakers of each target language,[2] all with a University degree and fluent in English, to post-edit the (non-overlapping sets of) automatic translations, i.e., fix the errors in automatic translations of utterances as well as slot values.

Since we carried out the automatic translation of the utterances independently of the automatic translation of the slot values, the translators were instructed to pay special attention to the alignment between each translated utterance and translations of slot value annotations for that utterance.

Quality Control. Human post-editors worked on disjunct sets of dialogs; we thus carried out an additional quality assurance step. Two new annotators for each target language judged the correctness of the translations on the random sample of 200 dialogs (10% of all translated dialogs, 100 from the development and test set each), containing 2,962 utterances in total. The annotators had to independently answer the following questions for each translated utterance from the sample: (1) *Is the utterance translation acceptable?* and (2) *Do the translated slot values match the translated utterance?* On average, across all target languages, both quality annotators for the respective language answered affirmatively to both questions for 99% of all utterances. Adjusting for chance agreement, we measured the Inter-Annotator Agreement (IAA) in terms of Cohen's κ (Cohen, 1960), observing the almost perfect agreement[3] of $\kappa = 0.824$ for the development set and $\kappa = 0.838$ for test set.

Annotation Duration and Cost. We hired 16 annotators in total – 4 per language: 2 for translation and 2 for quality assessment. The overall effort spanned almost full 5 months (from July to

[1] Relying on its Python API: https://pypi.org/project/googletrans

[2] In order to reduce the translation costs, we initially attempted to post-edit the translations via crowdsourcing. We tried this for Russian using the popular platform Toloka (toloka.yandex.com); however, the translation quality remained unsatisfactory even after several post-editing rounds.

[3] According to Landis and Koch (1977), if $\kappa \geq 0.81$.

November 2021), and amounted to 1,083 person-hours. With the remuneration rate of 16 $/h, MULTI^{2}WOZ creation cost $17,328 in total.

2.2 Comparison with Concurrent Work

Two concurrent works also derive multilingual datasets from MultiWOZ (Ding et al., 2021; Zuo et al., 2021), with different strategies and properties, discussed in what follows.

GlobalWOZ (Ding et al., 2021) encompasses Chinese, Indonesian, and Spanish datasets. The authors first create *templates* from dialog utterances by replacing slot-value strings in the utterances with the slot type and value index (e.g., *"...and the post code is cb238el"* becomes the template *"...and the post code is* `[attraction-postcode-1]`*"*. They then *automatically* translate all templates to the target languages. Next, they select a subset of 500 test set dialogs for human post-editing with the following heuristic: dialogs for which the sum of corpus-level frequencies of their constitutive 4-grams (normalized with the dialog length) is the largest.[4] Since this selection step is independent for each language, each GlobalWOZ portion contains translations of a different subset of English dialogs: this prevents any direct comparison of downstream TOD performance across languages. Even more problematically, the selection heuristic directly reduces linguistic diversity of dialogs chosen for the test set of each language, as it favors the dialogs that contain the same globally most frequent 4-grams. Due to this artificial homogeneity of its test sets, GlobalWOZ is very likely to overestimate downstream TOD performance for target languages.

Unlike GlobalWOZ, AllWOZ (Zuo et al., 2021) does automatic translation of a *fixed small* subset of MultiWOZ plus post-editing in seven target languages. However, it encompasses only 100 dialogs and 1,476 turns; as such, it is arguably too small to draw strong conclusions about performance of cross-lingual transfer methods. Its usefulness in joint domain and language transfer evaluations is especially doubtful, since it covers individual MultiWOZ domains with an extremely small number of dialogs (e.g., only 13 for the *Taxi* domain). Finally, neither Ding et al. (2021) nor Zuo et al. (2021) provide any estimates of the quality of their final datasets nor do they report their annotation costs.

[4] Interestingly, the authors do not provide any motivation or intuition for this heuristic. It is also worth noting that 4-gram frequencies are counted on the noisy automatic translations.

In contrast to GlobalWOZ, MULTI^{2}WOZ is a parallel corpus – with the exact same set of dialogs translated to all four target languages; as such it directly enables performance comparisons across the target languages. Further, containing translations of *all* dev and test dialogs from MultiWOZ (i.e., avoiding sampling heuristics), MULTI^{2}WOZ does not introduce any confounding factors that would distort estimates of cross-lingual transfer performance in downstream TOD tasks. Finally, MULTI^{2}WOZ is 20 times larger (per language) than AllWOZ: experiments on MULTI^{2}WOZ are thus much more likely to yield conclusive findings.

3 Cross-lingual Transfer for TOD

The parallel nature and sufficient size of MULTI^{2}WOZ allow us to benchmark and compare a number of established and novel cross-lingual transfer methods for TOD. In particular, (1) we first inject general conversational TOD knowledge into XLM-RoBERTa (XLM-R; Conneau et al., 2020), yielding TOD-XLMR (§3.1); (2) we then propose several variants for conversational specialization of TOD-XLMR for target languages, better suited for transfer in downstream TOD tasks (§3.2); (3) we investigate zero-shot and few-shot transfer for two TOD tasks: DST and RR (§3.3).

3.1 TOD-XLMR: A Multilingual TOD Model

Recently, Wu et al. (2020) demonstrated that specializing BERT (Devlin et al., 2019) on conversational data by means of additional pretraining via a combination of masked language modeling (MLM) and response selection (RS) objectives yields improvements in downstream TOD tasks. Following these findings, we first (propose to) conversationally specialize XLM-R (Conneau et al., 2020), a state-of-the-art multilingual PrLM covering 100 languages, in the same manner: applying the RS and MLM objectives on the same English conversational corpus consisting of nine human-human multi-turn TOD datasets (see Wu et al. (2020) for more details). As a result, we obtain TOD-XLMR – a massively multilingual PrLM specialized for task-oriented conversations. Note that TOD-XLMR is not yet specialized (i.e., fine-tuned) for any concrete TOD task (e.g., DST or Response Generation). Rather, it is enriched with general task-oriented conversational knowledge (in English), presumed to be beneficial for a wide variety of TOD tasks.

EN Subtitle	ZH Subtitle
- Professor Hall. - Yes. - I think your theory may be correct. - Walk with me. Just a few weeks ago, I monitored the strongest hurricane on record. The hail, the tornados, it all fits. Can your model factor in storm scenarios?	-霍尔教授 -是的 -我认为你的理论正确 -跟我来 上周我观测到史上最大的飓风 雹暴和龙卷风也符合你的理论 你能预测暴风雨的形成吗?

Translation LM (TLM)	- Professor Hall. - Yes. - I think your theory may be [MASK]. - Walk with...-霍尔教授 -是的 -我认为你的[MASK]正确...		
Response Selection (RS)	*Context:* 上周我观测到史上最大的飓风 雹暴和龙卷风也符合你的理论	**Monolingual (RS-Mono)** *True Response:* 你能预测暴风雨的形成吗? *False Response:* 你有彼得的电脑断层扫描吗?	**Cross-lingual (RS-X)** *True Response:* Can your model factor in storm scenarios? *False Response:* Do you have Peter's CT scan results?

Table 1: Examples of training instances for conversational specialization for the target language created from OpenSubtitles (OS). Top row: an example of a dialog created from OS, parallel in English and Chinese. Below are training examples for different training objectives: (1) *Translation Language Modelling* (**TLM**) on the interleaved English-Chinese parallel utterances; (2) two variants of *Response Selection* (RS) – (a) monolingual in the target language (**RS-Mono**) and (b) cross-lingual (**RS-X**).

3.2 Target-Language Specialization

TOD-XLMR has been conversationally specialized only on English data. We next hypothesize that a further conversational specialization for a concrete target language X can improve the transfer EN→X for all downstream TOD tasks. Accordingly, similar to Moghe et al. (2021), we investigate several intermediate training procedures that further conversationally specialize TOD-XLMR for the target language X (or jointly for EN and X). For this purpose, we (i) compile target-language-specific as well as cross-lingual corpora from the CCNet (Wenzek et al., 2020) and OpenSubtitles (Lison and Tiedemann, 2016) datasets and (ii) experiment with different mono-, bi-, and cross-lingual training procedures. Here, we propose a novel cross-lingual response selection (RS) objective and demonstrate its effectiveness in downstream TOD transfer.

Training Corpora. We collect two types of data for language specialization: (i) *"flat" corpora* (i.e., without any conversational structure): we simply randomly sample 100K sentences for each language from the respective monolingual portion of CCNet (we denote with *Mono-CC* the individual 100K-sentence portions of each language; with *Bi-CC* the concatenation of the English and each of target language *Mono-CCs*, and with *MultiCC* the concatenation of all five Mono-CC portions); (ii) *parallel dialogs* (in EN and target language X) from OpenSubtitles (OS), a parallel conversational corpus spanning 60 languages, compiled from subtitles of movies and TV series. We leverage the parallel OS dialogs to create two different cross-lingual specialization objectives, as described next.

Training Objectives. We directly use the CC portions (Mono-CC, Bi-CC, and Multi-CC) for standard **MLM** training. We then leverage the parallel OS dialogs for two training objectives. First, we carry out translation language modeling (**TLM**) (Conneau and Lample, 2019) on the synthetic dialogs which we obtain by interleaving K randomly selected English utterances with their respective target language translations; we then (as with MLM), dynamically mask 15% of tokens of such interleaved dialogs; we vary the size of the context the model can see when predicting missing tokens by randomly selecting K (between 2 and 15) for each instance. Second, we use OS to create instances for both monolingual and cross-lingual Response Selection (RS) training. RS is a simple binary classification task in which for a given pair of a *context* (one or more consecutive utterances) and *response* (a single utterance), the model has to predict whether the response utterance immediately follows the context (i.e., it is a *true* response) or not (i.e., it is a *false* response). RS pretraining has been proven beneficial for downstream TOD in monolingual English setups (Mehri et al., 2019; Henderson et al., 2019, 2020; Hung et al., 2021).

In this work, we leverage the parallel OS data to introduce the cross-lingual RS objective, where the context and the response utterance are not in the same language. In our experiments, we carry out both (i) monolingual RS training in the target language (i.e., both the context and response utterance are, e.g., in Chinese), denoted **RS-Mono**, and (ii) cross-lingual RS between English (as the source language in downstream TOD tasks) and the target language, denoted **RS-X**. We create *hard RS negatives*, by coupling contexts with non-immediate responses from the same movie or episode (same `imdbID`), as well as *easy negatives* by randomly sampling $m \in \{1, 2, 3\}$ responses from a different

movie of series episode (i.e., different `imdbID`). Hard negatives encourage the model to reason beyond simple lexical cues. Examples of training instances for OS-based training (for EN-ZH) are shown in Table 1.

3.3 Downstream Cross-lingual Transfer

Finally, we fine-tune the various variants of TOD-XLMR, obtained through the above-described specialization (i.e., intermediate training) procedures, for two downstream TOD tasks (DST and RR) and examine their cross-lingual transfer performance. We cover two cross-lingual transfer scenarios: (1) *zero-shot transfer* in which we only fine-tune the models on the English training portion of MultiWOZ and evaluate their performance on the MULTI^{2}WOZ test data of our four target languages; and (2) *few-shot transfer* in which we sequentially first fine-tune the models on the English training data and then on the small number of dialogs from the development set of MULTI^{2}WOZ, in similar vein to (Lauscher et al., 2020). In order to determine the effect of our conversational target language specialization (§3.2) on the downstream sample efficiency, we run few-shot experiments with different numbers of target language training dialogs, ranging from 1% to 100% of the size of MULTI^{2}WOZ development portions.

4 Experimental Setup

Evaluation Tasks and Measures. We evaluate different multilingual conversational PrLMs in cross-lingual transfer (zero-shot and few-shot) for two prominent TOD tasks: *dialog state tracking (DST)* and *response retrieval (RR)*.

DST is commonly cast as a multi-class classification task, where given a predefined ontology and dialog history (a sequence of utterances), the model has to predict the output state, i.e., *(domain, slot, value)* tuples (Wu et al., 2020).[5] We adopt the standard *joint goal accuracy* as the evaluation measure: at each dialog turn, it compares the predicted dialog states against the manually annotated ground truth which contains slot values for all the *(domain, slot)* candidate pairs. A prediction is considered correct if and only if all predicted slot values exactly match the ground truth.

RR is a ranking task that is well-aligned with the RS objective and relevant for retrieval-based TOD systems (Wu et al., 2017; Henderson et al., 2019): given the dialog context, the model ranks N dataset utterances, including the *true response* to the context (i.e., the candidate set includes the one *true* response and N-1 *false* responses). We follow Henderson et al. (2020) and report the results for $N = 100$, i.e., the evaluation measure is recall at the top 1 rank given 99 randomly sampled false responses, denoted as R_{100}@1.

Models and Baselines. We briefly summarize the models that we compare in zero-shot and few-shot cross-lingual transfer for DST and RR. As baselines, we report the performance of the vanilla multilingual PrLM XLM-R (Conneau et al., 2020)[6] and its variant further trained on the English TOD data from (Wu et al., 2020): TOD-XLMR (§3.1). Comparison between XLM-R and TOD-XLMR quantifies the effect of conversational English pretraining on downstream TOD performance, much like the comparison between BERT and TOD-BERT done by Wu et al. (2020); however, here we extend the comparison to cross-lingual transfer setups. We then compare the baselines against a series of our target language-specialized variants, obtained via intermediate training on CC (Mono-CC, Bi-CC, and Multi-CC) by means of MLM, and on OS jointly via TLM and RS (RS-X or RS-Mono) objectives (see §3.2 again).

Hyperparameters and Optimization. For training TOD-XLMR (§3.1), we select the effective batch size of 8. In target-language-specific intermediate training (§3.2), we fix the maximum sequence length to 256 subword tokens; for RS objectives, we limit the context and response to 128 tokens each. We train for 30 epochs in batches of size 16 for MLM/TLM, and 32 for RS. We search for the optimal learning rate among the following values: $\{10^{-4}, 10^{-5}, 10^{-6}\}$. We apply early stopping based on development set performance (patience: 3 epochs for MLM/TLM, 10 epochs for RS). In downstream fine-tuning, we train in batches of 6 (DST) and 24 instances (RR) with the initial learning rate fixed to $5 \cdot 10^{-5}$. We also apply early stopping (patience: 10 epochs) based on the development set performance, training maximally for 300 epochs in zero-shot setups, and for 15 epochs in target-language few-shot training. In all experiments, we use Adam (Kingma and Ba, 2015) as the optimization algorithm.

[5] The model is required to predict slot values for each *(domain, slot)* pair at each dialog turn.

[6] We use `xlm-roberta-base` from HuggingFace.

Model	DE	AR	ZH	RU	Avg.
w/o intermediate specialization					
XLM-R	1.41	1.15	1.35	1.40	1.33
TOD-XLMR	1.74	1.53	1.75	2.16	1.80
with conversational target-lang. specialization					
MLM on Mono-CC	3.57	2.71	3.34	5.17	3.70
Bi-CC	3.66	2.17	2.73	3.73	3.07
Multi-CC	3.65	2.35	2.06	5.39	3.36
TLM on OS	7.80	2.43	3.95	6.03	5.05
TLM + RS-X on OS	**7.84**	**3.12**	4.14	6.13	5.31
TLM + RS-Mono on OS	7.67	2.85	**4.47**	**6.57**	**5.39**

Table 2: Performance of multilingual conversational models in zero-shot cross-lingual transfer for Dialog State Tracking (DST) on MULTI^{2}WOZ, with joint goal accuracy (%) as evaluation metric. Reference English DST performance of TOD-XLMR: 47.86%.

Model	DE	AR	ZH	RU	Avg.
w/o intermediate specialization					
TOD-XLMR	3.3	2.9	1.9	2.7	2.7
with conversational target-lang. specialization					
MLM on Mono-CC	22.9	25.5	24.5	33.4	26.6
TLM on OS	**44.4**	30.3	34.1	39.3	37.0
TLM + RS-Mono on OS	44.3	**30.9**	**34.8**	**39.6**	**37.4**

Table 3: Performance of multilingual conversational models in zero-shot cross-lingual transfer for Response Retrieval (RR) on MULTI^{2}WOZ with R_{100}@1 (%) as the evaluation metric. Reference English RR performance of TOD-XLMR: 64.75%

5 Results and Discussion

We now present and discuss the downstream cross-lingual transfer results on MULTI^{2}WOZ for DST and RR in two different transfer setups: zero-shot transfer and few-shot transfer.

5.1 Zero-Shot Transfer

Dialog State Tracking. Table 2 summarizes zero-shot cross-lingual transfer performance for DST. First, we note that the transfer performance of all models for all four target languages is extremely low, drastically lower than the reference English DST performance of TOD-XLMR, which stands at 47.9%. These massive performance drops, stemming from cross-lingual transfer are in line with findings from concurrent work (Ding et al., 2021; Zuo et al., 2021) and suggest that reliable cross-lingual transfer for DST is much more difficult to achieve than for most other language understanding tasks (Hu et al., 2020; Ponti et al., 2020).

Despite low performance across the board, we do note a few emerging and consistent patterns. First, TOD-XLMR slightly but consistently outperforms the vanilla XLM-R, indicating that *conversational* English pretraining brings marginal gains. All of our proposed models from §3.2 (the lower part of Table 2) substantially outperform TOD-XLMR, proving that intermediate conversational specialization for the target language brings gains, irrespective of the training objective.

Expectedly, TLM and RS training on parallel OS data brings substantially larger gains than MLM-ing on flat monolingual target-language corpora (Mono-CC) or simple concatenations of corpora from two (Bi-CC) or more languages (Multi-CC). German and Arabic seem to benefit slightly more from the cross-lingual Response Selection training (RS-X), whereas for Chinese and Russian we obtain better results with the monolingual (target language) RS training (RS-Mono).

Response Retrieval. The results of zero-shot transfer for RR are summarized in Table 3. Compared to DST results, for the sake of brevity, we show the performance of only the stronger baseline (TOD-XLMR) and the best-performing variants with intermediate conversational target-language training (one for each objective type): MLM on Mono-CC, TLM on OS, and TLM + RS-Mono on OS. Similar to DST, TOD-XLMR exhibits a near-zero cross-lingual transfer performance for RR too, across all target languages. In sharp contrast to DST results, however, conversational specialization for the target language – with any of the three specialization objectives – massively improves the zero-shot cross-lingual transfer for RR. The gains are especially large for the models that employ the parallel OpenSubtitles corpus in intermediate specialization, with the monolingual (target language) Response Selection objective slightly improving over TLM training alone.

Given the parallel nature of MULTI^{2}WOZ, we can directly compare transfer performance of both DST and RR across the four target languages. In both tasks, the best-performing models exhibit stronger performance (i.e., smaller performance drops compared to the English performance) for German and Russian than for Arabic and Chinese. This aligns well with the linguistic proximity of the target languages to English as the source language.

5.2 Few-Shot Transfer and Sample Efficiency

Next, we present the results of few-shot transfer experiments, where we additionally fine-tune the

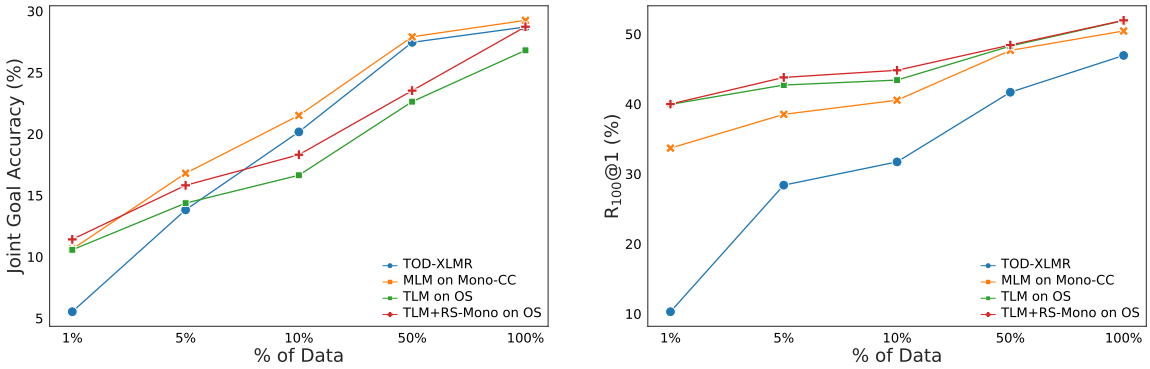

Figure 1: Few-shot cross-lingual transfer results for Dialog State Tracking (left figure) and Response Retrieval (right figure), averaged across all four target languages (detailed per-language results available in the Appendix). Results shown for different sizes of the training data in the target-language (i.e., different number of *shots*): 1%, 5%, 10%, 50% and 100% of the MULTI^{2}WOZ development sets (of respective target languages).

		DST					RR				
Lang	**Model**	**1%**	**5%**	**10%**	**50%**	**100%**	**1%**	**5%**	**10%**	**50%**	**100%**
DE	TOD-XLMR	7.68	19.26	28.08	33.17	34.10	10.25	32.47	35.56	45.39	49.46
	TLM+RS-Mono on OS	15.88	24.14	28.38	32.57	35.45	46.08	48.94	49.98	53.43	55.72
AR	TOD-XLMR	1.48	1.57	6.18	15.62	17.63	6.36	18.72	23.57	36.04	42.69
	TLM+RS-Mono on OS	4.42	6.79	8.27	14.39	21.48	33.45	37.09	38.01	41.89	47.15
ZH	TOD-XLMR	8.63	12.55	16.40	23.45	25.49	15.69	31.10	33.22	41.97	48.14
	TLM+RS-Mono on OS	11.63	14.90	17.97	22.81	28.84	38.45	43.71	45.27	48.50	51.81
RU	TOD-XLMR	4.34	21.89	30.01	37.58	37.61	8.90	31.31	34.51	43.33	47.45
	TLM+RS-Mono on OS	13.74	17.44	18.63	24.33	29.15	41.97	45.44	46.02	49.90	53.16

Table 4: Per-language few-shot transfer performance (sample efficiency results) on DST and RR for the baseline TOD-XLMR and the best specialized model (TLM+RS-Mono on OS).

task-specific TOD model on a limited number of target-language dialogs from the development portion of MULTI^{2}WOZ, after first fine-tuning it on the complete English training set from MultiWOZ (see §4). Few-shot cross-lingual transfer results, averaged across all four target languages, are summarized in Figure 1. The figure shows the performance for different sizes of the target-language training data (i.e., number of target-language shots, that is, percentage of the target-language development portion from MULTI^{2}WOZ). Detailed per-language few-shot results are given in Table 4, for brevity only for TOD-XLMR and the best target-language-specialized model (TLM+RS-Mono on OS). We provide full per-language results for all specialized models from Figure 1 in the Appendix.

The few-shot results unambiguously show that the intermediate conversational specialization for the target language(s) *drastically improves the target-language sample efficiency in the downstream few-shot transfer*. The baseline TOD-XLMR – not exposed to any type of conversational pretraining for the target language(s) – exhibits substantially lower performance than all three models (MLM on Mono-CC, TLM on OS, and TLM+RS-Mono on OS) that underwent conversational intermediate training on respective target languages. This is evident even in the few-shot setups where the three models are fine-tuned on merely 1% (10 dialogs) or 5% (50 dialogs) of the MULTI^{2}WOZ development data (after prior fine-tuning on the complete English task data from MultiWOZ).

As expected, the larger the number of task-specific (DST or RR) training instances in the target languages (50% and 100% setups), the closer the performance of the baseline TOD-XLMR gets to the best-performing target-language-specialized model – this is because the size of the in-language training data for the concrete task (DST or RR) becomes sufficient to compensate for the lack of conversational target-language intermediate training that the specialized models have been exposed to. The sample efficiency of the conversational target-language specialization is more pronounced for RR than for DST. This seems to be in line with the zero-shot transfer results (see Tables 2 and 3), where the specialized models displayed much larger cross-lingual transfer gains over TOD-XLMR on RR than

on DST. We hypothesize that this is due to the intermediate specialization objectives (especially RS) being better aligned with the task-specific training objective of RR than that of DST.

6 Related Work

TOD Datasets. Research in task-oriented dialog has been, for a long time, limited by the existence of only monolingual English datasets. While earlier datasets focused on a single domain (Henderson et al., 2014a,b; Wen et al., 2017), the focus shifted towards the more realistic multi-domain task-oriented dialogs with the creation of the MultiWOZ dataset (Budzianowski et al., 2018), which has been refined and improved in several iterations (Eric et al., 2020; Zang et al., 2020; Han et al., 2021). Due to the particularly high costs of creating TOD datasets (in comparison with other language understanding tasks) (Razumovskaia et al., 2021), only a handful of monolingual TOD datasets in languages other than English (Zhu et al., 2020) or bilingual TOD datasets have been created (Gunasekara et al., 2020; Lin et al., 2021). Mrkšić et al. (2017b) were the first to translate 600 dialogs from the single-domain WOZ 2.0 (Mrkšić et al., 2017a) to Italian and German. Concurrent work (Ding et al., 2021; Zuo et al., 2021), which we discuss in detail in §2.2 and compare thoroughly with our MULTI^{2}WOZ, introduces the first multilingual multi-domain TOD datasets, created by translating portions of MultiWOZ to several languages.

Language Specialization and Cross-lingual Transfer. Multilingual transformer-based models (e.g., mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020)) are pretrained on large general-purpose and massively multilingual corpora (over 100 languages). While this makes them versatile and widely applicable, it does lead to suboptimal representations for individual languages, a phenomenon commonly referred to as the "curse of multilinguality" (Conneau et al., 2020). Therefore, one line of research focused on adapting (i.e., *specializing*) those models to particular languages (Lauscher et al., 2020; Pfeiffer et al., 2020). For example, Pfeiffer et al. (2020) propose a more computationally efficient approach for extending the model capacity for individual languages: this is done by augmenting the multilingual PrLM with language-specific adapter modules. Glavaš et al. (2020) perform language adaptation through additional intermediate masked language modeling in the target languages with filtered text corpora, demonstrating substantial gains in downstream zero-shot cross-lingual transfer for hate speech and abusive language detection tasks. In a similar vein, Moghe et al. (2021) carry out intermediate fine-tuning of multilingual PrLMs on parallel conversational datasets and demonstrate its effectiveness in zero-shot cross-lingual transfer for the DST task.

Lauscher et al. (2020) show that few-shot transfer, in which one additionally fine-tunes the PrLM on few labeled task-specific target-language instances leads to large improvements for many task-and-language combinations, and that labelling few target-language examples is more viable than further LM-specialization for languages of interest under strict zero-shot conditions. This finding is also corroborated in our work for two TOD tasks.

7 Conclusion

Task-oriented dialog (TOD) has predominantly focused on *English*, primarily due to the lack of robust TOD datasets in other languages (Razumovskaia et al., 2021), preventing systematic investigations of cross-lingual transfer in this crucial NLP application area. In this work, we have presented MULTI^{2}WOZ – a robust multilingual multi-domain TOD dataset. MULTI^{2}WOZ encompasses gold-standard dialogs in four languages that are directly comparable with development and test portions of the English MultiWOZ dataset, thus allowing for the most reliable and comparable estimates of cross-lingual transfer performance for TOD to date. Further, we presented a framework for *multilingual conversational specialization* of pretrained language models that facilitates cross-lingual transfer for downstream TOD tasks. Our experiments on MULTI^{2}WOZ for two prominent TOD tasks – Dialog State Tracking and Response Retrieval – reveal that the cross-lingual transfer performance benefits from both (i) intermediate conversational specialization for the target language and (ii) few-shot cross-lingual transfer for the concrete downstream TOD task. Crucially, we show that our novel conversational specialization for the target language leads to *exceptional sample efficiency* in downstream few-shot transfer.

In hope to steer and inspire future research on multilingual and cross-lingual TOD, we make MULTI^{2}WOZ publicly available at: `URL-HIDDEN`. We will extend the resource to further languages from yet uncovered families (e.g., Turkish).

References

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Bosheng Ding, Junjie Hu, Lidong Bing, Sharifah Aljunied Mahani, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2021. Globalwoz: Globalizing multiwoz to develop multilingual task-oriented dialogue systems. *CoRR*, abs/2110.07679.

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.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*.

Ting Han, Ximing Liu, Ryuichi Takanabu, 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 *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 206–218. Springer.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014a. 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.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014b. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy. Association for Computational Linguistics.

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

Chia-Chien Hung, Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavas. 2021. DS-TOD: efficient domain specialization for task oriented dialog. *CoRR*, abs/2110.08395.

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

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

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*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. Cross-lingual intermediate fine-tuning improves dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017a. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017b. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Osman Ramadan, Paweł Budzianowski, and Milica Gašić. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 432–437, Melbourne, Australia. Association for Computational Linguistics.

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

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: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented

dialogue systems for online shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4618–4625. AAAI Press.

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.

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*, 8:281–295.

Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. Allwoz: Towards multilingual task-oriented dialog systems for all. *CoRR*, abs/2112.08333.

A Appendix

		DST					RR				
Lang	**Model**	**1%**	**5%**	**10%**	**50%**	**100%**	**1%**	**5%**	**10%**	**50%**	**100%**
DE	TOD-XLMR	7.68	19.26	28.08	33.17	34.10	10.25	32.47	35.56	45.39	49.46
	MLM on Mono-CC	13.75	25.15	34.12	38.01	38.26	34.37	42.13	43.51	49.10	52.80
	TLM on OS	14.17	19.45	21.62	27.28	29.91	47.21	48.59	48.96	53.01	55.30
	TLM+RS-Mono on OS	15.88	24.14	28.38	32.57	35.45	46.08	48.94	49.98	53.43	55.72
AR	TOD-XLMR	1.48	1.57	6.18	15.62	17.63	6.36	18.72	23.57	36.04	42.69
	MLM on Mono-CC	4.41	5.74	7.02	14.10	17.22	28.54	31.50	32.82	41.09	44.26
	TLM on OS	4.18	6.33	6.89	13.60	17.77	32.19	35.04	37.02	41.39	47.04
	TLM+RS-Mono on OS	4.42	6.79	8.27	14.39	21.48	33.45	37.09	38.01	41.89	47.15
ZH	TOD-XLMR	8.63	12.55	16.40	23.45	25.49	15.69	31.10	33.22	41.97	48.14
	MLM on Mono-CC	11.64	19.73	25.46	34.93	35.61	34.40	37.65	39.65	48.01	50.97
	TLM on OS	11.48	17.43	21.95	28.52	32.51	38.17	42.82	42.91	49.29	51.63
	TLM+RS-Mono on OS	11.63	14.90	17.97	22.81	28.84	38.45	43.71	45.27	48.50	51.81
RU	TOD-XLMR	4.34	21.89	30.01	37.58	37.61	8.90	31.31	34.51	43.33	47.45
	MLM on Mono-CC	12.70	16.56	19.45	24.58	25.90	37.43	42.80	46.19	52.43	53.73
	TLM on OS	12.45	14.26	16.10	21.13	27.04	42.23	44.40	44.78	49.43	53.76
	TLM+RS-Mono on OS	13.74	17.44	18.63	24.33	29.15	41.97	45.44	46.02	49.90	53.16

Table 5: Full per-language few-shot cross-lingual transfer results for Dialog State Tracking and Response Retrieval. Results shown for different sizes of the training data in the target-language (i.e., different number of *shots*): 1%, 5%, 10%, 50% and 100% of the MULTI^{2}WOZ development sets (of respective target languages).