Towards Zero-Shot Multilingual Transfer for Code-Switched Responses

Ting-Wei Wu[1]*, **Changsheng Zhao**[2], **Ernie Chang**[2], **Yangyang Shi**[2],
Pierce Chuang[2], **Vikas Chandra**[2], **Biing Juang**[1]
[1]Georgia Institute of Technology
[2]Meta Reality Labs
waynewu@gatech.edu, juang@ece.gatech.edu,
{cszhao, erniecyc, yyshi, pichuang, vchandra}@meta.com

Abstract

Recent task-oriented dialog systems obtained great successes in building personal assistants for high resource language such as English, but extending these systems to a global audience is challenging due to the need for annotated data or machine translation systems in the target language. An alternative approach is to leverage existing data in a high-resource language to enable cross-lingual transfer in low-resource language models. However, this type of transfer has not been widely explored in natural language response generation. In this research, we investigate the use of state-of-the-art multilingual models such as mBART and T5 to facilitate zero-shot and few-shot transfer of code-switched responses. We propose a new adapter-based framework that allows for efficient transfer by learning jointly the task-specific, source and target language representations. Our framework is able to successfully transfer language knowledge even when the target language corpus is limited. We present both quantitative and qualitative analyses to evaluate the effectiveness and limitations of our approach[1].

1 Introduction

Recent task-oriented dialog systems (ToD) have achieved great success in intelligently communicating with humans in natural languages (Chen et al., 2017; Bohus and Rudnicky, 2009). They are designed to fully assist users with widely heralded applications such as music playing, ticket ordering, or customer servicing (Zhang et al., 2020c). However, most ToD systems are primarily established for English due to its ubiquity and the abundance of high-quality human annotations (Serban et al., 2015). Extending these services to global users may take tremendous efforts, especially in low-resource languages where the collection of training corpus is labor-intensive.

On the other hand, given a sizeable English dialog corpus with standard dialog features shared across other languages, it is possible to transfer the knowledge and logic between languages via machine translation or cross-task alignments. Data-driven approaches (Schuster et al., 2019; Xiang et al., 2021) perform standard supervised training with translated dialogs, known as *Translate-Train*. Different pseudo-data pairs could be leveraged to enhance the multilingual model's robustness. Nevertheless, a fine-grained machine translation system may not exist in an extremely low-resource language. The translation errors of entities in ground truth annotations (e.g. *Indian* could be translated in Chinese either to an adjective of Indian or Indian people in different contexts.) can drastically influence how model is supervised. This primarily happens in dialog tasks like dialog state tracking (DST) or natural language response generation (NLG) with language-sensitive outputs.

Another line of approaches instead investigates cross-lingual transfer directly in pretrained multilingual language models (Tang et al., 2021; Gritta et al., 2022). In particular, the multilingual sequence-to-sequence model family (mSeq2seq), which learns to encode the hidden representation of a given input and generates relevant outputs, can achieve promising multi-task performance in different languages. Pretraining a multilingual encoder with machine translation (Schuster et al., 2019), mask context learning (Colombo et al., 2021) or learning a word alignment matrix (Liu et al., 2019) could possibly transfer knowledge between languages. These methods mostly share language-agnostic outputs (Shrivastava et al., 2021) for typical classification tasks to avert *off-target* problem, i.e., models (partially) translate its prediction into the wrong language during *zero-shot* transfer due to spurious correlation (Gu et al., 2019; Zhang et al.,

*This work was done during an internship at Meta Reality Labs.

[1]Code available at https://github.com/waynewu6250/zero-shot-multilingual-transfer-ACL-2023

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7551–7563
July 9-14, 2023 ©2023 Association for Computational Linguistics

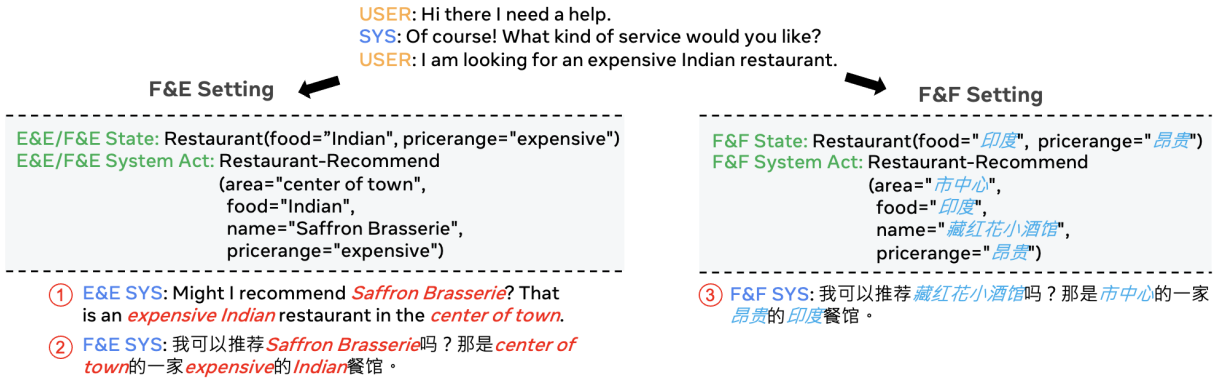

Figure 1: Examples of data formats for multilingual ToD systems. Each system response will come with a system act with different forms depending on use cases. ① E&E setting will have English sentences with English entities and dialog acts. ② F&E setting will still have entire English dialog acts but foreign responses with English entities embedded. ③ F&F will have all foreign responses but with code-switched dialog acts.

2020a). However, the cross-lingual performance of mSeq2seqs in more challenging response generation with language-specific or code-switched outputs remains mysteriously unexplored.

Herein, we present a study on the cross-lingual transferability of mSeq2seqs and quantify how well these models could adapt to reasonable multilingual response generation under meager availability of dialog annotation in a target language (*few-shot*). Given a pair of designed input-output sequences, we propose Cross-lingual Dialog Fusion (**XDFusion**) that employs mSeq2seqs to quickly adapt to downstream NLG tasks in target low-resource languages by inserting denoising-trained language adapters and a knowledge fusion module. In particular, we first fine-tune mSeq2seq models with the English dialog generation task. Then we insert both pretrained source and target language adapters and an additional fusion module within the fine-tuned models to merge the language-specific knowledge and fine-tune with target languages. We conduct our experiments on a multilingual multi-domain ToD dataset: GlobalWOZ (Ding et al., 2022). It is a multilingual extension of an English ToD dataset for DST, MultiWoZ (Budzianowski et al., 2018). Both quantitative and qualitative results show that our proposed adapter-based framework benefits from multilingual pretraining power and abundant English resources as it outperforms several baselines with deficient target language availability.

To this end, our contributions are the following:

1. We investigate and benchmark the transferability of large multilingual pretrained models in the low-resource dialog generation task.
2. We propose an adapter-based learning framework that shows large improvements in BLEU and the slow error rate by preserving English entities in a code-switched foreign language response.
3. The proposed method allows quick adaptation of training a new fusion module to support a new language while ameliorating the limited parameter capacity of pretrained models.

2 Problem Formulation

2.1 Data Format

We mainly follow Madotto et al. (2021) to model ToD systems as a Seq2seq generation module using annotated formats in existing ToD dialog datasets that can generate natural responses in an allocated target language. As shown in Figure 1 of a data sample, each dialog will contain several turns of user utterances (USER) and system utterances (SYS). We first define the dialog history H as the concatenation of the alternating utterances from the user and system turns, respectively, without the last system utterance which we denote as S. Each system utterance comes with a system dialog act S_{ACT} denoted as the concatenation of the intent $\mathbf{I}$ and slot-value pairs (s, v) as follows:

$$S_{ACT} = \mathbf{I}(s_1 = v_1, \ldots, s_k = v_k) \quad (1)$$

Without loss of generality, we define the modularized system response generation task as input-output pairs to benchmark the transferability performance of mSeq2seqs:

$$H + \underbrace{\mathbf{I}(s_1 = v_1, \ldots, s_k = v_k)}_{S_{ACT}} \rightarrow S \quad (2)$$

S_{ACT} could be empty sometimes where the task becomes a direct mapping between dialog history to the ideal system response $H \rightarrow S$.

We first refer the dataset of English sentences with English entities as **E&E**. Due to frequent code-switching phenomena, besides English-only sentences, the GlobalWOZ dataset also provides two additional use cases for other foreign dialogs: Foreign sentences with foreign local entities (**F&F**) and foreign sentences with local English entities (**F&E**). The key discrepancy lies in whether local name entities in the sentences remain in English, which will determine a language-agnostic/specific S_{ACT}, as shown in Figure 1. **E&E** and **F&E** will have language-agnostic acts while **F&F** will have language-specific acts which is considered more challenging in cross-lingual transfer.

2.2 Seq2seq Model & Setting

Based on the input-output definition in Section 2.1, we can prepare the dialog dataset as $\mathcal{D}_K = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, where $(x^{(i)}, y^{(i)})$ is a pre-defined input-output pair from one of the three settings in consideration (**E&E**, **F&F**, **F&E**) and K is the language of a dataset (e.g., Chinese). In this paper, we mainly employ mSeq2seqs (e.g., mBART (Tang et al., 2021), mT5 (Xue et al., 2020)), which provide suitable parameter initialization to model the new conditional distribution. Given the input text sequence $x^{(i)} = (x_1^{(i)}, ..., x_L^{(i)})$ with length L, we leverage the Seq2seq encoder-decoder architecture to maximize the conditional log-likelihood $\log p_\theta(y|x)$ where $y^{(i)} = (y_1^{(i)}, ..., y_T^{(i)})$ with length T is the output text sequence:

$$\mathcal{L}_{MLE}(\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T} \log p_\theta(y_t^{(i)}|y_{<t}^{(i)}, x^{(i)}) \quad (3)$$

$$p_\theta(y_t^{(i)}|y_{<t}^{(i)}, x^{(i)}) = softmax(\mathbf{W}h_t^{(i)} + b) \quad (4)$$

$$h_t^{(i)} = g(y_{t-1}^{(i)}, f(x^{(i)}; \theta); \theta) \quad (5)$$

Following the standard taxonomy for *Zero-shot cross-lingual transfer* and *Few-shot cross-lingual transfer* setting (Ding et al., 2022), we investigate the model transfer capability based on the available resources during training. In *zero-shot* setting, we are only given a high-quality set of human-annotated English ToD data $\mathcal{D}_{En}$. We directly train the Seq2seq model with the defined input-output pairs, including English data and data translated from English using a machine translation system. In *few-shot* setting where we have further access to a small budget of foreign ToD data $\mathcal{D}_{Fo}$ during training to induce *few-shot* learning. Particularly, we include a small set (100 dialogs) of foreign ToD data in a target language during training and evaluate multilingual models' performance on NLG tasks. In summary, we mainly have three experimental settings (*Train data* $\rightarrow$ *Test data*) for benchmarking based on different language datasets to use († indicates only 100 dialogs available):

- Zero-shot F&F: $\mathcal{D}_{En} \rightarrow \mathcal{D}_{Fo}^{F\&F}$
- Few-shot F&F: $\mathcal{D}_{En} + \mathcal{D}_{Fo}^{F\&F\dagger} \rightarrow \mathcal{D}_{Fo}^{F\&F}$
- Few-shot F&E: $\mathcal{D}_{En} + \mathcal{D}_{Fo}^{F\&E\dagger} \rightarrow \mathcal{D}_{Fo}^{F\&E}$

3 Model Adaptation for Cross-lingual Dialog Transfer

3.1 Structural Fine-tuning

In the last section, we describe how we induce cross-lingual transfer by directly fine-tuning large mSeq2seqs on labeled data of response generation task in English and very few in a target language. However, models trained with extremely imbalanced data distribution may fail to generate reasonable target language responses and suffer from spurious correlation to source language (Gu et al., 2019). How to adequately extract relevant source language knowledge while preserving spaces for target language adaptation becomes crucial and challenging, more than just simple fine-tuning.

We instead split the training steps into separate phases that allows more exclusive parameter updates on source and target languages independently. In the first phase, we care more about learning the task-centralized knowledge agnostic of languages. We retain the original fine-tuning step of training large mSeq2seqs with English data only that can explicitly performs well on generating high-quality responses in English.

3.2 Language Adapters

Since the emphasis is on target language adaptation as well as avoiding catastrophic forgetting of the multilingual and task knowledge acquired from Section 3.1, adapter module is a great fit for parameter-efficient and quick fine-tuning to new tasks and domains (Rebuffi et al., 2017). Following MAD-X (Pfeiffer et al., 2020c) for cross-lingual transfer, we employ a recent efficient adapter structure to learn language-specific information for each

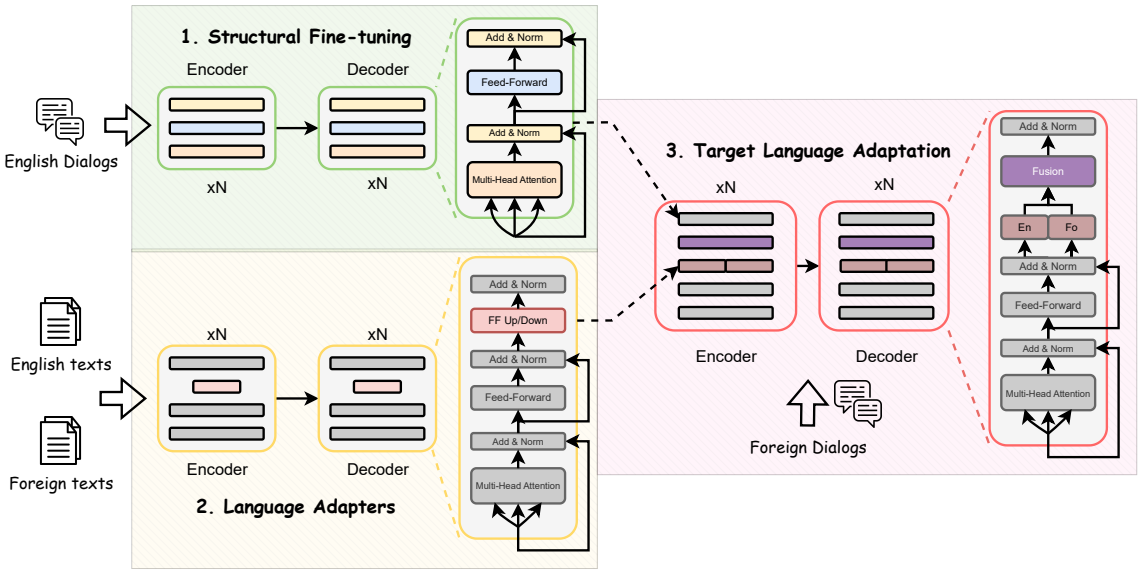

Figure 2: The overview of our proposed cross-lingual transfer framework: XDFusion. We first fine-tune parameters of large pretrained mSeq2seq models with English dialogs to learn syntactic information. Additional language adapters are trained via BART/T5 denoising task while the pretrained multilingual model is kept frozen. Finally, we insert both English (En) and Foreign (Fo) language adapters in the fine-tuned Seq2seq models from Structural Fine-tuning while training the new inserted fusion module only on target language dialogs.

language, independent from the original large fine-tuned model. Each adapter module contains a simple down- and up-projection combined with a residual connection:

$$\texttt{Adapter}_l(\mathbf{h}_l, \mathbf{r}_l) = \mathbf{U}_l(\texttt{ReLU}(\mathbf{D}_l(\mathbf{h}_l))) + \mathbf{r}_l \quad (6)$$

where h_l is the hidden representation of subsequent layer normalization output after feed-forward layer in the transformer layer l, U_l and D_l are up- and down-projection matrices, r_l is the hidden state directly from feed-forward layer. During training, we insert the language adapters into original large pretrained multilingual models and update their parameters only with others kept fixed. However, instead of training language adapters using MLM tasks like Pfeiffer et al. (2020c), to better align the original pretraining objective and learn Seq2seq-fashion language knowledge, we train them on unlabeled data of a language using the BART denoising task.

3.3 Target Language Adaptation

With fine-tuned Seq2seq model from Section 3.1 as well as both source and target language adapters from Section 3.2, we could perform task- and language-specific learning to boost the performance of a specific target language with very few annotations available. To achieve the knowledge sharing between languages, we fix the parameters of large fine-tuned model Θ and source/target language adapters ϕ_s, ϕ_t, we additionally introduce an AdapterFusion module (Pfeiffer et al., 2020a) with parameters Ψ to combine two language adapters with cross attention and facilitate dynamic knowledge allocation to the downstream task by training target language data D_t.

$$\Psi_t \leftarrow \underset{\Psi}{\text{argmin}} \mathcal{L}_t(D_t; \Theta, \phi_s, \phi_t, \Psi) \quad (7)$$

By employing two phases of knowledge extraction and composition, we only train the AdapterFusion layer which averts catastrophic forgetting on task-related knowledge reserving from large fine-tuned models and interference between separate language tasks and target-language adaptation. The use of parameter-efficient structure is language agnostic and seamlessly extendable to other low-resource languages by efficiently training a lightweight target language adapter and a fusion module with easily fetched unlabeled data (bitext pairs are not required). It can allow fast alignment with other languages without much parameter updating.

4 Experimental Settings

4.1 Dataset

We conduct our experiments on *GlobalWOZ dataset* (Ding et al., 2022), a large-scale multilingual ToD dataset globalized from an English-based ToD benchmark: MultiWoZ (Budzianowski et al.,

2018) with four different multilingual use cases, based on the tongue of speakers and countries they travel. We mainly adopt three of all four GlobalWOZ settings: an English speaker in an English country (**E&E**), a Foreign speaker in an English country (**F&E**) and a Foreign speaker in a Foreign country (**F&F**), described in Section 2.1 and Figure 1. There are 10,437 dialogs for each language use in GlobalWOZ. To better compare the observations in GlobalWoZ (Ding et al., 2022) experiments, we follow Ding et al. (2022) to choose English as the high-resource *source* language and other three languages: Chinese (Zh), Spanish (Es), Indonesian (Id) as the low-resource *target (foreign)* languages. In each of four languages, we split 10,437 dialogs into train/validation/test sets with ratio 8:1:1 and we further subsample 100 dialogs from Zh, Es, Id train sets for *few-shot* training. Finally we remain Zh, Es, Id test sets untouched during training and only for testing purpose.

4.2 Baselines

In our first set of experiments, we explore the following *zero-shot* baselines and strategies for training models in Chinese (Zh), Spanish (Es), Indonesian (Id) given a large amount of English training data:

- **E&E**: Fine-tune mSeq2seq with E&E training data only.
- **Translate-Train** (Ding et al., 2022): Translate E&E data with label sequence translation in Ding et al. (2022) using an external machine translation system.
- **Translate-Back**: Directly translate response outputs predicted from English-trained model back into the target language.
- **Adapter** (Pfeiffer et al., 2020c): Insert and fine-tune adapter modules both at encoder and decoder side only.
- **Freeze-Decoder** (Chi et al., 2019): Freeze the decoder part and fine-tune encoder side only.
- **Multi-task learning: NMT & Denoise** (Liu et al., 2020): Include external out-of-domain corpus to perform NMT or Denoising task training simultaneously with the main dialog response generation task.

Then we consider the following *few-shot* baselines by adding a small amount of Zh, Es, Id training data along with English training data.

- **F&F**: Fine-tune mSeq2seq with few F&F training data (100 dialogs) only.
- **E&E + F&F**: Fine-tune Seq2seq model with both E&E and few F&F training data.
- **SPImpMem** (Chen et al., 2019): Insert shared and private memory modules within Seq2seq model to induce cross-lingual transfer.
- **Adapter** (Pfeiffer et al., 2020c): Fine-tune Seq2seq model with E&E training data; then insert and fine-tune adapter modules both at encoder and decoder side only.
- **XDFusion**: Our proposed approach to insert Adapter cross-lingual fusion module which combines pretrained language adapters together.

4.3 Experimental Details

Task Our experiments are mainly conducted on Natural Language Response Generation task (NLG), a critical component in a ToD system to accurately generate relevant system responses given the dialog history and system acts, where large pretrained models serve an ideal purpose.

mBART-50-large-NMT We choose mBART as our base Seq2seq pretrained model with 590M parameters from HuggingFace with a hidden_size = 1,024, which is also first fine-tuned on 50-language translation tasks (mBART-50-large-NMT) (Tang et al., 2021). We then employ the defined data format to train base models in *few-shot* and *zero-shot* setting depicted in Section 2.2.

Evaluation We use sacreBLEU to evaluate the overall n-gram match between generated and ground truth responses and Slot Error Rate (SER) to measure the percentage of correct predicted slots in a generated response.

Implementation details We implement our framework and all baselines within the Transformers (Wolf et al., 2019) and Adapter-Transformers (Pfeiffer et al., 2020b) library. We mainly use mBART (mBART-large-50, mBART-50-large-NMT) and mT5 (mT5-small, mT5-base) for our base pretrained multilingual models. For fine-tuning via mBART denoising task on unlabelled data for language adapters, we train the same amount of **mC4** dataset (Xue et al., 2020) from the public Common Crawl web scrape as GlobalWOZ training data of the corresponding language for 10 epochs, with a batch size of 6 and learning rate $5e-5$. For fine-tuning pretrained models with large training dialog corpus, we train each model for 10 epochs with a batch size of 16 and learning rate $1e-4$. Finally, in *few-shot* training, we train the final model for 60 epochs with the same

Task			NLG							
Metrics			sacreBLEU (%) ↑				SER (%) ↓			
ID	Model	Setting	zh	es	id	avg	zh	es	id	avg
①	E&E (Tang et al., 2021)	Zero-shot	4.44	7.34	11.00	7.59	51.94	33.02	30.97	42.48
②	Translate-Train (Ding et al., 2022)	Zero-shot	10.80	11.81	12.40	11.67	42.32	38.61	37.12	39.35
③	Translate-Back	Zero-shot	14.89	14.10	16.70	15.23	44.19	34.79	29.24	36.07
④	F&F (Tang et al., 2021)	Few-shot	22.56	15.86	20.03	19.48	17.50	30.11	16.91	21.51
⑤	① + ④ (Tang et al., 2021)	Few-shot	22.76	19.47	22.34	21.52	17.31	20.69	17.00	18.33
⑥	⑤ + SPImpMem (Chen et al., 2019)	Few-shot	6.78	8.51	4.31	6.53	78.77	75.31	83.04	79.04
⑦	⑤ + Adapter (Pfeiffer et al., 2020c)	Few-shot	23.82	21.28	23.22	22.77	15.78	21.69	15.62	17.70
⑧	⑦ + Fusion (XDFusion)	Few-shot	**26.71**	**21.39**	**23.78**	**23.96**	**9.76**	**18.46**	**12.51**	**13.58**

Table 1: SacreBLEU and Slot Error Rate (SER) of different cross-lingual methods in NLG task of three target languages. Best scores are highlighted in **bold**. ↑ indicates the higher the better while ↓ indicates the lower the better. avg implies the average result of three languages.

batch size and learning rate. For *zero-shot* baseline (Multi-task NMT), we include **CCMatrix** dataset (Schwenk et al., 2021) for additional NMT training. We choose the best checkpoint for evaluation based on validation performance. We use the Adam optimizer for all parameter optimization. We follow the hidden size of pretrained models with dimensionalities of 512 (mt5-small), 768 (mt5-base), and 1024 (mBART-large-50). We run each experiment with three random seeds and take the average as the results on 8 NVIDIA A100 40GB GPUs.

5 Results & Discussion

5.1 Main Results

In Table 1, we demonstrate the main results of cross-lingual transfer capability by fine-tuning mBART on the GlobalWOZ response generation task. The inferior performance of multilingual mBART in *zero-shot* setting ① reflects the *off-target* problem where generated outputs are undesirably code-switched and missing accurate slot values. Although *Translate-Train* ameliorates the problem by training models with pseudo-labeled translated data, noisy machine-translated entities without context-aware translation still deteriorates its performance on generating accurate local entities. From ③, we found sacreBLEU increases which alludes that multilingual encoders could implicitly learn to encode language-agnostic representations that are reasonable to decode even the decoder messes up the target language generation.

For *few-shot* setting, we observe that the performance increases significantly if we introduce even a small set of annotated foreign dialogs ④. Co-training with English data directly that transfers English knowledge ⑤ will be more useful for the same Indo-European language family like Spanish. SPImpMem ⑥ does not exhibit its power in disentangling language agnostic/specific information in our case with an extremely imbalanced dataset. The additional private memory is not well-trained with only few foreign dialogs. Eventually, our proposed adapter framework ⑧, beats all above baselines including introducing a single adapter ⑦, by efficiently manipulating denoising-trained adapters to quickly adapt language models to a target language without sacrificing much task-specific knowledge learned in the previous phase. We also found that our approach shows larger improvements in Chi-

Model	sacreBLEU (%) ↑				SER (%) ↓			
	zh	es	id	avg	zh	es	id	avg
mT5-small	15.4	7.7	8.7	10.6	38.9	51.4	56.2	48.8
w/ XDFusion	19.1	9.8	11.8	13.6	25.4	39.8	42.9	36.0
mT5-base	13.0	11.2	14.2	12.8	37.6	36.7	33.9	36.1
w/ XDFusion	23.0	14.1	16.5	17.9	14.2	29.5	28.0	23.9
mBART-50-large	24.6	17.9	22.4	21.6	13.6	17.6	14.7	15.3
w/ XDFusion	**26.8**	20.0	**24.0**	23.6	9.8	**17.5**	12.8	**13.3**
mBART-50-large-NMT	22.8	19.5	22.3	21.5	17.3	20.7	17.0	18.3
w/ XDFusion	26.7	**21.4**	23.8	**24.0**	**9.8**	18.5	**12.5**	13.6

Table 2: Comparison of using different pretrained models for F&F testing dataset in three languages. Best scores are highlighted in **bold**.

Model	sacreBLEU (%) ↑				SER (%) ↓			
	zh	es	id	avg	zh	es	id	avg
mT5-small	5.6	7.6	7.6	6.9	59.5	47.4	45.6	50.8
w/ XDFusion	8.9	9.8	7.7	8.8	41.9	38.5	40.6	40.3
mT5-base	9.2	11.4	8.2	9.6	46.4	29.8	51.0	42.4
w/ XDFusion	14.5	15.4	16.0	15.3	35.4	31.9	25.3	30.9
mBART-50-large	18.8	15.0	21.8	18.5	24.2	26.9	**7.5**	19.5
w/ XDFusion	**21.6**	**22.4**	23.4	**22.5**	23.3	16.3	14.0	17.9
mBART-50-large-NMT	17.8	20.2	22.8	20.2	24.7	13.1	8.3	15.4
w/ XDFusion	16.3	21.0	**23.5**	20.3	**21.1**	**11.3**	10.4	**14.3**

Table 3: Comparison of using different pretrained models for F&E testing dataset in three target languages. Best scores are highlighted in **bold**.

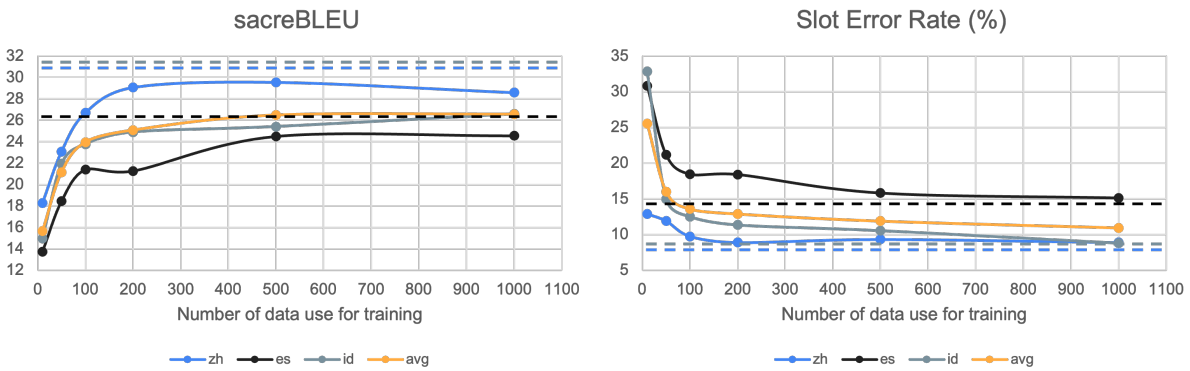

Figure 3: Performance difference of varying available foreign data amount for training. Dashed lines are the results of using all available foreign dialogs in GlobalWOZ of a target language, which are considered as the upper bound.

nese data, which indicates our treatment in disentangling structure and language learning is more important when source and target languages linguistically share less common.

5.2 Comparison to pretrained language models

Table 2 and 3 summarize the results of our proposed framework performance with different base models against the baseline in ⑤ of Table 1. Non-surprisingly, using mT5-small with fewer parameters have limited capacity to learn complicated structures of the fine-tuning task which leads to unsatisfying results. Interestingly, using mBART is more effective than mT5-base while they have similar amount of parameters. We conjecture that the use of special language tokens in mBART may induce better model awareness of language-specific knowledge in *few-shot* setting. The effectiveness of Pretraining mBART with the machine translation task has alternative trends in three languages which may conclude that it will highly depend on the domain intimacy between machine translation corpus and downstream dialogs. For **F&E** setting, overall we have poorer sacreBLEU (code-switched response quality) and SER (predicting English entities) than **F&F** setting where we could deduce that code-switched phenomena make the models harder to generalize between two languages especially with extremely imbalanced datasets. However, we still observe a larger improvement by adopting our proposed framework in **F&E** setting.

5.3 Further Analyses

Data variation. In Figure 3, we vary the number of foreign dialogs to train in the final phase of language adaption. We observe each language saturates around 1k dialogs where the dashed lines are the upper bound of performance when we include all foreign dialogs for training. It demonstrates a good *few-shot* performer of our model by fine-tuning the adapter fusion module that could enhance the overall performance in the low-resource language setting.

Model	sacreBLEU (%) ↑				SER (%) ↓			
	zh	es	id	avg	zh	es	id	avg
Baseline	4.4	7.3	11.0	7.6	51.9	33.0	31.0	42.5
Adapter	2.3	5.5	8.4	5.4	54.7	39.3	42.1	45.3
Freeze-Decoder	7.1	6.5	9.5	7.7	44.3	35.6	30.3	36.8
NMT	1.7	5.4	9.9	5.7	54.1	40.5	34.1	42.9
Denoise	7.2	6.8	9.4	7.8	**44.2**	**33.2**	**28.3**	**35.2**
Translate-Back	**14.9**	**14.1**	**16.7**	**15.2**	44.2	34.8	29.2	36.1

Table 4: Comparison of using different *zero-shot* approaches for F&F testing dataset in three target languages.

Zero-shot observation. Table 4 summarizes our extended experiment results of mBART-50-large-NMT performance on *zero-shot* transfer to understand the effects of some common techniques. Directly applying English-trained model to testing low-resource languages has the lowest BLEU and SER. Since the problem mainly rises from the catastrophic forgetting on the decoder side, we have proposed different additional approaches to mitigate. However, we found neither of them work better except a slight increase in decoder freezing and denoising (row 3 and 5). Special input-output formatting seems to require more efforts for adapters to digest and transform where limiting parameter updates will restrict such capability and still suffer from *off-target* problem. We then focus our contributions in unfreezing our limited budget where extremely few foreign dialog annotation is acceptable.

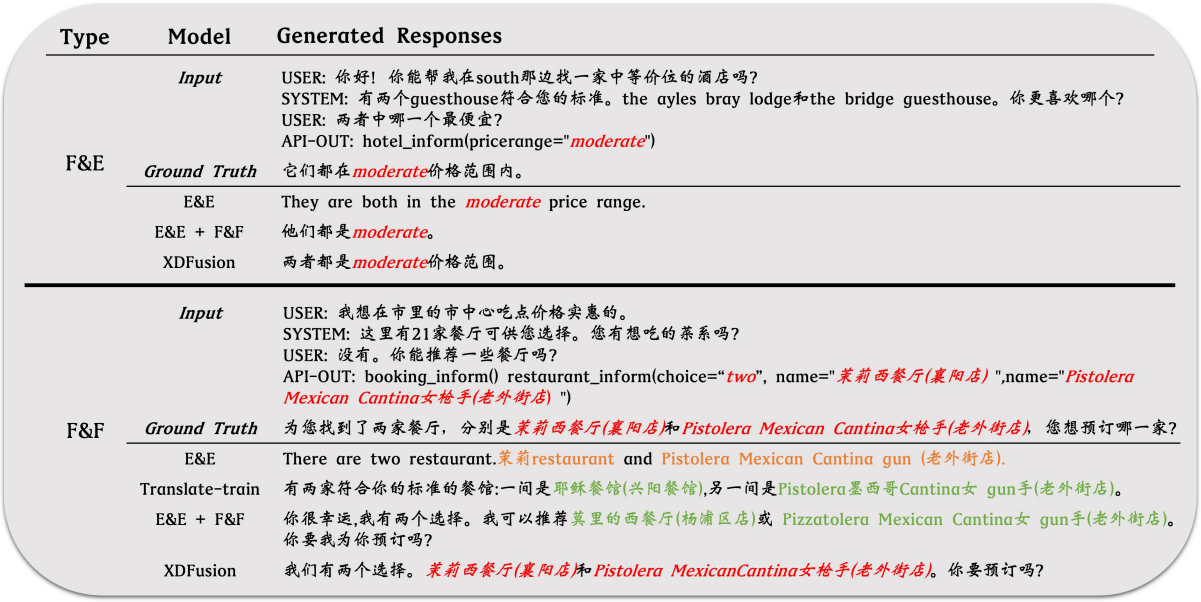

Figure 4: Examples of generated system responses from different models, along with its corresponding input and ground truth responses. The first example is sampled from MultiWOZ F&E Chinese (Zh) test dataset and the other is from F&F. Red words indicate the correct local entities (F&E has English entities; F&F has Foreign entities). Orange words indicate wrong code-switched responses. Green words indicate wrong foreign entities.

Qualitative analysis. Figure 4 shows the generated response examples from different models on the **F&E** and **F&F** test sets. We first observe that if we directly employ mSeq2seq models trained with only **E&E** data to low-resource language tasks, we can see the off-target problem causes models to generate English responses where the target language indicator is omitted. Instead, *Translate-Train* method generates Chinese correctly except the entities are erroneous due to wrong-translated entities from model supervision. Both XDFusion and the few-shot baseline (**E&E + F&F**) generate reasonable responses that correctly follow the given system acts. The results further elucidate XDFusion's high flexibility to generalize to new target language with very limited training data, by generating more fruitful responses with consistent local entities.

6 Related Work

Response generation is one of critical components in ToD systems. Extensive works have proposed to enhance response quality with RNNs (Wen et al., 2015), large pretrained models (Zhang et al., 2020b; Peng et al., 2020), augmentation (Xu et al., 2021) or new learning objectives (Mi et al., 2019; Zhu, 2020). They are either dealing with monolingual data or still require large amounts of annotated data which cannot allow *few-shot* foreign language generation – a vast majority of existing multilingual systems mostly consider language-agnostic task outputs like semantic parsing or ignore real code-switched sentences in real cases (Ding et al., 2022). Instead, DeltaLM (Ma et al., 2021) pretrains interleaved multilingual decoders for text summarization and question generation and CSRL (Wu et al., 2022) learns language-agnostic structure-aware representations for semantic role labeling. Often, due to the high cost of collecting low-resource task-oriented dialog annotations, data-based (Yi and Cheng, 2021; Xiang et al., 2021; Li et al., 2021) and model-based transfer approaches (Schuster et al., 2019; Colombo et al., 2021) are popular to take advantage of high-resource language corpus for cross-lingual transfer. Nevertheless, *few-shot* response generation is yet largely unexplored to induce cross-lingual transfer. The most related prior work is Chen et al. (2019) which extends the Seq2seq models for response generation with private and local memory to accommodate new languages, which nevertheless cannot learn good memory modules when language data is highly imbalanced. Our work continues to explore the possibility of cross-lingual response generation with large Seq2seq models under low-resource language constraint more effectively.

7 Conclusion

In this paper, we explore the pretrained mSeq2seq's capability to induce high-resource language dialog knowledge for low-resource language response generation. By introducing a few foreign high-quality annotated dialogs, we observe that it is possible to learn a dynamic adapter fusion module to fuse all related knowledge in a single large multilingual model, while preserving multilingual power from high-resource language fine-tuning. We have shown that by fine-tuning on very few dialogs of a target language, our proposed model-agnostic framework is capable of producing reasonable responses and more effective than several common baselines, which could quickly adapt to a new target language without further parameter.

Limitations

While we observe marked improvements in the proposed multilingual language transfer with adapters, we recognize that there are several limitations still in the experiments. The first limitation is that the entity translation remains difficult, which is especially severe in the generated responses in the **E&E** setting. We think that name-entity translation is itself a task to be explored in-depth for future works. On the other hand, we think that while knowledge of language is one aspect for the transfer, the structural information of the semantic representation is also another important aspect – models need to acquire the important semantic structural information on top of the language-specific syntactic information. We think that this would further improve the resulting performance.

Ethics Statement

We recognize and take seriously the ethical principles of avoiding harm, trustworthiness, fairness and non-discrimination, and privacy. We take steps to minimize the potential negative impacts of our research and we are committed to ensuring that the use of our findings and technology is done in an ethical and responsible manner. We are committed to ensuring that our research and the use of machine translation technology do not perpetuate language biases, discrimination or any form of inequality.

References

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

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

Chen Chen, Lisong Qiu, Zhenxin Fu, Dongyan Zhao, Junfei Liu, and Rui Yan. 2019. Multilingual dialogue generation with shared-private memory.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Code-switched inspired losses for spoken dialog representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8320–8337, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci. 2022. CrossAligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4048–4061, Dublin, Ireland. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Bing Li, Yujie He, and Wenjin Xu. 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

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.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *ArXiv*, abs/2106.13736.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3151–3157. International Joint Conferences on Artificial Intelligence Organization.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterfusion: Non-destructive task composition for transfer learning.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020b. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020c. 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.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Iulian Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *ArXiv*, abs/1512.05742.

Akshat Shrivastava, Pierce Chuang, Arun Babu, Shrey Desai, Abhinav Arora, Alexander Zotov, and Ahmed Aly. 2021. Span pointer networks for non-autoregressive task-oriented semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1873–1886, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Han Wu, Haochen Tan, Kun Xu, Shuqi Liu, Lianwei Wu, and Linqi Song. 2022. Zero-shot cross-lingual conversational semantic role labeling. *ArXiv*, abs/2204.04914.

Lu Xiang, Yang Zhao, Junnan Zhu, Yu Zhou, and Chengqing Zong. 2021. Zero-shot deployment for cross-lingual dialogue system. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II*, page 193–205, Berlin, Heidelberg. Springer-Verlag.

Xinnuo Xu, Guoyin Wang, Young-Bum Kim, and Sungjin Lee. 2021. Augnlg: Few-shot natural language generation using self-trained data augmentation. In *ACL-IJCNLP 2021*.

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

Huixiong Yi and Jin Cheng. 2021. Zero-shot entity recognition via multi-source projection and unlabeled data. *IOP Conference Series: Earth and Environmental Science*, 693(1):012084.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, Minlie Huang, and Xiaoyan Zhu. 2020c. Recent advances and challenges in task-oriented dialog system.

Chenguang Zhu. 2020. Boosting naturalness of language in task-oriented dialogues via adversarial training. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 265–271, 1st virtual meeting. Association for Computational Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.