XPersona: Evaluating Multilingual Personalized Chatbot

Zhaojiang Lin*, **Zihan Liu***, **Genta Indra Winata***, **Samuel Cahyawijaya***, **Andrea Madotto***, **Yejin Bang, Etsuko Ishii, Pascale Fung**
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{zlinao,zliucr,giwinata,scahyawijaya,amadotto}@connect.ust.hk

Abstract

Personalized dialogue systems are an essential step toward better human-machine interaction. Existing personalized dialogue agents rely on properly designed conversational datasets, which are mostly monolingual (e.g., English), which greatly limits the usage of conversational agents in other languages. In this paper, we propose a multi-lingual extension of Persona-Chat (Zhang et al., 2018), namely XPersona. Our dataset includes persona conversations in six different languages other than English for evaluating multilingual personalized agents. We experiment with both multilingual and cross-lingual trained baselines, and evaluate them against monolingual and translation-pipeline models using both automatic and human evaluation. Experimental results show that the multilingual trained models outperform the translation-pipeline and that they are on par with the monolingual models, with the advantage of having a single model across multiple languages. On the other hand, the state-of-the-art cross-lingual trained models achieve inferior performance to the other models, showing that cross-lingual conversation modeling is a challenging task. We hope that our dataset and baselines [1] will accelerate research in multilingual dialogue systems.

1 Introduction

Personalized dialogue agents have been shown efficient in conducting human-like conversation. This progress has been catalyzed thanks to existing conversational dataset such as Persona-chat (Zhang et al., 2018; Dinan et al., 2019a). However, the training data are provided in a single language (e.g., English), and thus the resulting systems can perform conversations only in the training language. Commercial dialogue systems are required to handle a large number of languages since the smart home devices market is increasingly international (Etherington, 2019). Therefore, creating multilingual conversational benchmarks is essential, yet challenging since it is costly to perform human annotation of data in all languages.

A possible solution is to use translation systems before and after the model inference. This comes with three major problems: 1) amplification of translation errors since the current dialogue systems are far from perfect, especially with noisy input; 2) the three-stage pipeline system is significantly slower in terms of inference speed; and 3) high translation costs since the current state-of-the-art models, especially in low resources languages, are only available using costly APIs.

In this paper, we analyze two possible workarounds to alleviate the aforementioned challenges. The first is to build a cross-lingual transferable system by aligning cross-lingual representations, as in Conneau et al. (2018), in which the system is trained on one language and zero-shot to another language. The second is to learn a multilingual system directly from noisy multilingual data (e.g., translated data), thus getting rid of the translation system dependence at inference time.

To evaluate the aforementioned solutions, we propose a dataset called Multilingual Persona-Chat, or XPersona, by extending the Persona-Chat corpora (Dinan et al., 2019a) to six languages: Chinese, French, Indonesian, Italian, Korean, and Japanese. In XPersona, the training sets are automatically translated using translation APIs with several human-in-the-loop passes of mistake correction. In contrast, the validation and test sets are annotated by humans to facilitate both automatic and human evaluations in multiple languages.

Furthermore, we propose competitive baselines in two training settings, namely, cross-lingual and multilingual, and compare them with translation pipeline models. Our baselines leverage pre-

* Equal contributions. Listing order is random

[1] The code and dataset are available in https://github.com/HLTCHKUST/Xpersona.

Proceedings of the Third Workshop on Natural Language Processing for Conversational AI, pages 102–112
November 10, 2021.

Persona	I helped design the game starcraft. I am a famous twitch streamer..
En	**Sys:** do you know the game starcraft ? i helped designing it ! **Usr:** that s awesome , i m sure my ex boyfriend would like that game **Sys:** i usually stream the game on my twitch channel that s great , i also love twitch as well
It	**Sys:** conosci il gioco starcraft? ho aiutato a progettarlo! **Usr:** è fantastico, sono sicuro che al mio ex ragazzo gli piacerebbe **Sys:** di solito faccio streaming mentre gioco sul mio canale Twitch
Fr	**Sys:** connaissez-vous le jeu starcraft? j'ai aidé à le concevoir! **Usr:** c'est génial, je suis sûr que mon ex petit ami aimerait ce jeu **Sys:** Je diffuse généralement le jeu sur ma chaîne Twitch
Id	**Sys:** apakah anda tahu game starcraft? saya yang mendesainnya! **Usr:** itu luar biasa, saya yakin mantan pacar saya suka game itu. **Sys:** saya biasanya memainkan game itu di channel twitch saya.
Zh	**Sys:** 你知道游戏《星际争霸》吗？我帮忙设计了它！ **Usr:** 好厉害，我觉得我的前男友会喜欢那个游戏 **Sys:** 我经常在我的直播频道上直播游戏
Ko	**Sys:** 너 게임 스타크래프트를 아니? 나는 그것을 디자인하는 것을 도왔어! **Usr:** 멋진데, 내 전 남자친구가 그 게임을 좋아할 거라고 확신해. **Sys:** 나는 보통 내 트위치 채널로 그 게임을 스트리밍해.
Jp	**Sys:** ゲームのスタークラフトを知っていますか？私はそれを設計するのを助けました！ **Usr:** それはすごいです、私は私の元彼がそのゲームを好きになると確信しています **Sys:** 私は通常、twitchチャンネルでゲームをストリーミングします

Table 1: Multi-turn annotated dialogue samples from test set in seven languages. For simplicity, we only show three turns for each dialogue and the persona in English.

trained cross-lingual (Chi et al., 2019) and multilingual (Devlin et al., 2018) models.

An extensive automatic and human evaluation (Li et al., 2019) of our models shows that a multilingual system is able to outperform strong translation-based models and on par with or even improve the monolingual model. The cross-lingual performance is still lower than other models, which indicates that cross-lingual conversation modeling is very challenging. The main contributions of this paper are summarized as follows:

- We present the first multilingual non-goal-oriented dialogue benchmark for evaluating multilingual generative chatbots.
- We provide both cross-lingual and multilingual baselines and discuss their limitations to inspire future research.
- We show the potential of multilingual systems to understand the mixed language dialogue context and generate coherent responses.

2 Related Work

Dialogue Systems are categorized as goal-oriented and chit-chat. Interested readers may refer to Gao et al. (2018) for a general overview. In this paper, we focus on the latter, for which, in recent years, several tasks and datasets have been proposed to ground the conversation on knowledge (Dinan et al., 2019b; Gopalakrishnan et al., 2019; Fan et al., 2019; Reddy et al., 2019; Moon et al., 2019) such as Wiki-Articles, Reddit-Post, and CNN-Article. In this work, we focus on personalized dialogue agents where the dialogues are grounded on persona information.

Li et al. (2016a) was the first to introduce a persona-grounded dialogue dataset for improving response consistency. Later on, Zhang et al. (2018) and Dinan et al. (2019a) introduced Persona-chat, a multi-turn conversational dataset, where two speakers are paired, and a persona description (4–5 sentences) is randomly assigned to each of them. By conditioning the response generation on the persona descriptions, a chit-chat model is able to produce a more persona-consistent dialogue (Zhang et al., 2018). Several works have improved on the initial baselines with various methodologies, especially using large pre-trained models (Wolf et al., 2019).

Multilingual Extensive approaches have been introduced to construct multilingual systems, for example, multilingual semantic role labeling (Akbik et al., 2015), multilingual machine trans-

	Valid.				Test			
Lang	***#Dial.***	***#Utt.***	***Edit***	***BLEU***	***#Dial.***	***#Utt.***	***Edit***	***BLEU***
Fr	248	3868	21.23	94.45	249	3900	24.29	94.19
It	140	2160	83.01	80.45	140	2192	81.93	80.08
Id	484	7562	157.58	60.46	484	7540	156.19	60.66
Jp	275	4278	71.41	53.66	275	4322	75.83	49.56
Ko	299	4684	74.04	61.25	300	4678	70.96	62.49
Zh	222	3440	30.33	59.89	222	3458	33.07	64.61

Table 2: The statistics of the collected dataset. We report the number of dialogues (#Dial.) and utterances (#Utt.) of the validation and test set in six languages. Edit distance per dialogue (Edit) and BLEU score are computed to show the difference between the human-annotated dataset and auto-translated dataset (Training set is reported in Appendix A). The BLEU score also reflects the quality of machine translated dialogues.

lation (Johnson et al., 2017), and multilingual automatic speech recognition (Toshniwal et al., 2018). Multilingual deep contextualized model, such as Multilingual BERT (M-BERT) (Devlin et al., 2018), MT5 (Xue et al., 2021), MBART (Liu et al., 2020) have been commonly used to represent multiple languages and elevate the performance in many NLP applications, such as classification tasks (Pires et al., 2019), textual entailment, named entity recognition (K et al., 2020), and natural language understanding. Multilingual datasets have also been created for a number of NLP tasks, such as named entity recognition or linking (Pan et al., 2017; Aguilar et al., 2018), question answering (Liu et al., 2019; Lewis et al., 2019), dialogue state tracking (Mrkšić et al., 2017), and natural language understanding (Schuster et al., 2019). However, none of these datasets include the multilingual chit-chat task.

Cross-lingual Cross-lingual adaptation learns the inter-connections among languages and circumvents the requirement of extensive training data in target languages (Wisniewski et al., 2014; Zhang et al., 2016). Cross-lingual transfer learning methods have been applied to multiple NLP tasks, such as named entity recognition (Ni et al., 2017), dialogue state tracking (Chen et al., 2018), part-of-speech tagging (Wisniewski et al., 2014; Zhang et al., 2016; Kim et al., 2017), and dependency parsing (Ahmad et al., 2019). Meanwhile, Lample and Conneau (2019) and Conneau et al. (2019) proposed pre-trained cross-lingual language models to align multiple language representations, achieving state-of-the-art results in many cross-lingual classification tasks. The aforementioned tasks focused on classification and sequence labeling, while instead, Chi et al. (2019) proposed to pre-train both the encoder and decoder of a sequence-to-sequence model (XNLG) to conduct cross-lingual generation tasks, namely, question generation and abstractive summarization. The latter is the closest to our task since it focuses on language generation; however cross-lingual dialogue generation has not yet been explored.

3 Data Collection

The proposed XPersona dataset is an extension of the persona-chat dataset (Zhang et al., 2018; Dinan et al., 2019a). Specifically, we extend ConvAI2 (Dinan et al., 2019a) to six languages: Chinese, French, Indonesian, Italian, Korean, and Japanese. Since the test set of ConvAI2 is hidden, we split the original validation set into a new validation set and test sets. Then, we firstly automatically translate the training, validation, and test set using APIs (PapaGo for Korean, Google Translate for other languages). For each language, we hired native speaker annotators with at least a bachelor degree and a fluent level of English and asked them to revise the machine-translated dialogues and persona sentences in the validation set and test set according to original English dialogues. The main goal of human annotation is to ensure the revised conversations are coherent and fluent in target language despite the cultural discrepancy in different languages. Therefore, annotators are not restricted to translate the English dialogues. They are also **allowed** to customize dialogues and persona sentences. The annotated dialogues can deviate from original translation while **retain persona and conversation consistency**. The full annotation instructions are reported in Appendix A.

Compared to collecting new persona sentences and dialogues in each language, human-annotating

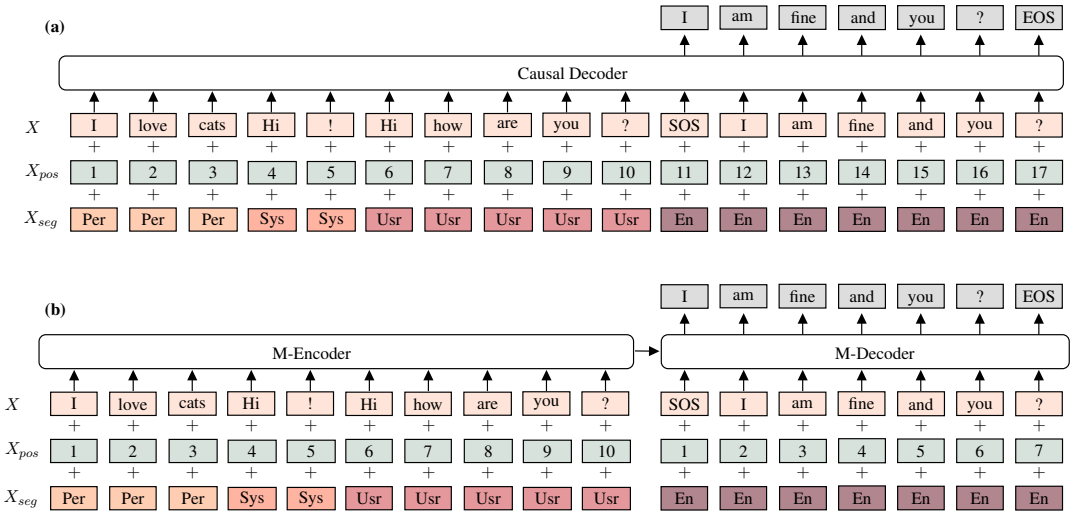

Figure 1: (a) Multilingual Causal Decoder model. (b) Multilingual Encoder-Decoder model.

the dialogues by leveraging translation APIs has multiple advantages. First, it increases the data distribution similarity across languages (Conneau et al., 2018), which can better examine the system's cross-lingual transferability. Second, revising the machine-translated dialogues based on the original English dialogue improves the data construction efficiency. Third, it leverages the well-constructed English persona conversations as a reference to ensure the dialogue quality without the need for training a new pool of workers to generate new samples (Conneau et al., 2018).

On the other hand, human-translating the entire training-set (∼130K utterances) in six languages is expensive. Therefore, we propose an iterative method to improve the quality of the automatically translated training set. We firstly sample 200 dialogues from the training set (∼2600 utterances) in each language, and we assign human annotators to list all frequent translation mistakes in the given dialogues. For example, daily colloquial English expressions such as "cool", "I see", and "lol" are usually literally translated. After that, we use a simple string matching to revise the inappropriate translations [2] in the whole training-set and return a revision log, which records all the revised utterances. Then, we assign human annotators to check all the revised utterances and list translation mistakes again. We repeat this process at least twice for each language. Finally, we summarize the statistics of the collected dataset in Table 2.

[2]The list of corrections and matching rules are reported in Appendix.

4 Multilingual Personalized Conversational Models

Let us define a dialogue $\mathcal{D} = \{U_1, S_1, U_2, S_2, \ldots, U_n, S_n\}$ as an alternating set of utterances from two speakers, where U and S represent the user and the system, respectively. Each speaker has its corresponding persona description that consists of a set of sentences $\mathcal{P} = \{P_1, \ldots, P_m\}$. Given the system persona sentences $\mathcal{P}_s$, dialogue history $U_{\leq k}$, $S_{<k}$, and response language l, we are interested in predicting the next system utterances S_k with model $f(\theta)$.

$$S_k = f(U_{\leq k}, S_{<k}, l; \theta) \quad (1)$$

4.1 Model Architecture

We explore both encoder-decoder and causal decoder architectures, and we leverage existing pretrained contextualized multilingual language models as weights initialization. Hence, we firstly define the multilingual embedding layer and then the two multilingual models used in our experiments.

Embedding We define three embedding matrices: word embedding $E^W \in \mathbb{R}^{|V| \times d}$, positional embedding $E^P \in \mathbb{R}^{M \times d}$, and segmentation embedding $E^S \in \mathbb{R}^{|S| \times d}$, where $|.|$ denotes set cardinality, d is the embedding size, V denotes the vocabulary, M denotes the maximum sequence length, and S denotes the set of segmentation tokens. Segmentation embedding (Wolf et al., 2019) is used to indicate whether the current token is part of i) **Persona** sentences, ii) System (**Sys.**) utter-

ances, iii) **User** utterances, iv) response in **Language** l. The language embedding l_{id} is used to inform the model which language to generate. Hence, given a sequence of tokens X, the embedding functions E are defined as:

$$E(X) = E^W(X) \oplus E^P(X_{pos}) \oplus E^S(X_{seg}), \quad (2)$$

where $\oplus$ denotes the positional sum, $X_{pos} = \{1, \ldots, |X|\}$ and X_{seg} is the sequence of segmentation tokens, as in Wolf et al. (2019). Figure 1 shows a visual representation of the embedding process. A more detailed illustration is reported in Appendix B.

Encoder-Decoder To model the response generation, we use a Transformer (Vaswani et al., 2017) based encoder-decoder (Vinyals and Le, 2015). As illustrated in Figure 1, we concatenate [3] the system persona $\mathcal{P}_s$ with the dialogue history $U_{\leq k}$, $S_{<k}$. Then we use the embedding layer E to finally pass it to the encoder. In short, we have:

$$H = \text{Encoder}(E([\mathcal{P}_s, U_{\leq k}, S_{<k}])), \quad (3)$$

where $H \in \mathbb{R}^{L \times d_{model}}$ is the hidden representation computed by the encoder, and L denotes the input sequence length. Then, the decoder attends to H and generates the system response S_k token by token. In the decoder, segmentation embedding is the language ID embedding (e.g., we look up the embedding for Italian to decode Italian). Thus:

$$S_k = \text{Decoder}(H, l), \quad (4)$$

Causal Decoder As an alternative to encoder-decoders, the causal-decoders (Radford et al., 2018, 2019; He et al., 2018) have been used to model conversational responses (Wolf et al., 2019; Zhang et al., 2019) by giving as a prefix the dialogue history. In our model, we concatenate the persona $\mathcal{P}_s$ and the dialogue history $U_{\leq k}$, $S_{<k}$ as the language model prefix, and autoregressively decode the system response S_k based on language embedding:

$$S_k = \text{Decoder}(E([\mathcal{P}_s, U_{\leq k}, S_{<k}]), l). \quad (5)$$

Figure 1 shows the conceptual differences between the encoder-decoder and casual decoder. Note that in both multilingual models, the dialogue history encoding process is language-agnostic, while decoding language is controlled by the language embedding. Such design allows the model to understand mixed-language dialogue contexts and to respond in the desired language (details in Section 5.3.2).

[3] $[a; b]$ denotes concatenating the vectors a and b

4.2 Training Strategy

We consider two training strategies to learn a multilingual conversational model: multilingual training and cross-lingual training.

Multilingual Training jointly learns personalized conversations in multiple languages. We follow a transfer learning approach (Wolf et al., 2019) by initializing our models with the weights of the large multilingual pretrained model M-Bert (Pires et al., 2019). For the causal decoder, we add the causal mask into self-attention layer to convert M-Bert encoder to decoder. For encoder-decoder model, we randomly initialize the cross encoder-decoder attention (Rothe et al., 2019). Then, we train the both models on the combined training set in all 7 languages using cross-entropy loss.

Cross-lingual Training transfers knowledge from the source language data to the target languages. In this setting, the model is trained on English (source language) conversational samples, and evaluated on the other 6 languages. Following the methodology proposed by Chi et al. (2019), we align the embedded representations of different languages into the same embedding space by applying cross-lingual pre-training to the encoder-decoder model. The pre-training procedure consists of two stages:

- pre-training the encoder and the decoder independently utilizing masked language modeling, as in Lample and Conneau (2019);
- jointly pre-training the encoder-decoder by using two objective functions: Cross-Lingual Auto-Encoding (XAE) and Denoising Auto-Encoding (DAE) (Chi et al., 2019).

For instance, DAE adds perturbations to the input sentence of encoder and tries to reconstructs the original sentence using the decoder, whereas, XAE uses parallel translation data to pre-train both the encoder and decoder with machine translation objective. As in the multilingual models, the language IDs are fed into the decoder to control the language of generated sentences. Both pre-training stages require both parallel and non-parallel data in the target language.

	Bert2Bert		M-Bert2Bert		CausalBert		M-CausalBert		XNLG	
	ppl.	***BLEU***	***ppl.***	***BLEU***	***ppl.***	***BLEU***	***ppl.***	***BLEU***	***ppl.***	***BLEU***
En	21.99	1.53	25.99	0.57	16.08	1.79	**15.62**	1.97	54.74*	**2.25***
Zh	21.35	3.36	13.24	1.25	**8.69**	5.51	9.27	**5.7**	3482.27	2.16
It	50.36	0.6	24.16	0.31	18.41	1.32	**15.12**	**1.3**	917.63	0.41
Jp	10.09	5.23	10.64	0.79	11.00	**6.74**	**7.13**	4.53	999.81	0.0
Ko	12.81	0.24	34.31	0.00	9.66	1.06	**9.56**	**1.08**	331.23	0.32
Id	21.37	0.11	22.83	0.22	14.77	**2.1**	**14.61**	1.92	844.98	0.15
Fr	13.22	0.35	15.58	0.50	**10.39**	1.97	10.59	**2.17**	640.33	0.09

Table 3: Results of automatic evaluation score on test set in seven languages. We compute the BLEU score and perplexity (ppl.) for monolingual, multilingual, and cross-lingual models.

After the two stages of pre-training, the model is fine-tuned using just the source language samples (i.e., English) with the same cross-entropy loss as for the multilingual training. However, as suggested in Chi et al. (2019), only the encoder parameters are updated with back-propagation and both the decoder and the word embedding layer remain frozen. This retains the decoders' ability to generate multilingual output while still being able to learn new tasks using only the target language.

5 Experiments

5.1 Evaluation Metrics

Evaluating open-domain chit-chat models is challenging, especially in multiple languages and at the dialogue-level. Hence, we evaluate our models using both automatic and human evaluation. In both cases, human-annotated dialogues are used, which show the importance of the provided dataset.

Automatic For each language, we evaluate responses generated by the models using perplexity (ppl.) and BLEU (Papineni et al., 2002) with reference to the human-annotated responses. Although these automatic measures are not perfect (Liu et al., 2016), they help to roughly estimate the performance of different models under the same test set. More recently, Adiwardana et al. (2020) has shown the correlation between perplexity and human judgment in open-domain chit-chat models.

Human Asking humans to evaluate the quality of a dialogue model is challenging, especially when multiple models have to be compared. The likert score (a.k.a. 1 to 5 scoring) has been widely used to evaluate the interactive experience with conversational models (Venkatesh et al., 2018; See et al., 2019; Zhang et al., 2018; Dinan et al., 2019a). In such evaluation, a human interacts with the systems for several turns, and then they assign a score from 1 to 5 based on three questions (Zhang et al., 2018) about fluency, engagingness, and consistency. This evaluation is both expensive to conduct and requires many samples to achieve statistically significant results (Li et al., 2019). To cope with these issues, Li et al. (2019) proposed ACUTE-EVAL, an evaluation for dialogue systems. The authors proposed two modes: human-model chats and self-chat (Li et al., 2016b; Ghandeharioun et al., 2019). In this work, we opt for the latter since it is cheaper to conduct and achieves similar results (Li et al., 2019) to the former. Another advantage of using this method is the ability to evaluate multi-turn conversations instead of single-turn responses.

Following ACUTE-EVAL, the annotator is provided with two full dialogues made by self-chat or human-dialogue. The annotator is asked to choose which of the two dialogues is better in terms of engagingness, interestingness, and humanness. For each comparison, we sample 60–100 conversations from both models. In Appendix C, we report the exact questions and instructions given to the annotators, and the user interface used in the evaluation. We hired at least 10 annotators for each considered language, the annotators are either native speakers or linguists in corresponding language. The annotators were different from the dataset collection annotators to avoid any possible bias.

5.2 Implementation Details

Multilingual Models We use the "BERT-Base, Multilingual Cased" checkpoint, and we denote the multilingual encoder-decoder model as **M-Bert2Bert** (∼220M parameters) and causal decoder model as **M-CausalBert** (∼110M parameters). We fine-tune both models in the combined

Multi Wins %	Lang	*Engageness*			*Interestingness*			*Humanness*		
		Human	**Mono**	**Poly**	**Human**	**Mono**	**Poly**	**Human**	**Mono**	**Poly**
	En	**23.33**	**68.57**	36.36	**23.33**	**64.29**	**32.73**	**30.00**	**62.86**	42.73
	Fr	32.00	55.17	42.86	**16.00**	53.45	48.21	**28.00**	50.00	44.64
	Id	**21.67**	51.67	**65.45**	**23.33**	46.67	55.45	**25.00**	46.67	**65.45**
	It	**35.00**	48.33	56.36	**30.00**	48.33	53.64	**30.00**	40.00	57.27
	Jp	**18.33**	50.00	**61.82**	**13.33**	43.33	45.45	**18.33**	51.67	59.09
	Ko	**30.00**	52.46	**62.39**	**26.67**	50.82	59.63	**28.33**	52.46	**64.22**
	Zh	**36.67**	55.00	**65.45**	**36.67**	60.00	**61.82**	**36.67**	55.00	**70.91**

Table 4: Results of ACUTE-EVAL human evaluation. Tests are conducted pairwise between M-CausalBert (Multi.) and other models (Human, Poly-encoder (Poly), Monolingual CausalBert (Mono)). **Numbers indicate the winning rate of M-CausalBert**. Numbers in bold are statistically significant ($p < 0.05$).

training set (English in Persona-chat (Zhang et al., 2018), six languages in Xpersona) for five epochs with AdamW optimizer and a learning rate of $6.25e$-5.

Monolingual Models To verify whether the multilingual agent will under-perform the monolingual agent in the monolingual conversational task, we build a monolingual encoder-decoder model and causal decoder model for each language. For a fair comparison, we initialize the monolingual models with a pre-trained monolingual BERT [4] (Devlin et al., 2018; Cui et al., 2019; Martin et al., 2019). We denote the monolingual encoder-decoder model as **Bert2Bert** (∼220M parameters) and causal decoder model as **CausalBert** (∼110M parameters). Then we fine-tune each model in each language independently for the same number of epoch and optimizer as the multilingual model. Our **CausalBert** model achieve 16.08 perplexity, which is similar to 17.51 of the GPT based models (Wolf et al., 2019).

Translation-based Models Another strong baseline we compare with is Poly-encoder (Humeau et al., 2019), a large-scale pre-trained retrieval model that **fine-tuned** on English Persona-chat, has shown state-of-the-art performance in the ConvAI dataset (Li et al., 2019; Humeau et al., 2019). We adapt this model to the other languages by using the Google Translate API to translate target languages (e.g., Chinese) query to English as the input to the model, then translate the English response back to the target language. Thus, the response generation flow is: target query → English query → English response → target response. We denote this model as **Poly**.

[4]The monolingual BERT pre-trained models are available in https://github.com/huggingface/transformers

Cross-lingual Models. In the first pre-training stage, we use the pre-trained weights from XLMR-base (Conneau et al., 2019). Then, we follow the second pre-training stage of XNLG (Chi et al., 2019) for pre-training Italian, Japanese, Korean, Indonesia cross-lingual transferable models. For Chinese and French, we directly apply the pre-trained XNLG (Chi et al., 2019) weights[5]. Then, the pre-trained models are fine-tune on English PersonaChat training set and early stop based on the perplexity on target language validation set.

5.3 Results and Discussion

5.3.1 Quantitative Analysis

Table 3 compares monolingual, multilingual, and cross-lingual models in terms of BLEU and perplexity in the human-translated test set. On both evaluation matrices, the causal decoder models outperform the encoder-decoder models. We observe that the encoder-decoder model tends to overlook dialogue context and generate digressive responses. (Generated samples are available in Appendix D) We hypothesize that this is because the one-to-many problem (Zhao et al., 2017) in open-domain conversation weakens the relation between encoder and decoder; thus the well pre-trained decoder (Bert) easily converges to a local optimum, and learns to ignore the dialogue context from the encoder and generate the response in an unconditional language model way. We leave the investigation of this problem to future work. On the other hand, M-CausalBert achieves a comparable or slightly better performance compared to CausalBert, which suggests that M-CausalBert leverages the data from other languages. As expected, we observe a significant gap between the cross-lingual model and

[5]Available in https://github.com/CZWin32768/XNLG

other models, which indicates that cross-lingual zero-shot conversation modeling is very challenging.

Table 4 shows the human evaluation result of comparing M-CausalBert (Multi) against the human, translation-based Poly-encoder (Poly), and monolingual CausalBert (Mono). The results illustrate that Multi outperforms Mono in English and Chinese, and is on par with Mono in other languages. On the other hand, Poly shows a strong performance in English as it was pre-trained with a large-scale English conversation corpus. In contrast, the performance of Poly drops in other languages, which indicates that the imperfect translation affects translation-based systems. We also conduct M-CausalBert against XNLG (cross) human evaluation, and Multi achieve nearly 100 percent winning rate.

5.3.2 Qualitative Analysis and Discussion

We randomly sample 7 self-chat dialogues for each baseline model in the seven languages and report them in Appendix D., And we summarize the generation of each model as follows:

Poly Poly-encoder, pretrained on 174 million Reddit data and fine tuned on English Persona-Chat, can accurately retrieve coherent and diverse responses in English. However, in the other six languages, some of the retrieved responses are digressive due to translation errors.

Monolingual & Multilingual We observe that both the monolingual and multilingual models can generate fluent responses. Compared to Bert2Bert and M-Bert2Bert, CausalBert and M-CausalBert can generate more on-topic responses but sometimes repeat through turns. CausalBert and M-CausalBert are on par with each other in monolingual conversational tasks, while M-CausalBert shows the advantage of handling a mixed-language context. For multilingual speakers, the conversation may involve multiple languages. Therefore, we test M-CausalBert with extra-Sentential code-switching context. Table 5 shows that M-CausalBert can understand the mixed-language context, and decode coherent responses. In Appendix C.1, we report more generation examples, and we also show that the response language of M-CausalBert can be control by language embeddings.

System Persona	
	My favorite activity is rock climbing. I'm a mechanical engineer. I like to go hiking in the spring. I live in colorado.
Dialogue	
Usr	studio inglese oggi e tu? (I am studying English today, you?)
Sys	just watching netflix with friends
Usr	我喜欢攀岩，你呢？ (I like rock climbing, what about you?)
Sys	nice, I enjoy it while in the summer. what are you up to?
Usr	saya ingin melakukan panjat tebing (I want to go to hike)
Sys	I'm going hiking the most time of year
Usr	スポーツがお好きなようですね (It seems you like sports)
Sys	I love the outdoors too, do you have any hobbies?
Usr	나는 티비 보는 것을 좋아해.(I like to watch tv)
Sys	i really like hiking and listening to music

Table 5: M-CausalBert handles extra-Sentential code-switching context without any language identifiers.

Cross-lingual. The current state-of-the-art cross-lingual generation approach XNLG (Chi et al., 2019) shows inferior performance on multi-turn dialogue tasks, and generates repetitive responses. Although cross-lingual dialogue generation is challenging, it reduces the human effort for data annotation in different languages. Therefore, the cross-language transfer is an important direction to investigate.

6 Conclusion

In this paper, we studied both cross-lingual and multilingual approaches in end-to-end personalized dialogue modeling. We presented the XPersona dataset, a multilingual extension of Persona-Chat, for evaluating the multilingual personalized chatbots. We further provided both cross-lingual and multilingual baselines and compared them with the monolingual approach and two-stage translation approach. Extensive automatic evaluation and human evaluation were conducted to examine the models' performance. The experimental results showed that multilingual trained models, with a single model across multiple languages, can outperform the two-stage translation approach and is on par with monolingual models. On the other hand, the current state-of-the-art cross-lingual approach XNLG achieved

lower performance than other baselines. In future work, we plan to research a more advanced cross-lingual generation approach and construct a mixed-language conversational benchmark for evaluating multilingual systems.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 414–424.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training. *arXiv preprint arXiv:1909.10481*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019a. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Darrell Etherington. 2019. Amazon launches multilingual mode for using alexa in multiple languages at once.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374. ACM.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, pages 13658–13669.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems*, pages 7944–7954.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *CoRR*

abs/1905.01969. External Links: Link Cited by, 2:2–2.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016b. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. Association for Computational Linguistics.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. Xqa: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368.

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

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

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

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1946–1958.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *arXiv preprint arXiv:1907.12461*.

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.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4904–4908. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*, 4:60–68.

Oriol Vinyals and Quoc V Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1779–1785.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.