ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240950

ChatZero: Zero-Shot Cross-Lingual Dialogue Generation via Pseudo-Target Language

Yongkang Liu[1,2], Feng Shi[1], Daling Wang[1,*], Yifei Zhang[1] and Hinrich Schütze[2]

[1]Northeastern University, China
[2]Center for Information and Language Processing and Munich Center for Machine Learning, LMU Munich

Abstract. Although large language models(LLMs) show amazing capabilities, among various exciting applications discovered for LLMs fall short in other low-resource languages. Besides, most existing methods depend on large-scale dialogue corpora and thus building systems for dialogue generation in a zero-shot scenario remains a considerable challenge. To address this challenge, we propose a novel end-to-end zero-shot dialogue generation model ChatZero based on cross-lingual code-switching method. First, we construct code-switching language and pseudo-target language with placeholders. Then for cross-lingual semantic transfer, we employ unsupervised contrastive learning to minimize the semantics gap of the source language, code-switching language, and pseudo-target language that are mutually positive examples in the high dimensional semantic space. Experiments on the multilingual DailyDialog and DSTC7-AVSD datasets demonstrate that ChatZero can achieve more than 90% of the original performance under the zero-shot case compared to supervised learning, and achieve state-of-the-art performance compared with other baselines.

1 Introduction

Open domain dialogue generation techniques have achieved significant progress thanks to the availability of large-scale dialogue datasets. Particularly, the pre-trained large models of dialogue generation can generate informative and fluent responses [30, 26], which have huge potential in various applications such as emotional companionship, mental health support, and social chatbots.

The availability of large-scale datasets is a double-edged sword, which also brings a worrying phenomenon that most existing dialogue systems excessively rely on large-scale dialogue corpus [30, 4]. This phenomenon greatly limits the popularity of dialogue systems due to large-scale corpora unavailable in most cases [10]. For example, there are more than 7,000 languages worldwide, but only about 1% have an available corpus [35]. When it comes to dialogue tasks, there are even fewer languages with an available corpus. Zero-shot dialogue techniques usually utilize non-target language corpus for knowledge transfer, which can significantly alleviate the dependence on the target language corpus [18].

The multilingual code-switching method has been proven to be effective in low- and zero-shot generation in NMT (Neural Machine Translation) [3, 13]. Unfortunately, zero-shot generation methods in NMT are difficult to apply to dialogue tasks. The main reason is that the source and target languages of NMT have the same semantics, while there is no similar semantic phenomenon between dialogue history and response, as their semantics are different.

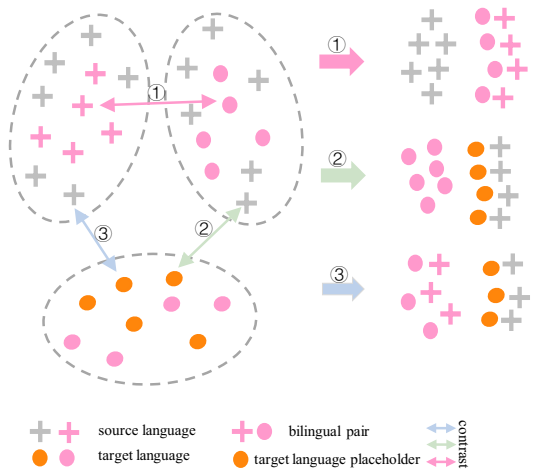

Figure 1. Schematic diagram of ChatZero.

The most existing low-resource dialogue studies pay little attention to the problem of missing corpora. Low-resource knowledge-grounded dialogue generation methods focus on the problem of knowledge deficiency with sufficient corpus [14, 38], which fail to work in zero-shot scenarios. Liu et al. [18] proposes a utterance level code-switching method for zero-shot dialogue generation, which depends on massive translated target language utterances. Their proposed model exposes a large amount of target language corpus in exchange for performance improvement, which, strictly speaking, could not be regarded as a zero-shot generation. Although employing large-scale pre-trained language models, such as GPT-3 [1] and BlenderBot [26], can reduce the dependence on target corpus, these models are usually limited to the language of pre-trained corpus and fail to work well on other languages.

It is known that different language representations of the same utterance are similar in high-dimensional semantic space [11]. Accordingly, we construct a pseudo-target language corresponding to the source language by dictionaries. The pseudo-target language refers to a language that contains target language words and placeholder **[MASK]**. For example, "*Hier [MASK] ein [MASK]*" is pseudo-German language. The main reason is that [MASK] can be considered as an unrevealed token with actual semantics in masked language model, such as mBERT, which is determined by the pre-training tasks. Besides, we build code-switching languages consist-

* Corresponding Author

ing of source and target languages, through bilingual dictionaries to reduce the difficulty of cross-lingual learning. An example of code-switching between English and German is "*Here is ein Beispiel*". In summary, the semantics of the source language, pseudo-target language, and code-switching language of the same utterance are similar in semantic space.

We propose a novel end-to-end zero-shot dialogue generation model ChatZero based on a pseudo-target language. ChatZero employs unsupervised contrastive learning to minimize the representation gap of same utterances in different languages and maximize that of irrelevant utterances. As shown in Figure 1, we pull the semantic representations of the source language, the code-switching language and the pseudo-target language closer for same utterances. At the same time, we push away the utterances that are irrelevant in the same batch. Cross-lingual knowledge transfer is realized by semantic approximation in high-dimensional semantic space, which includes three aspects: (i) aligning the semantics between source and target languages ($\leftrightarrow$ and $\leftrightarrow$); (ii) aligning the semantics of the placeholders with the source language ($\leftrightarrow$ and $\leftrightarrow$); (iii) aligning the semantics of the placeholders to the target language ($\leftrightarrow$). The process (iii) depends on processes (i) and (ii). Because the input in the code-switching form contains limited vocabulary in the target language, (i) can only play a limited role in semantic transfer. The implicit semantic alignment between source and target languages is achieved through (ii) and (iii). Models can adapt to the input of the target language and improve the semantic transfer ability by (ii) and (iii). These three aspects promote each other to transfer knowledge from the source to the target language. ChatZero allows placeholders to be included in the generated responses. Finally, we employ mBERT [8] to convert the placeholders into actual words. To summarize, we make the following contributions:

- We propose the idea of constructing a pseudo-target language by introducing placeholders. As far as we know, we are the first to study zero-shot dialogue generation task without using massive target language utterances.
- We propose a novel end-to-end zero-shot dialogue generation model ChatZero, which achieves cross-lingual knowledge transfer by minimizing representations in different languages through unsupervised contrastive learning.
- Extensive experiments on two multilingual benchmark datasets demonstrate that ChatZero can achieve more than 90% of the original performance under zero-shot conditions compared to supervised learning and achieve state-of-the-art performance compared with other baselines. Code associated with our work are available on gitHub repository[1].

2 RELATED WORK

2.1 Dialogue Generation

Dialogue generation systems aim to produce informative and fluent responses and have attracted considerable attention in academia. Early studies [31, 29] employing seq2seq [32] structure tend to generate dull and generic responses. Since the emergence of the transformer [34], it has gradually become the go-to method. The popularity of Transformers brings a new problem of heavy reliance on large-scale corpus, such as DialogGPT [37], BlenderBot [26], and LaMDA [33]. Although they achieve promising performances, depending on large-scale corpora severely limits the usability of dialogue systems. Most languages that are recorded have no corpora available [10, 7, 35]. This means that these methods fail to work in these languages. We propose a novel end-to-end zero-shot dialogue generation model to alleviate this problem.

2.2 zero-shot Learning

Dialogue generation has enjoyed a great boost utilizing neural network models. However, this is not the case for most languages, especially zero-shot ones with insufficient training corpus [5, 19]. Zero-shot learning is a method of learning without any target language training samples. One of the solutions to zero-shot learning is cross-lingual transfer learning method, which improves the performance in the zero-shot target language by leveraging data from other (source) languages, typically with the help of cross-lingual resources. Cross-lingual transfer methods have been widely adopted in natural language processing tasks such as machine translation [9, 6, 16]. In this paper, we propose a cross-lingual, zero-shot generative model for dialogue generation, which does not depend on the target corpus.

3 Problem Statement

3.1 Problem Formalization

Given the source-language dialogue corpus and source-to-other language bilingual dictionaries, our goal is to build dialogue generation systems for other languages. In this paper, English is the source language. We concatenate the dialogue history of source language into a continuous sequence, denoted as $\tilde{\boldsymbol{H}}$, and the response denoted as $\tilde{\boldsymbol{R}}$. We employ $\boldsymbol{D}_t$ to denote a bilingual dictionary from English to target language *t*.

3.2 Code-switching Languages

Constructing the code-switching language containing source and target language is a common method for cross-lingual semantic transfer. During the training process, input is provided in the form of code-switching, whereas during inference, the input is in the form of the target language. The gap makes models unable to adapt to zero-shot scenarios. On the other hand, the target language tokens included in the training is limited, resulting in poor semantic transfer ability of models. We propose constructing a pseudo-target code-switching language to alleviate this limitation.

We employ bilingual dictionaries[2] to build code-switching languages. The dictionary is English-centric, including En-Zh (English-Chinese), En-De (English-German), En-Ru (English-Russian), En-Es (English-Spanish), En-Fr (English-French), En-It (English-Italian). We also collect other bilingual dictionaries and expand the size of the dictionary to improve its coverage of the corpus. The statistical information of the dictionaries is shown in Table 1. An English word may have multiple counterparts in other languages with the same meaning. At the same time, we count the coverage of different dictionaries in the English corpus. The calculation method is shown as follows:

$$f = \frac{Count(Dict \cap Corpus)}{Count(Corpus)} \tag{1}$$

Count means no repeat count function. We remove stop words and punctuation. The results are shown in Table 1. It can be seen that the dictionaries have a high coverage rate for DSTC7-AVSD. The coverage rate has a direct impact on the performance of ChatZero.

[1] https://github.com/misonsky/ChatZero

[2] https://github.com/facebookresearch/MUSE

Next, we introduce the process of constructing code-switching languages. We mainly build two forms of code-switching languages: (i) code-switching languages consisting of source and target languages; (ii) fake target languages containing placeholders and target languages. We employ the "[MASK]" symbol for placeholders. We can employ mBERT to manifest "[MASK]" a concrete token. We adopt Algorithm 1 to represent the construction process of code-switching languages.

Symbolic descriptions: The input of Algorithm 1 is $\tilde{\boldsymbol{H}}$, $\tilde{\boldsymbol{R}}$, and $\boldsymbol{D}$. $\tilde{\boldsymbol{H}}$, $\tilde{\boldsymbol{R}}$, where $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{R}}$ are the source language.

We employ $\bar{\boldsymbol{H}} = \{\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots\}$ and $\bar{\boldsymbol{R}} = \{\boldsymbol{r_1}, \boldsymbol{r_2}, \cdots\}$ to denote the dialogue histories and responses sets output by Algorithm 1, where $\boldsymbol{h_i} = \{w_1, w_2, \cdots, w_s\}$ and $\boldsymbol{r_i} = \{w_1, w_2, \cdots, w_t\}$, i represents the i-th example generated, s and t denote the sequence length of dialogue history and response, respectively. $\boldsymbol{\tau}$ represents a threshold value. The output $\bar{\boldsymbol{H}}$ and $\bar{\boldsymbol{R}}$ contain pseudo-target and code-switching lang. The $\boldsymbol{k}$ represents the number of iterations.

To distinguish different languages, we add an additional language identification token preceding both dialogue history and response. Specifically, English, Chinese, German, Spanish, French, Italian, and Russian are respectively set as: <En>, <Zh>, <De>, <Es>, <Fr>, <It> and <Ru>. In particular, we employ [Cs] as the language identification token for the code-switching language. The dialogue history and response of an English example can be expressed as $\boldsymbol{h}_{en} = \{\text{<En>}, w_1, w_2, \cdots, w_s\}$ and $\boldsymbol{r}_{en} = \{\text{<En>}, w_1, w_2, \cdots, w_t\}$. Adding language identification tokens for other languages is similar to English.

Table 1. Information of Dictionaries. **#key** represents the number of key values in the dictionary, **#val** represents the number of values, **#cov-da** and **#cov-ds** represent the dictionary's coverage of english training corpus on data sets DailyDialog and DSTC7-AVSD respectively.

Items	De	Ru	Es	Fr	It	Zh
#key	33,104	42,930	38,902	28,605	34,480	18,183
#val	37,633	45,989	41,334	31,065	35,443	19,452
#cov-da	47.01	42.52	47.29	43.85	44.68	37.68
#cov-ds	73.69	73.52	79.10	70.26	70.01	63.22

4 METHODOLOGY

The overall framework is illustrated in Figure 2. We use contrastive learning to minimize the semantic gaps between source language (i.e., English), code-switching language, and pseudo-target language at the encoding and decoding ends. ChatZero enhances the cross-lingual semantic transfer ability by direct and implicit semantic alignment. Besides, it also adapts to different forms of input during the inference stage.

4.1 Multilingual Transformer

A multilingual dialogue generation model learns a function f to model the relation between dialogue history and response, which can be applied to different languages. Unfortunately, there are no pre-trained models available for multilingual dialogue generation. An alternative is to use the mBERT initialized Transformer as multilingual dialogue generation [27]. Specifically, the encoder and decoder are initialized by mBERT checkpoints.

Algorithm 1: Code-Switching Languages.

Input: history $\tilde{\boldsymbol{H}}$; response $\tilde{\boldsymbol{R}}$; dictionary D; parameter k, τ; placeholder symbol S.
Output: code-switching history $\bar{\boldsymbol{H}}$ and response $\bar{\boldsymbol{R}}$

```
Initialize H̄, R̄ and local variable h as empty set ∅;
while k>0 do
    foreach token in H̃ or R̃ do
        If token in D do
            tokens ← GetDictValues(D, token)
            selection ← RandGetTokens(tokens)
            AddOperation(h,selection)
        else
            AddOperation(h,S)
    UpdateSetOperation(H̄, R̄, h)
    ClearSetOperation(h)
    foreach token in H̃ or R̃ do
        If token in D do
            tokens ← GetDictValues(D, token)
            If RandomNumber() > τ do
            selection ← RandGetTokens(tokens)
            AddOperation(h,selection)
            else
            AddOperation(h,tokens)
        else
            AddOperation(h,S)
    UpdateSetOperation(H̄, R̄, h)
    k = k-1
```

4.2 Cross-lingual Contrastive Learning

Cross-lingual mechanisms enable the implicit learning of shared representations of different languages. ChatZero introduces contrastive loss to explicitly bring different languages together to map a shared semantic space. The core idea of contrastive learning is to minimize the representation gap of similar utterances and maximize that of irrelevant utterances. We leverage contrastive learning on the encoder side and decoder side, respectively.

We assume that the output at the encoder side is denoted as $\widetilde{\boldsymbol{h}} = \{\widetilde{h}_{cls}, \widetilde{h}_{lan}, \widetilde{h}_1, \widetilde{h}_2, \cdots, \widetilde{h}_s, \widetilde{h}_{sep}\}$, where h_{lan} represents the representation of language identification token. The mean of all token representations is considered the representation of dialogue history, denoted as c, and the calculation method is as follows:

$$c = \frac{1}{s}\sum_{i}\widetilde{h}_i \tag{2}$$

According to Algorithm 1, we will get $2\times k$ +1 examples that are mutually positive instances. For $2\times k$ +1 positive examples, we push their semantics close by maximizing the cosine similarity between them, which can be formally described as:

$$\ell_p^e = \frac{1}{2k+1}\sum_{i>j}^{2k+1}\frac{c_i^T c_j}{||c_i||||c_j||} \tag{3}$$

where ℓ_p^e represents the similarity score of multiple positive examples. Besides, we maximize the distinction between positive and negative examples by maximizing the cosine similarity between each positive and negative pair. Negative examples are other samples from the same batch. The loss is calculated as follows:

$$\ell_n^e = \frac{1}{2k+1}\sum_{i}^{2k+1}\sum_{j}\frac{c_i^T N_j}{||c_i||||N_j||} \tag{4}$$

where N_j represents the j-th negative example in the same batch.

On the decoder side, we assume that the probabilities obtained by decoding is $\boldsymbol{P} \in \mathbb{R}^{t\times v}$, where t is the decoded length and v stands

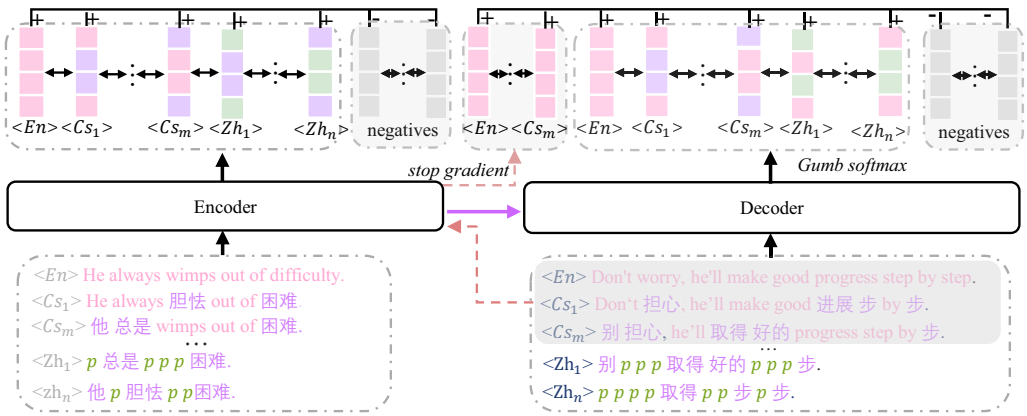

Figure 2. Overview of ChatZero. m represents the number of code-switching samples. n represents the number of fake target language samples. p means placeholder.

for the size of vocabulary $\boldsymbol{V}$. We employ *Gumb-Softmax* to sample the probability $\boldsymbol{P}$ to get the predicted probability distribution. The process can be formally described as follows:

$$\tilde{\boldsymbol{P}} = \text{Gumb-Softmax}(\boldsymbol{P}) \tag{5}$$

where $\tilde{\boldsymbol{P}} \in \mathbb{R}^{t \times v}$ and $\sum_j \tilde{\boldsymbol{P}}_{ij} = 1$. $\tilde{\boldsymbol{P}}$ not only represents the probability distribution of each word appearing in the response, but also can be regarded as the weight corresponding to each word. Therefore, we can obtain the semantic representation of the predicted response through the dot product of $\tilde{\boldsymbol{P}}$ and $\boldsymbol{V}$. The process can be formally described as follows:

$$\boldsymbol{r} = \tilde{\boldsymbol{P}}\boldsymbol{V} \tag{6}$$

where $\boldsymbol{r} \in \mathbb{R}^{t \times d}$ and d represents the hidden dimension of embedding layer. $\tilde{\boldsymbol{P}}$ assigns the correct candidate token a higher weight score, and the weight values of the remaining incorrect candidate tokens are close to zero. Equation 6 considers the information of incorrect tokens in a weighted approach, which mainly draws on the idea of label smoothing [22].

Note that $\boldsymbol{r}$ contains placeholders [MASK]. The representation of response is denoted as $\tilde{\boldsymbol{r}}$ by taking the mean of all tokens representations. For predicted responses, we employ contrastive learning to minimize the gap between positive examples and maximize that between positive and negative examples. To minimize the number of placeholders in the predictions, we introduce the ground truth responses as rectification signals to the positive examples. We only adopt source language (i.e., English) and code-switching responses as rectification signals. The reason is that fake target responses contains placeholders, which will encourage model to generate responses with more [MASK]. We only obtain representations of ground truth responses with encoder and do not compute the gradients. On the decoder side, ℓ_p^d and ℓ_n^d are calculated as follows on decoder side:

$$\ell_p^d = \frac{1}{3k+2} \sum_{i>j}^{3k+2} \frac{\tilde{\boldsymbol{r}}_i^T \tilde{\boldsymbol{r}}_j}{||\tilde{\boldsymbol{r}}_i|| ||\tilde{\boldsymbol{r}}_j||} \tag{7}$$

$$\ell_n^d = \frac{1}{3k+2} \sum_{i}^{3k+2} \sum_{j} \frac{\tilde{\boldsymbol{r}}_i^T N_j}{||\tilde{\boldsymbol{r}}_i|| ||N_j||} \tag{8}$$

where N_j represents the j-th negative example. We get the negative examples from other ground truth responses in the same batch. Besides, the loss for generative responses is cross-entropy, defined as:

$$\ell_g = \sum -log f_\theta(w_r | w_h) \tag{9}$$

The final total loss is defined as:

$$L = \ell_g + \frac{1}{4t}(\ell_n^e + \ell_n^d - \ell_p^d - \ell_p^e) \tag{10}$$

Since ℓ_g is calculated on the token-level, therefore contrastive loss should be multiplied by the averaged response length t.

5 Experiments

5.1 Datasets

We employ multilingual versions of **DailyDialog** and **DSTC7-AVSD** datasets [18], which include seven language versions i.e., English, Chinese, German, Russian, Spanish, French and Italian. The English version is the original corpus, and other language versions are obtained through bilingual dictionaries and translations. The supervised model for each language is trained on the corresponding language corpus. **DailyDialog** is a multi-turn dialogue dataset of daily life, which consists of 11,118 context-response pairs for training, 1,000 pairs for validation, and 1,000 pairs for testing. **DSTC7-AVSD** is a social media multi-turn dialogue dataset that consists of 76,590 context response pairs for training, 17,870 pairs for validation, and 1,710 pairs for testing.

5.2 Baselines

We compare the proposed model performance with the following baselines: **LVM** [20] which refines the aligned cross-lingual word-level representations by very few parallel word pairs. In this paper, we refine the cross-lingual word-level representations by bilingual dictionaries. **MLT** [21] which leverages parallel word pairs to generate code-switching sentences for learning the interlingual semantics across languages **OBPE** [24] which modifies the BPE algorithm to encourage more shared tokens between high-resource and low-resource languages tokens in the vocabulary. We obtain the OBPE word vocabulary based on bilingual dictionaries. We use mBERT as the base model and use mBERT tokenizer. This method does not

Table 2. Performance of the ChatZero comparison under zero-shot and supervised learning on DailyDialog. The **sup** and **zero** denote supervised learning performance and zero-shot performance, respectively. **Per** represents the percentage of zero-shot performance to supervised learning performance. **AVE** represents the average performance excluding PPL.

Language	Types	BLEU-1	BLEU-2	Rouge-L	Dist-1	Dist-2	Embed A/E/G			AVE	PPL
De	sup	34.22	27.91	37.71	18.75	53.84	77.27	61.17	84.31	49.40	101.57
	zero	32.11	25.13	34.72	15.89	49.13	72.81	56.75	80.36	45.86	109.26
	per(%)	93.83	90.03	92.07	84.74	91.25	94.22	92.77	95.31	91.78	91.21
Ru	sup	34.44	27.55	38.72	25.21	57.15	81.26	68.92	85.11	52.30	103.61
	zero	30.58	24.32	34.67	21.89	52.38	77.42	63.53	79.91	48.09	112.62
	per(%)	88.79	88.27	89.54	86.83	91.65	95.27	92.18	93.89	90.80	92.00
Es	sup	32.74	26.83	38.18	20.46	57.02	77.02	60.71	84.61	49.70	100.44
	zero	29.57	24.55	36.65	17.91	48.87	74.25	57.01	80.91	46.22	106.36
	per(%)	90.31	91.50	95.99	87.53	85.71	96.40	93.91	95.63	92.12	94.43
Fr	sup	33.12	27.43	37.58	16.84	48.18	80.32	62.68	86.27	49.05	101.56
	zero	30.02	24.82	34.71	14.33	43.15	77.21	58.26	81.09	45.45	110.19
	per(%)	90.64	90.48	92.36	85.10	89.56	96.12	92.94	94.00	91.40	92.16
Zh	sup	29.13	24.48	33.87	10.37	45.62	75.64	62.22	81.88	45.40	104.51
	zero	25.31	21.21	30.01	8.38	39.71	70.05	57.93	74.83	40.93	115.21
	per(%)	86.88	86.64	88.60	80.81	87.05	92.61	93.11	91.38	88.39	90.71
It	sup	33.18	26.74	37.29	21.85	58.68	75.91	60.01	84.26	49.74	100.55
	zero	30.06	24.85	34.44	19.84	53.86	72.26	55.21	79.73	46.28	108.35
	per(%)	90.60	92.93	92.36	90.80	91.79	95.19	92.00	94.62	92.54	92.80

work for source and target languages that have a big gap. Therefore, we also consider all source and target token pairs that appear in bilingual dictionaries share the same representation.

5.3 Implementation Details

We implement our ChatZero using PyTorch and train ChatZero on a server with an Intel(R) Xeon(R) Gold-5218R CPU 2.10GHz and 4×GeForce RTX 3090 GPU (24G). Adam [12] is utilized for optimization. The adam parameters beta1 and beta2 are set to 0.9 and 0.999, respectively. Note that we employ mBERT to perform placeholder (i.e., [MASK]) reprediction of ChatZero results. When predicting [MASK], we will concatenate the dialogue history and generated responses with [MASK] into a continuous sequence and input it into mBERT. The maximum length of the dialogue history is set to 512, and the maximum length of the response set to 50. We set the batch size to 64 and the learning rate to $5e^{-5}$. Beam search is used to generate responses. The beam size is set 6. We train word embeddings for embedding-based metrics for each language using Glove [25]. The unknown tokens are removed when computing embedding-based metrics, and the vectors of all unknown tokens are initialized to zero vector. We find that the parameter k is set to 2 for the best performance of ChatZero under zero-shot condition. The main reason is that most bilingual dictionaries have more than two candidate replacement words in target language with a ratio of 50%. Based on validation set experiments, the parameter τ is set to 0.4.

5.4 Evaluation metrics

We employ both automatic metrics and human evaluations. **Automatic Metrics**: Following previous studies [30, 18], we employ perplexity (PPL) and distinct-1/2 (i.e., Dist.1/2). A lower PPL means a more reliable result. Distinct-1/2 is a key metric to evaluate the diversity of responses, which can be calculated through the ratio of distinct uni-grams / bi-grams. Higher distinct means better diversity of responses generated by the model. The higher the diversity of responses generated by the model, the higher the value of the Distinct-1/2 metrics. Following previous studies [30, 18, 17], we also employ BLEU and ROUGE-L for evaluating response generation. BLEU [2] and ROUGE-L [2] metrics evaluate the response based on co-occurrence properties of tokens. Embedding-based metrics (Average, Exterma, and Greedy) [15, 36, 28] can reflect the quality of the generated responses at the semantic level.

Human Evaluation: Human evaluation mainly includes the following three aspects: (i) **Fluency** measures whether the generated responses are smooth or grammatically correct. (ii) **Diversity** evaluates whether the generated responses are informative, rather than generic and repeated information. (iii) **Relevance** evaluates whether the generated responses are relevant to the dialogue context. We select Chinese, French and German responses for human evaluation. We ask three crowdsourced graduate students to evaluate the quality of generated responses for 100 randomly sampled input contexts. We request annotators to score the response quality on a scale of [0,1,2] (0-bad, 1-neutral, 2-good) from three aspects fluency, diversity, and relevance. All annotators are unaware of the model corresponding to the generated results.

5.5 Results

Table 2 and Table 3 report the results of automatic metrics of ChatZero on DailyDialog and DSTC7-AVSD datasets under supervised and zero-shot conditions, respectively. First, ChatZero's performance under zero-shot conditions is inferior to that of supervised learning. However, the performance of zero-shot is exciting. Specifically, apart from Chinese datasets, ChatZero's zero-shot overall performance surpasses 90% of that achieved through supervised learning on DailyDialog. On DSTC7-AVSD dataset, ChatZero's overall zero-shot performance exceeds 90% of supervised learning across all languages. The results demonstrate the effectiveness of ChatZero.

We find an interesting phenomenon that ChatZero's cross-lingual capabilities are relatively weaker in Chinese and Russian compared to other languages. Specifically, ChatZero's zero-shot performance in Chinese is only 88.39% of supervised learning, and 90.8% in Russian on DailyDialog. On DSTC7-AVSD, ChatZero's zero-shot performance in Chinese is only 90.91% of supervised learning, and 92.38%in Russian. We believe this is related to the dictionary's cov-

Table 3. Performance comparison under zero-shot and supervised learning on DSTC7-AVSD.

Language	Types	BLEU-1	BLEU-2	Rouge-L	Dist-1	Dist-2	Embed A/E/G			AVE	PPL
De	sup	26.33	16.86	26.94	7.12	32.81	82.26	60.51	88.34	42.65	106.40
	zero	25.16	16.28	25.38	6.67	31.11	81.33	58.88	87.92	41.59	112.67
	per(%)	95.56	96.56	94.21	93.68	94.82	98.86	97.31	99.52	96.32	94.43
Ru	sup	32.01	20.56	37.11	7.56	22.74	83.33	65.31	89.74	44.80	142.53
	zero	29.27	18.67	34.10	6.54	20.88	80.12	62.01	86.34	42.24	156.37
	per(%)	91.44	90.08	91.88	86.51	91.82	96.14	94.95	96.21	92.38	91.14
Es	sup	30.28	20.42	34.21	7.04	28.01	84.58	67.61	89.12	45.16	105.93
	zero	28.64	18.43	32.47	6.56	26.68	82.89	64.53	87.61	43.48	111.86
	per(%)	94.58	90.25	94.91	93.18	95.25	98.00	95.44	98.31	94.99	98.31
Fr	sup	31.38	22.37	32.32	6.26	22.62	84.14	64.96	89.44	44.19	107.36
	zero	29.32	20.92	30.28	5.47	21.46	82.86	63.04	87.88	42.65	114.64
	per(%)	93.43	93.52	93.69	87.38	94.87	98.48	97.04	98.26	94.58	93.65
Zh	sup	25.26	16.07	31.61	6.47	20.31	77.56	64.91	85.78	41.00	152.69
	zero	23.04	14.64	29.23	5.24	18.04	73.52	61.47	79.92	38.14	164.47
	per(%)	91.20	91.12	92.47	81.00	88.84	94.79	94.70	93.16	90.91	92.83
It	sup	25.31	16.78	27.09	7.61	32.45	79.94	60.32	86.74	42.03	105.86
	zero	23.62	16.54	25.67	6.78	29.93	77.78	58.22	85.30	40.48	111.63
	per(%)	93.31	98.57	94.75	89.09	92.22	97.30	96.52	98.34	95.01	94.83

Table 4. Performances of baselines comparison on DailyDialog (up) and DSTC7-AVSD (down). **Bold** indicates the best result, and underline indicates the second best result.

Language	Model	BLEU-1	BLEU-2	Rouge-L	Dist-1	Dist-2	Embed A/E/G			AVE	PPL
De	LVM	29.53	25.55	30.05	13.77	45.44	68.55	51.77	78.54	42.90	117.88
	MLT	30.86	24.44	31.65	13.33	47.88	70.24	53.47	78.57	43.81	115.33
	OBPE	31.55	26.74	32.15	13.77	47.86	70.55	53.33	79.88	44.48	110.44
	ChatZero	32.11	25.13	34.72	15.89	49.13	72.81	56.75	80.36	**45.86**	**109.26**
Zh	LVM	27.66	22.54	33.64	14.29	45.88	72.88	55.62	78.35	43.86	115.77
	MLT	28.34	24.69	34.66	15.33	46.77	73.82	55.44	79.21	44.78	112.40
	OBPE	27.88	23.69	34.19	15.87	47.93	73.55	56.32	78.82	44.78	110.76
	ChatZero	29.57	24.55	36.65	17.91	48.87	74.25	57.01	80.91	**46.22**	**106.36**
Fr	LVM	28.34	21.65	33.48	11.38	39.59	75.58	55.87	78.66	43.07	119.43
	MLT	29.35	22.74	33.47	12.66	41.55	77.14	56.77	80.17	44.23	111.90
	OBPE	28.88	21.36	33.55	13.86	42.25	76.53	57.52	80.59	44.32	112.33
	ChatZero	30.02	24.82	34.71	14.33	43.15	77.21	58.26	81.09	**45.45**	**110.19**
De	LVM	22.05	13.80	21.86	4.76	28.06	77.22	56.04	83.26	38.38	121.68
	MLT	23.03	14.08	23.66	5.47	29.27	78.86	56.22	85.09	39.46	115.71
	OBPE	24.05	15.66	24.77	5.33	30.08	80.06	57.65	86.07	40.46	114.50
	ChatZero	25.16	16.28	25.38	6.67	31.11	81.33	58.88	87.92	**41.59**	**112.67**
Zh	LVM	20.21	12.38	26.87	4.44	17.54	69.08	58.46	76.07	35.63	172.36
	MLT	21.03	13.36	27.66	4.73	18.76	71.55	59.05	77.68	36.73	173.46
	OBPE	22.69	14.36	28.33	4.77	19.02	72.33	60.43	78.12	37.51	168.24
	ChatZero	23.04	14.64	29.23	5.24	18.04	73.52	61.47	79.92	**38.14**	**164.47**
Fr	LVM	27.43	18.56	28.86	3.77	18.50	80.09	61.86	85.36	40.55	123.33
	MLT	28.64	19.30	29.45	4.76	19.55	82.10	62.09	87.65	41.69	117.46
	OBPE	28.10	18.88	29.44	5.12	20.33	81.76	62.87	85.40	41.49	115.77
	ChatZero	29.32	20.92	30.28	5.47	21.46	82.86	63.04	87.88	**42.65**	**114.64**

erage of the corpus and the similarity between languages. The corresponding discussion is in Section 5.6. We can also observe similar phenomena on different metrics. On DailyDialog, the zero-shot BLEU-1/2 and Rouge-L results of ChatZero can achieve the performance of more than 90% of supervised learning in German, Spanish, French, and Italian. The performance of zero-shot learning on PPL and Embedding-based metrics have achieved more than 90% of supervised learning on all tested languages. On DSTC7-AVSD, we can observe that, except for the Distinct-1/2 metrics, the performance of other metrics under zero-shot reaches more than 91% of supervised learning. The results of some metrics under zero-shot are even close to that of supervised learning.

Table 4 reports the performance of ChatZero and other baselines on two datasets. We can observe that ChatZero enjoys the advantage of performance compared to other baselines. LVM achieves cross-lingual semantic transfer by training a shared embedding on word pairs. The disadvantage of LVM is that the size and coverage of word pairs strictly limit the model's semantic transfer capabili-

ties, which may cause models to appear OOV (out-of-vocabulary) on the target language. MLT achieves cross-semantic transfer by constructing code-switching sentences through corpus pairs. The inability to fully cover the target language often results in the generation of code-switching results. OBPE also constructs co-embedding between high-resource and low-resource languages, and faces the same problem as LVM. ChatZero avoids the challenge of generating code-switching forms by creating a pseudo-target language structure. It fills in placeholders in a manner that aligns with the language model's pre-training, compensating for the limited coverage of dictionaries.

5.6 Cross-lingual Analysis

We can observe that ChatZero demonstrates different cross-lingual abilities in different languages. The Chinese zero-shot performances of ChatZero on DailyDialog and DSTC7-AVSD have a gap compared to supervised learning than other languages. We believe this is mainly related to dictionary coverage and language similarity. From the dictionary coverage, the English-Chinese dictionary coverage of DailyDialog is 37.68%, and DSTC7-AVSD is 63.22%, which is the lowest compared to other dictionaries. The higher the dictionary coverage, the higher the bilingual pairs contained in the training corpus, which will significantly enhance the ability of ChatZero to transfer knowledge from the source language to the target language.

In order to explore the impact of language similarity on cross-lingual transfer learning, we calculate the similarity between English and other languages following previous study [23], we calculate the similarity between English and other languages, where Zh-En is 9.4%, Ru-En is 41.01%, Es-En is 54.51%, Fr-En is 54.67%, It-En is 68.65% and De-En is 91.84%. We find that Chinese and Russian are less similar to English than other languages, especially Chinese, align with our experimental observations, that is, ChatZero's cross-lingual capabilities are relatively weaker in Chinese and Russian compared to other languages. The similarity between languages can affect the cross-lingual ability of the model. We believe that cross-language transfer learning is easier between similar languages.

Table 5. Human evaluation results on DailyDialog (left) and DSTC7-AVSD (right).

Language	Models	Diversity	Relevance	Fluency	Diversity	Relevance	Fluency
De	LVM	0.850	0.580	1.096	0.635	0.629	0.996
	MLT	0.965	0.694	1.114	0.904	0.755	1.183
	OBPE	**1.110**	0.838	1.402	**1.100**	0.808	1.216
	ChatZero	1.066	**0.879**	**1.448**	1.035	**0.845**	**1.288**
Zh	LVM	0.678	0.644	0.908	0.557	0.683	1.104
	MLT	0.680	0.790	0.996	0.685	0.722	1.092
	OBPE	0.701	0.788	1.145	0.778	0.705	1.110
	ChatZero	**0.706**	**0.809**	**1.184**	**0.789**	**0.766**	**1.147**

5.7 Human Evaluation & Analysis

We compare the performance of ChatZero with other baselines from three dimensions through human evaluation. Table 5 reports the results. We can observe that ChatZero has obvious performance advantages in most evaluation dimensions. Specifically, on the German DailyDialog dataset, ChatZero has 30.8% advantage in diversity, 25.8% in relevance and 32.2% in fluency compared with LVM on DailyDialog. Compared with MLT, ChatZero has 11.6% advantage in diversity, 13.8% in relevance and 22.20% in fluency. On the DSTC7-AVSD dataset, we can observe a similar phenomenon. As we discussed previously, LVM suffers from OOV problems on the target language and MLT generates code-switching responses, which is the main reason for the performance inferiority of these models.

Context	Hey man, you wanna buy some weed? Some what? Weed! You know? Pot, Ganja, Mary Jane some chronic! Oh, no thanks. I also have blow if you prefer to do a few lines. ... Where do you get them from? I got my connections! Just tell me what you want and I'll even give you one ounce for free. Sounds good! Let's see, I want.
De	Ich [MASK], dass du deine [MASK] hinter deinen Kopf legen! Ich möchte, dass du deine hände hinter deinen Kopf legen!
Ru	чтобы ты положил руки за голову! чтобы ты положил руки за голову!
Es	¡ [MASK] que pongas tus manos detrás de tu [MASK]! ¡ Qui que pongas tus manos detrás de tu cabeza
Fr	Mettez [MASK] mains derrière la tête Mettez vos mains derrière la tête
It	[MASK] le mani dietro la testa, sei in arresto! Met le mani dietro la testa, sei in arresto!
Zh	我[MASK] 你把[MASK][MASK]头后面,你被捕了 我想你把手放头后面,你被捕了

Figure 3. The **context** stands for dialogue history. Here we only give the dialogue context in English. **De** represents the response in German under the zero resource condition, where the first row represents the response generated by ChatZero with placeholders, and the second row represents the responses of using mBERT to re-predict the placeholders. Other languages are similar. We highlight placeholders in blue, correct predictions by mBERT in red, and incomplete words in orange.

Although the performance of model OBPE is close to that of model ChatZero, it still has a slight disadvantage. The idea of OBPE maximizes word overlap is limited by the degree of similarity between source and target languages. This method is more effective when the source and target language are similar. We can observe that the performance of OBPE in German is closer to ChatZero than in Chinese, and even exceeds ChatZero in the diversity dimension.

To further evaluate the effectiveness of ChatZero, we count the proportion of placeholders in the responses generated by ChatZero. The ratio of placeholders [MASK] is obtained by dividing the number of placeholders by the number of all words in the responses. The placeholders in the generated responses are at a low level in both datasets, with an average of 8.74% placeholders on DailyDialog and 6.28% on DSTC7-AVSD. We find that the placeholders of the generated responses on DailyDialog are significantly higher than those on DSTC7-AVSD, which shows that the higher the dictionary coverage, the lower the placeholders in the generated responses.

5.8 Case Study

We further analyze the performance of the model through case Figure 3. It is an effective method to employ mBERT to predict the placeholders [MASK] when the generated responses contain fewer placeholders. We can observe that our approach still has some problems. The responses contains incomplete tokens. The main reason is that the number of tokens used to express the same semantics in different languages is inconsistent. We can not pre-set the number of placeholders when constructing pseudo-target language containing placeholders, which encourages ChatZero to assume that the number of tokens to express the same semantics is consistent in different languages. On the other hand, some words are splited into smaller tokens after word segmentation by the wordpiece tokenizer used in mBERT. These two reasons result in incomplete words being generated when making predictions for placeholders by mBERT.

6 Conclusion

We propose a novel zero-shot dialogue generation model ChatZero by introducing placeholders to build a pseudo-target language, which can avoid generating code-switching responses. ChatZero makes full use of the advantages of language models to make up for the shortcomings of incomplete dictionary coverage. Specifically, ChatZero utilizes unsupervised contrastive learning to minimize the semantic gap of source, code-switching and pseudo-target languages. Results on two multilingual dialogue datasets show that ChatZero achieves more than 90% of the supervised learning performance. Compared with baselines, ChatZero achieves state-of-the-art performance.

Acknowledgement

We would like to thank reviewers for their constructive comments. The project is supported by the National Natural Science Foundation of China (62172086,62272092) and DFG (grant SCHU 2246/14-1). The project is also supported by China Scholarship Council.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] B. Chen and C. Cherry. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367, 2014.

[3] G. Chen, S. Ma, Y. Chen, D. Zhang, J. Pan, W. Wang, and F. Wei. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157, 2022.

[4] W. Chen, Y. Gong, S. Wang, B. Yao, W. Qi, Z. Wei, X. Hu, B. Zhou, Y. Mao, W. Chen, et al. Dialogved: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4852–4864, 2022.

[5] X. Chen, A. H. Awadallah, H. Hassan, W. Wang, and C. Cardie. Zero-resource multilingual model transfer: Learning what to share. 2018.

[6] Y. Cheng. Joint training for pivot-based neural machine translation. In *Joint Training for Neural Machine Translation*, pages 41–54. Springer, 2019.

[7] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[9] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[10] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, 2020.

[11] J. Karlgren and P. Kanerva. High-dimensional distributed semantic spaces for utterances. *Natural Language Engineering*, 25(4), 2019.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] E.-S. Lee, S. Thillainathan, S. Nayak, S. Ranathunga, D. Adelani, R. Su, and A. D. McCarthy. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, 2022.

[14] L. Li, C. Xu, W. Wu, Y. Zhao, X. Zhao, and C. Tao. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485, 2020.

[15] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. 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, 2016.

[16] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[17] Y. Liu, S. Feng, D. Wang, H. Schütze, and Y. Zhang. Pvgru: Generating diverse and relevant dialogue responses via pseudo-variational mechanism. *arXiv preprint arXiv:2212.09086*, 2022.

[18] Y. Liu, S. Feng, D. Wang, and Y. Zhang. Mulzdg: Multilingual code-switching framework for zero-shot dialogue generation, 2022.

[19] Y. Liu, E. Nie, Z. Hua, Z. Ding, D. Wang, Y. Zhang, and H. Schütze. A unified data augmentation framework for low-resource multi-domain dialogue generation. *arXiv preprint arXiv:2406.09881*, 2024.

[20] Z. Liu, J. Shin, Y. Xu, G. I. Winata, P. Xu, A. Madotto, and P. Fung. 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, 2019.

[21] Z. Liu, G. I. Winata, Z. Lin, P. Xu, and P. Fung. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440, 2020.

[22] R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[23] E. Nie, S. Liang, H. Schmid, and H. Schütze. Cross-lingual retrieval augmented prompt for low-resource languages. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[24] V. Patil, P. Talukdar, and S. Sarawagi. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, 2022.

[25] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[26] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, et al. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, 2021.

[27] S. Rothe, S. Narayan, and A. Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280, 2020.

[28] J. Sedoc, D. Ippolito, A. Kirubarajan, J. Thirani, L. Ungar, and C. Callison-Burch. Chateval: A tool for chatbot evaluation. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 60–65, 2019.

[29] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[30] H. Song, Y. Wang, K. Zhang, W. Zhang, and T. Liu. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. 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 167–177, 2021.

[31] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *proceedings of the 24th ACM international on conference on information and knowledge management*, pages 553–562, 2015.

[32] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[33] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[35] X. Wang, S. Ruder, and G. Neubig. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, 2022.

[36] X. Xu, O. Dušek, I. Konstas, and V. Rieser. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, 2018.

[37] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL (demo)*, 2020.

[38] X. Zhao, W. Wu, C. Tao, C. Xu, D. Zhao, and R. Yan. Low-resource knowledge-grounded dialogue generation. *International Conference on Learning Representations*, 2020.