XDailyDialog: A Multilingual Parallel Dialogue Corpus

Zeming Liu[1]*, Ping Nie[2]*, Jie Cai[2]*, Haifeng Wang[3]†, Zheng-Yu Niu[3], Peng Zhang[4], Mrinmaya Sachan[5], Kaiping Peng[4]

[1]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, China

[2]Peking University [3]Baidu Inc., Beijing, China [4]Tsinghua University [5]ETH Zurich

zmliu@ir.hit.edu.cn {ping.nie, caijie}@pku.edu.cn

{wanghaifeng, niuzhengyu}@baidu.com

Abstract

High-quality corpora are significant to the development of dialogue models. However, most existing corpora for open-domain dialogue modeling are limited to a single language. The absence of multilingual open-domain dialog corpora not only limits the research on multilingual or cross-lingual transfer learning but also hinders the development of robust open-domain dialogue systems that can be deployed in other parts of the world. In this paper, we provide a multilingual parallel open-domain dialog dataset, **XDailyDialog**,[1] to enable researchers to explore the challenging task of multilingual and cross-lingual open-domain dialogue. **XDailyDialog** includes 13K dialogues aligned across 4 languages (52K dialogues and 410K utterances in total). We then propose a dialogue generation model, **kNN-Chat**, which has a novel kNN-search mechanism to support unified response retrieval for monolingual, multilingual, and cross-lingual dialogue. Experiment results show the effectiveness of this framework.

1 Introduction

Developing high-quality open-domain dialogue systems is one of the key challenges in Artificial Intelligence. Unlike closed-domain dialogue systems which deal with specific kinds of conversations (like a chatbot for customer support), open-domain dialogue systems can engage in conversation on any topic. In recent years, there has been a significant increase in research on dialogue due to the rise of voice-based bots, such as Meena (Adiwardana et al., 2020), BlenderBot (Roller et al., 2021) and XiaoIce (Zhou et al., 2020). To advance the quality of open-domain dialogue systems, many large-scale corpora have been created (Sordoni et al., 2015; See et al., 2019; Yang et al., 2018; Mazaré et al., 2018; Keskar et al., 2019).

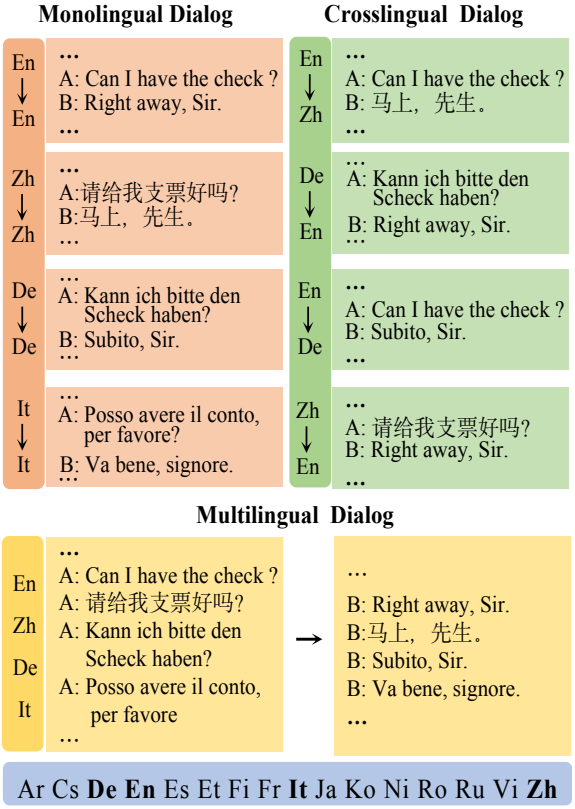

Figure 1: Illustration of XDailyDialog with the monolingual, multilingual, and crosslingual dialogue.

However, to the best of our knowledge, almost all existing large-scale corpora for open-domain dialogue modeling are limited to a single language, such as English (Sordoni et al., 2015; See et al., 2019; Yang et al., 2018; Mazaré et al., 2018; Keskar et al., 2019), or Chinese (Shang et al., 2015; Wu et al., 2017; Cai et al., 2019; Wang et al., 2020). The absence of multilingual open-domain dialogue corpora not only limits the research on multilingual or cross-lingual transfer learning (Lin et al., 2021) but also hinders the development of robust open-domain dialogue systems that can be deployed in other parts of the world. Previous work on various NLP tasks has shown that multilingual corpora can bring performance improvements in multilingual

* Equal contribution

† Corresponding author: Haifeng Wang.

[1]https://github.com/liuzeming01/XDailyDialog

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12240–12253
July 9-14, 2023 ©2023 Association for Computational Linguistics

or cross-lingual transfer learning. This includes tasks such as conversational recommendation (Liu et al., 2021), task-oriented dialog (Schuster et al., 2019b), semantic parsing (Li et al., 2021), QA and reading comprehension (Jing et al., 2019; Lewis et al., 2020; Artetxe et al., 2020; Clark et al., 2020; Hu et al., 2020; Hardalov et al., 2020), machine translation (Johnson et al., 2017b), document classification (Lewis et al., 2004; Klementiev et al., 2012; Schwenk and Li, 2018), semantic role labelling (Akbik et al., 2015) and NLI (Conneau et al., 2018). Thus, we believe that multilingual training data might enhance multilingual or cross-lingual transfer learning for open-domain dialogue as well.

To facilitate the study of multilingual and cross-lingual dialogue, we present a multilingual **parallel** dialog dataset, **XDailyDialog**, for multilingual and cross-lingual open-domain dialogue. *XDailyDialog* consists of 13K dialogues aligned across 4 languages (52K dialogues and 410k utterances in total). **The most significant advantage of parallel data over non-parallel data, such as XPersona(Lin et al., 2020), is that it can support cross-lingual tasks.**

We define 3 task settings using *XDailyDialog*. As shown in Figure 1, the first task is monolingual dialogue, where dialogue context and response are in the same language. It aims at investigating the performance variation of the same model across different languages. The figure also illustrates another task that is called multilingual dialogue. Here we directly mix training instances of the 4 languages into a single training set and train a single model to handle multilingual dialogue at the same time. Finally, the last task is cross-lingual dialogue, where model input and output are in different languages, e.g. dialog context is in English and the generated response is in Chinese.

To address these tasks, we build a model using k-Nearest Neighbors (kNN) and mBART (Liu et al., 2020a). We conduct an empirical study of the proposed model on *XDailyDialog*. Our experimental results indicate that the use of cross-lingual data can bring performance improvement in the monolingual dialogue.

Our work makes the following contributions:

- To facilitate the study of multilingual and cross-lingual dialogue, we create a novel corpus *XDailyDialog*, the first publicly available **multilingual parallel open-domain** dialogue corpus.
- We define 3 tasks, including monolingual, multilingual, and crosslingual dialog, based on *XDailyDialog*. Automatic evaluation and human evaluation results confirm the benefits of this corpus for monolingual dialogue.
- We propose a dialog generation framework, **kNN-Chat**, with a novel kNN-search mechanism that can support unified token retrieval for monolingual, multilingual, and cross-lingual dialogue generation. Our experimental results confirm the effectiveness of this framework.

2 Related Work

Multilingual and Cross-lingual Datasets for dialog Multilingual dialog datasets are relatively scarce. Lin et al. (Lin et al., 2020) propose a Multilingual Persona-Chat dataset, XPersona, by extending the Persona-Chat corpora (Dinan et al., 2019) to 6 languages: Chinese, French, Indonesian, Italian, Korean, and Japanese. In XPersona, the training sets are automatically translated using translation APIs, while the validation and test sets are annotated by humans. XPersona focuses on cross-lingual personalized dialog and is not parallel, while *XDailyDialog* focuses on multilingual and cross-lingual dialog generation and is parallel. Liu et al(Liu et al., 2021) construct a Multilingual conversational recommendation dataset, DuRecDial 2.0, by Crowdsourcing translation based on DuRecdial (Liu et al., 2020b) to 2 languages: Chinese, and English. DuRecDial 2.0 focuses on conversational recommendation, while *XDailyDialog* focuses on dialog generation and has more languages.

Multilingual and Cross-lingual Datasets for Task-oriented Dialog Several multilingual task-oriented dialogue datasets have been published (Mrkšić et al., 2017b; Schuster et al., 2019a), enabling evaluation of the approaches for cross-lingual dialogue systems. (Lin et al., 2021) introduces the first bilingual multi-domain dataset for task-oriented dialogue modeling, which has only two languages and is not parallel. mrksi et al .(Mrkšić et al., 2017b) annotated two languages (German and Italian) for the dialogue state tracking dataset WOZ 2.0 (Mrkšić et al., 2017a) and trained a unified framework to cope with multiple languages. Meanwhile, Schuster et al. (Schuster et al., 2019a) introduced a multilingual NLU dataset and

highlighted the need for more sophisticated cross-lingual methods. Those datasets mainly focus on multilingual NLU and DST for task-oriented dialogue and are not parallel. In comparison with them, *XDailyDialog* is a multilingual dataset for open-domain dialog, which has 4 languages and is parallel.

3 Dataset Collection

XDailyDialog aims to collect high-quality parallel data for the research of monolingual, multilingual, and cross-lingual open-domain dialog. It is translated from DailyDialog (Li et al., 2017), which is a human-written, well-formatted English dataset. In this section, we describe how *XDailyDialog* is collected.

3.1 Data Collection

Human Translation We select 3 common languages (Italian, Chinese, and German) from 3 different language families to be translated by professional translators.[2] To guarantee the quality of translation, we use a strict quality control procedure.

First, we randomly sample 100 conversations from DailyDialog and assign them to more than 50 professional translators. Following (Liu et al., 2021), all translation results are assessed by 3 data specialists with translation experience after translation. Specifically, data specialists randomly select 20% of each translator's translation results for assessment. The assessment is done at a word-level, utterance-level, and session-level. For word-level assessment, they assess whether the choice of words is appropriate, and whether there are typos. For utterance-level assessment, they assess whether the utterance is accurate and colloquial. For session-level assessment, they assess whether the session is coherent and parallel to DailyDialog. Only if the error rate is less than 5%, the translators can pass. Finally, we pick 20 translators.

Then, the 20 translators translate about 500 utterances at a time. After data translation, data specialists randomly select 10-20% of each translator's translation results for assessment in the same way as above. The translators can continue to translate only after they pass the assessment.

[2]http://www.ethnologue.com

3.2 Dataset Quality Analysis and Statistics

Quality Analysis of Human Translation We conduct human evaluations for data quality. A dialog will be rated "1" if all utterances are accurate and colloquial and the dialogue session is coherent., otherwise "0". Then we ask 3 data specialists with translation experience to judge the quality of 100 randomly sampled dialogues (about 800 dialogue utterances). [3] Finally, we obtain an average score of 0.96, 0.97, and 0.98 for German, Italian, and Chinese, respectively.

#Parallel dialogues	52K
#Parallel utterances	411K
#Average Utterance Per Dialogue	7.9
#Average Tokens Per Utterance	11.5
Languages	En, De, Zh, It

Table 1: Languages and statistics of the *XDailyDialog*.

Dataset Statistics Table 1 provides statistics of *XDailyDialog*, indicating rich dialog languages. We believe that *XDailyDialog* would better facilitate the study of multilingual and cross-lingual dialog.

4 Task Formulation on *XDailyDialog*

Let $D = \{(X_i, \mathcal{T}_i, \mathcal{A}_i, \mathcal{E}_i, Y_i)\}_{i=1}^{n}$ denote a set of dialogues in *XDailyDialog*, where $\forall i \in \{1, \ldots, n\}$, X_i refers to a dialog context, Y_i is a response to X_i, and $\mathcal{T}_i$, $\mathcal{A}_i$, $\mathcal{E}_i$ is the dialogue topic, dialogue act, dialogue emotion corresponding to Y_i, respectively. Given a context $X = \{u_j\}_{j=0}^{i-1}$ associated with a dialogue topic $\mathcal{T}$, a dialogue act $\mathcal{A}$ and a dialogue emotion $\mathcal{E}$, the aim is to produce a proper response $Y = u_i$, where u_j and u_i are dialogue utterances. As shown in Table 1, the dialogues in XDailyDialog include languages $L = (De, En, It, Zh)$.

Monolingual dialog: **Task 1:** (X_l, $\mathcal{T}_l$, $\mathcal{A}_l$, , $\mathcal{E}_l$, Y_l), where $\forall l \in L$. With these 4 monolingual dialogue forms, we can investigate the performance variation of the same model trained on 4 separate datasets in different languages. In our experiments, we train 4 models respectively for the 4 monolingual tasks. Then we can evaluate their performance variation across all languages to see how the changes between languages can affect model performance.

[3]We calculate the averaged weighted kappa value for all sampled dialogues and get a high score of 0.81, demonstrating good agreements between data specialists.

Multilingual dialog: **Task 2:** (X_L, $\mathcal{T}_L$, $\mathcal{A}_L$, , $\mathcal{E}_L$, Y_L). Similar to multilingual conversational recommendation (Liu et al., 2021), multilingual neural machine translation (Johnson et al., 2017b) and multilingual reading comprehension (Jing et al., 2019), we directly mix training instances of the 4 languages into a single training set and train a single model to handle the 4 languages dialogue at the same time. This task setting can help us investigate if the use of additional training data in other languages can bring performance benefits for a model of the current language.

Cross-lingual dialog: The cross-lingual dialogue is **Task 3:** (X_{l2}, $\mathcal{T}_{l1}$, $\mathcal{A}_{l1}$, $\mathcal{E}_{l1}$, Y_{l1}), where $l1$ and $l2$ are two different languages in L. In **Task 3**, when given a related dialogue topic, dialogue act, and dialogue emotion (e.g., in Engish), the model takes dialog context in one language (e.g., in Chinese) as input, and then produces responses in another language (e.g., in Engish) as output. Understanding the mixed-language dialog context is a desirable skill for end-to-end dialog systems. This task setting can help evaluate if a model has the capability to perform cross-lingual tasks.

5 Our Approach

We propose a non-parametric method named **kNN-Chat** for monolingual, multilingual, and cross-lingual dialogue generation, as shown in Figure 2. kNN-Chat is inspired by kNN language models such as kNN-LM (Khandelwal et al., 2020b) and kNN-MT(Khandelwal et al., 2020a) and implemented for monolingual, multilingual, and cross-lingual dialogue by using dialogue context, response, and extra information (dialogue emotion, topic, action, target language). Compared to recent dialogue systems with kNN (Fan et al., 2020), our experimental results show that kNN-Chat, a non-parametric method (without extra training except for pre-training mBART), could also be effective for monolingual, multilingual, and cross-lingual dialogue generation.

5.1 Model Architecture

We utilize a unified generative framework for the given 4 monolingual tasks, 4 cross-lingual tasks, and 1 multilingual dialogue task. We choose the unified generative framework as it has a) interpretable responses, b) flexible language control, and c) efficient training and inference.

Specifically, **kNN-Chat**, contains three modules: (1) a generative module to encode multiple languages and decode coarse-grained target language responses (mBART is used in our experiments). (2) a datastore module to store key-value pairs, where the key is the representation of dialogue context and extra information, and the value is the corresponding response token. (3) a kNN-search module (*faiss* (Johnson et al., 2017a) is used in our experiments) to search similar dialogue tokens from the datastore according to the generative model's representation of the next token.

Moreover, **kNN-Chat** is *interpretable* as similar dialogue contexts are retrieved for generation and is *flexible* as it is adaptable to different language settings with corresponding arbitrary amount of data size by using *faiss* (Johnson et al., 2017a). Besides, it is *efficient* as it can extend to any trained generative model such as mBART or GPT without extra training processes.

5.2 Generative Module

As shown in Figure 2, we use mBART (Liu et al., 2020a) as our generative model for all tasks. We do not modify the mBART encoder-decoder architecture but design the input-output for mBART. Specifically, we concatenate dialogue emotion, dialogue topic, and dialogue act with dialogue context for better response generation as the input of mBART. Since the response could be in different languages, we also append a language identifier of the response to the end of the input. The output of mBART is the coarse-grain dialogue response in the target language specified by the language identifier.

We use the pre-trained mBART model from (Liu et al., 2020a) and fine-tune it on task-specific corpus using the faiseq (Ott et al., 2019). When mBART is finetuned, mBART can be directly used to generate dialogue response and served as our baseline model.

5.3 Datastore Module

After mBART is finetuned, we construct the datastore of kNN-Chat based on trained mBART before response generation. Datastore consists of a set of key-value pairs. Let $I = \{X, \mathcal{A}, \mathcal{T}, \mathcal{E}\}$, given an input-response pair $(in, y) \in (I, Y)$ from the training set, we use the trained mBART model to generate the t-th token y_t based on the input and generated tokens $(in, y_{<t})$. When mBART generates t-th token y_t, it also produces a high-dimensional

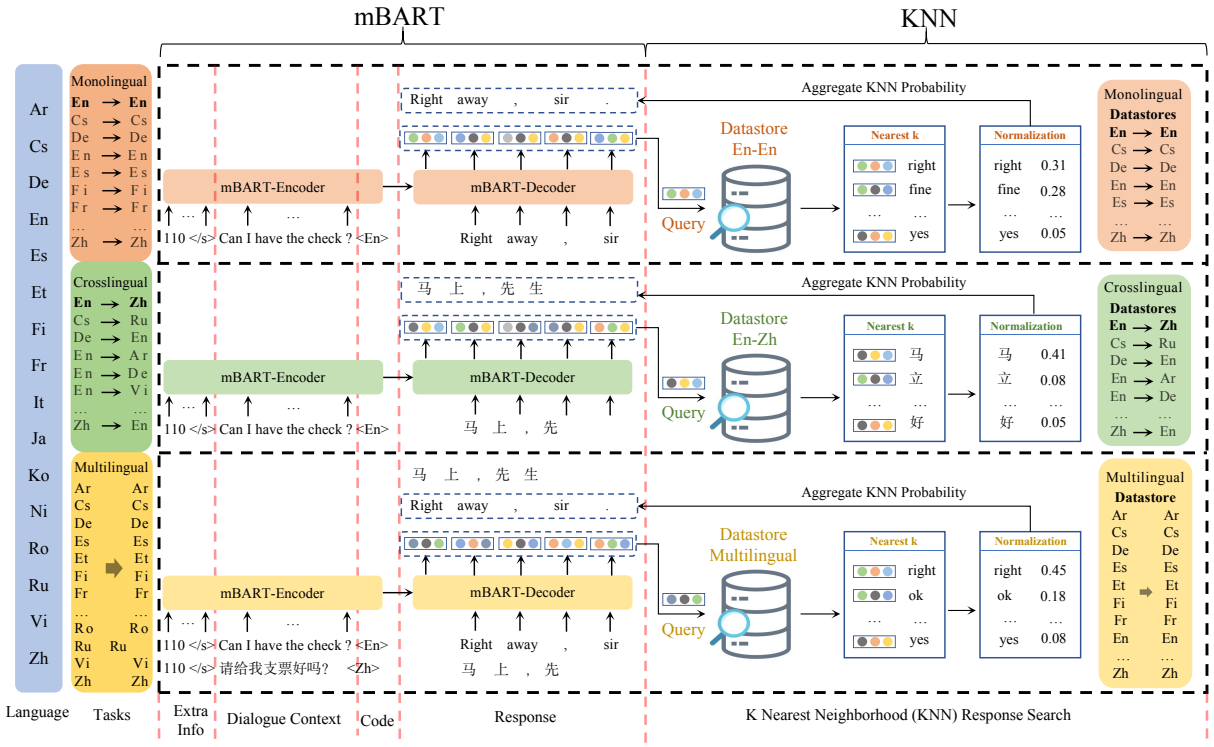

Figure 2: Architecture of our KNN-Chat model.

representation vector. The vector corresponding to y_t can be denoted as $f(in, y_{<t})$. Finally, each token of response in the training set has a representation vector. The representation $f(in, y_{<t})$ and response token y_t are then used as key and value in the datastores respectively:

$$(K, V) = \bigcup_{(in,y)\in(I,Y))} \{(f(in, y_{<t}), y_{<t}), \forall y_t \in y\}$$

Datastore can be created by one forward pass of mBART from the training set. For monolingual and cross-lingual tasks, we build one datastore for one language setting. For multilingual tasks, we build datastores for each language separately. The datastore size (number of training set response tokens) is set to 2 million for monolingual and cross-lingual experiments and 20 million for the multilingual experiment. We train *faiss* (Johnson et al., 2017a) index on datastore and then search potential response tokens from these datastores by kNN-search module.

5.4 KNN-search Module

When kNN-Chat predicts y_t at time step t, mBART produces the representation $f(in, y_{<t})$ for y_t according to the generated tokens $y_{<t}$ and the dialogue input. kNN-Chat uses the representation of y_t to search k nearest neighbors. Suppose the queried neighbors for $f(in, y_{<t})$ are $N^t = \{(k_i, v_i) \in (K, V), i \in \{1, 2, \ldots, k\}\}$, where k_i and v_i are i-th key vector and i-th tokens in nearest neighbors. Then the distribution of neighbors over the decoder vocabulary can be calculated as:

$$p_{knn}(y_t|in, y_{<t}) \propto \sum_{(k_i,v_i)} \mathbb{1}_{y_t=v_i} exp(\frac{-d(k_i, f(in, y_{<t}))}{T}))$$

where T is the temperature and $d()$ denotes the distance between representation vectors. After we get the p_{knn}, the final probability of y_t can be computed as the interpolation of two distributions with a hyper-parameter λ:

$$p(y_t|in, y_{<t}) = \lambda p_{knn}(y_t|in, y_{<t}) + (1-\lambda) p_{mBART}(y_t|in, y_{<t})$$

where $p_{mBART}(y_t|in, y_{<t})$ denotes the vanilla mBART prediction probability in section 5.2.

6 Experiments and Results

6.1 Experiment Setting

Dataset For the train/development/test set of *XDailyDialog*, we follow the split of (Li et al., 2017),

with one notable difference that we filtered the duplicate data in (Li et al., 2017).

We conduct both automatic and human evaluation for kNN-Chat and baselines on *XDailyDialog*.

Automatic Evaluation Metrics: For automatic evaluation, we follow the setting in previous work (Li et al., 2017) to use several common metrics such as F1, BLEU (Papineni et al., 2002) (BLEU1-4), and DISTINCT (DIST-1 and DIST-2) (Li et al., 2016) to measure the relevance, fluency, and diversity of generated responses.

Human Evaluation Metrics: The human evaluation is conducted at the level of both turns and dialogues. For turn-level human evaluation, we ask each model to produce a response conditioned on a given context. The generated responses are evaluated by 4 evaluators in terms of fluency, appropriateness, and informativeness. For dialogue-level human evaluation, we let each model converse with evaluators. For each model, we collect 30 dialogues. These dialogues are then evaluated by 4 evaluators in terms of coherence that examines fluency, relevancy, and logical consistency of each response when given context. The evaluators rate the dialogues on a scale of 0 (poor) to 2 (good) in terms of each human metric.[4]

6.2 Baselines

We carefully select two strong baselines for multilingual and crosslingual natural language generation.

mBART (Liu et al., 2020a) is a multilingual sequence-to-sequence denoising auto-encoder pre-trained on CC25 (Wenzek et al., 2020; Conneau et al., 2020). It provides a set of parameters that can be fine-tuned for any of the language pairs in CC25, including all languages in *XDailyDialog*. We treat our 3 tasks as Machine Translation tasks as (Liu et al., 2021). The mBART model can serve as a strong baseline for multilingual and cross-lingual dialogue generation.

mT5 (Xue et al., 2021) is a massively multilingual pre-trained text-to-text transformer model, trained following a similar recipe as T5 (Raffel et al., 2020). It can be fine-tuned for 100 languages, including all languages in *XDailyDialog*.

6.3 Experiment Results

Table 2, 3, 4, 5, 6, and 7 report automatic evaluation results and human evaluation results of our model and all baselines.

6.3.1 Experiment Results for Monolingual dialog

According to the evaluation results in Table 2, and Table 3, **kNN-Chat** outperforms mBART and mT5 across almost all the tasks or metrics, which confirms the effectiveness of kNN-Chat for monolingual dialog generation. The possible reason is that kNN-search can retrieval more appropriate tokens for dialog generation. Furthermore, Chinese dialog(1(Zh→Zh)) is more challenging than other tasks. One possible reason is that both kNN-Chat and baselines can not well model the dialogue context of those two languages.

6.3.2 Experiment Results for Multilingual dialog

Our model vs. Baselines: According to the evaluation results in Table 4, and Table 5, **kNN-Chat** outperforms mBART and mT5 across almost all the tasks or metrics, which also demonstrates the effectiveness of kNN-Chat for multilingual dialog generation and the flexibility for different language settings.

Monolingual vs. Multilingual: Based on the results in Table 2, 4, 3, and 5, we can find that all languages get worse results in the multilingual task. It indicates that multilingual dialogue generation is a more challenging task than monolingual dialogue generation. The possible reason is that it is more difficult to train the model to generate in multiple languages than to generate in a single language. Moreover, we can find that multilingual tasks are better than monolingual tasks for mT5. It indicates that the use of additional corpora can improve mT5's performance for multilingual dialog. But the other two models for other language multilingual tasks can not outperform the monolingual tasks. The possible reason is that the pre-trained models can not perform well in the modeling of 4 languages dialog utterances, resulting in poor model performance.

6.3.3 Experiment Results for Cross-lingual dialog

Our model vs. Baselines: According to the evaluation results in Table 6 and Table 7, **kNN-Chat** outperforms mBART and mT5 across almost all the tasks or metrics, which indicates the effectiveness of kNN-Chat for cross-lingual dialog generation and the flexibility for different language settings.

[4]Please see Appendix A.2 and A.3 for more details.

Tasks	Methods	F1	BLEU1	BLEU2	BLEU3	BLEU4	DIST-1	DIST-2
1(De→De)	mBART	36.73%	0.105	0.020	0.007	0.002	0.034	0.093
1(De→De)	mT5	48.46%	0.095	0.015	0.004	0.002	0.109	0.287
1(De→De)	kNN-Chat	**53.61%**	**0.206**	**0.143**	**0.132**	**0.128**	**0.206**	**0.554**
1(En→En)	mBART	33.95%	0.327	0.054	0.021	0.009	0.012	0.047
1(En→En)	mT5	40.46%	0.146	0.040	0.018	0.007	0.093	0.252
1(En→En)	kNN-Chat	**53.68%**	**0.351**	**0.170**	**0.151**	**0.144**	**0.087**	**0.377**
1(It→It)	mBART	31.03%	0.089	0.022	0.012	0.008	0.030	0.072
1(It→It)	mT5	38.21%	0.092	0.018	0.007	0.003	0.133	0.330
1(It→It)	kNN-Chat	**50.91%**	**0.161**	**0.108**	**0.101**	**0.097**	**0.204**	**0.555**
1(Zh→Zh)	mBART	16.40%	0.293	0.047	0.011	0.005	0.024	0.082
1(Zh→Zh)	mT5	20.21%	0.342	0.066	0.024	0.008	0.067	0.217
1(Zh→Zh)	kNN-Chat	**26.57%**	**0.337**	**0.157**	**0.123**	**0.109**	**0.119**	**0.432**

Table 2: Automatic evaluation results for monolingual dialog on *XDailyDialog*. All results with p-value < 0.01.

Tasks	Methods	Fluency	Appro.	Info.	Coherence
1(De→De)	mBART	**1.96**	0.90	0.76	0.83
1(De→De)	mT5	1.94	0.88	0.80	0.85
1(De→De)	kNN-Chat	1.95	**0.93**	**1.34**	**0.89**
1(En→En)	mBART	**1.99**	0.74	0.74	0.74
1(En→En)	mT5	1.93	0.71	0.82	0.72
1(En→En)	kNN-Chat	1.91	**0.82**	**1.29**	**0.83**
1(It→It)	mBART	**1.99**	0.88	0.70	0.81
1(It→It)	mT5	1.92	0.86	0.91	0.78
1(It→It)	kNN-Chat	1.94	**0.99**	**1.25**	**0.87**
1(Zh→Zh)	mBART	**1.98**	0.83	0.75	0.83
1(Zh→Zh)	mT5	1.93	0.79	0.86	0.78
1(Zh→Zh)	kNN-Chat	1.89	**0.86**	**1.38**	**0.89**

Table 3: Human evaluation results for monolingual on *XDailyDialog* at the level of turns and dialogues. "Appro.", and "Info."stand for appropriateness, and informativeness respectively.

Tasks	Methods	F1	BLEU1	BLEU2	BLEU3	BLEU4	DIST-1	DIST-2
2(De→De)	mBART	26.00%	**0.350**	0.035	0.010	0.003	0.003	0.004
2(De→De)	mT5	49.67%	0.123	0.026	0.008	0.002	0.113	0.263
2(De→De)	kNN-Chat	**50.11%**	0.272	**0.040**	**0.022**	**0.015**	**0.127**	**0.295**
2(En→En)	mBART	31.59%	0.289	0.037	0.010	0.003	0.003	0.007
2(En→En)	mT5	45.41%	0.160	0.045	0.019	0.007	**0.055**	0.177
2(En→En)	kNN-Chat	**46.22%**	**0.293**	**0.045**	**0.020**	**0.012**	0.049	**0.181**
2(It→It)	mBART	19.98%	0.342	0.028	0.009	0.003	0.004	0.008
2(It→It)	mT5	45.65%	0.115	0.028	0.010	0.004	0.123	0.277
2(It→It)	kNN-Chat	**45.98%**	**0.259**	**0.030**	**0.017**	**0.012**	**0.124**	**0.284**
2(Zh→Zh)	mBART	12.04%	0.284	0.034	0.002	0.000	0.002	0.004
2(Zh→Zh)	mT5	24.96%	**0.346**	0.029	0.031	0.010	0.072	0.188
2(Zh→Zh)	kNN-Chat	**26.18%**	0.249	**0.033**	**0.015**	**0.011**	**0.074**	**0.191**

Table 4: Automatic evaluation results for multilingual dialog on *XDailyDialog*. All results with p-value < 0.01.

Monolingual vs. Cross-lingual:

According to the results in Table 2, 3, 6, and 7, the model performs surprisingly better on almost all cross-lingual tasks than the corresponding monolingual tasks (for example, 3(En→De) is better than 1(De→De)) in terms of almost all the automatic and human metrics, which is similar to (Liu et al., 2021). It indicates that the use of multilingual corpora can consistently bring performance improvement for monolingual dialog. One possible reason is that kNN-Chat or mBART can fully exploit the multilingual dataset, resulting in better model performance.

6.3.4 Human Evaluation

Our human evaluation results are reported in Table 3, 5, and 7, which show that all the models can generate fluent responses. In general, kNN-Chat is also better than mBART and mT5 in this evaluation, which further confirms the effectiveness of kNN-Chat for monolingual, multilingual, and cross-lingual dialog generation. However, the

Tasks	Methods	Fluency	Appro.	Info.	Coherence
2(De→De)	mBART	**2.00**	0.37	0.63	0.32
2(De→De)	mT5	1.96	0.45	0.69	0.46
2(De→De)	kNN-Chat	1.92	**0.47**	**0.73**	**0.51**
2(En→En)	mBART	1.93	0.29	0.29	0.29
2(En→En)	mT5	**1.94**	0.73	0.66	0.59
2(En→En)	kNN-Chat	1.89	**0.87**	**0.82**	**0.87**
2(It→It)	mBART	1.96	0.31	0.56	0.42
2(It→It)	mT5	**1.97**	0.49	0.69	0.58
2(It→It)	kNN-Chat	1.94	**0.52**	**0.71**	**0.63**
2(Zh→Zh)	mBART	1.81	0.35	0.54	0.49
2(Zh→Zh)	mT5	1.92	0.46	0.44	0.59
2(Zh→Zh)	kNN-Chat	**1.94**	**0.54**	**0.56**	**0.61**

Table 5: Human evaluation results for multilingual on *XDailyDialog* at the level of turns and dialogues. "Appro.", and "Info."stand for appropriateness, and informativeness respectively.

Tasks	Methods	F1	BLEU1	BLEU2	BLEU3	BLEU4	DIST-1	DIST-2
3(En→De)	mBART	45.44%	0.116	0.025	0.009	0.003	0.074	0.221
3(En→De)	mT5	47.40%	0.099	0.015	0.004	0.001	0.065	0.180
3(En→De)	kNN-Chat	**59.13%**	**0.262**	**0.200**	**0.189**	**0.185**	**0.218**	**0.588**
3(De→En)	mBART	44.32%	0.281	0.053	0.024	0.012	0.023	0.117
3(De→En)	mT5	36.84%	0.111	0.022	0.008	0.002	0.058	0.145
3(De→En)	kNN-Chat	**58.67%**	**0.392**	**0.233**	**0.216**	**0.210**	**0.094**	**0.411**
3(Zh→En)	mBART	42.73%	0.290	0.057	0.025	0.012	0.024	0.121
3(Zh→En)	mT5	36.09%	0.104	0.020	0.007	0.002	0.041	0.105
3(Zh→En)	kNN-Chat	**58.65%**	**0.396**	**0.235**	**0.217**	**0.211**	**0.098**	**0.417**
3(En→Zh)	mBART	18.01%	0.278	0.050	0.017	0.010	0.038	0.148
3(En→Zh)	mT5	19.28%	0.351	0.067	0.024	0.008	0.041	0.147
3(En→Zh)	kNN-Chat	**30.09%**	**0.353**	**0.182**	**0.147**	**0.128**	**0.121**	**0.449**

Table 6: Automatic evaluation results for cross-ingual dialog on *XDailyDialog*. All results with p-value < 0.01.

Tasks	Methods	Fluency	Appro.	Info.	Coherence
3(En→De)	mBART	**1.88**	0.61	1.23	0.42
3(En→De)	mT5	1.81	0.53	1.11	0.39
3(En→De)	kNN-Chat	1.83	**0.82**	**1.62**	**0.75**
3(De→En)	mBART	**1.96**	0.84	1.36	0.61
3(De→En)	mT5	1.86	0.77	1.23	0.55
3(De→En)	kNN-Chat	1.93	**0.99**	**1.56**	**0.77**
3(Zh→En)	mBART	1.90	0.87	0.88	0.84
3(Zh→En)	mT5	**1.92**	0.71	0.73	0.68
3(Zh→En)	kNN-Chat	1.89	**1.10**	**1.09**	**1.06**
3(En→Zh)	mBART	1.52	0.72	0.66	0.62
3(En→Zh)	mT5	1.41	0.57	0.55	0.61
3(En→Zh)	kNN-Chat	**1.70**	**0.83**	**0.80**	**0.81**

Table 7: Human evaluation results for cross-lingual on *XDailyDialog* at the level of turns and dialogues. "Appro.", and "Info."stand for appropriateness, and informativeness respectively.

models score much lower on appropriateness, informativeness, and coherence. These results highlight the challenges in building a multilingual or cross-lingual dialog system and opportunities for future progress.

7 Conclusion

To facilitate the study of multilingual and cross-lingual dialog, we created the first publicly available multilingual parallel dataset, *XDailyDialog*, for dialog and proposed three challenging tasks for the community based on it. Furthermore, We also built a new conversation generation framework, **kNN-Chat**, with a novel kNN-search mechanism that can support unified response retrieval for monolingual, multilingual, and cross-lingual dialog generation. Expensive experiment results confirm the effectiveness of this framework. We hope that *XDailyDialog* would help push forward research in the unified end-to-end monolingual, multilingual,

and cross-lingual conversational modeling.

Limitations

The main limitation of this work is that the pre-training model can not deal with numerical knowledge well. In future work, we will try to enhance the ability of the pre-training model to better deal with numerical knowledge.

Ethics Statement

We make sure that *XDailyDialog* has been collected in a manner that is consistent with the terms of use of any sources and the intellectual property and privacy rights of the original authors of the texts. And crowd workers were treated fairly. This includes but is not limited to, compensating them fairly and ensuring that they were able to give informed consent, which includes but is not limited to, ensuring that they were voluntary participants who were aware of any risks of harm associated with their participation. In this paper, we propose a novel multilingual corpus for end-to-end dialog training and evaluation. Our corpus neither introduces any social/ethical, since we generate data by human translation or machine translation. We do not foresee any direct social consequences or ethical issues.

Acknowledgments

Thanks for the insightful comments from the reviewers. This work was supported by the National Key R&D Program of China (2021ZD0110501).

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

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, Beijing, China. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. 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 1866–1875, Hong Kong, China. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*.

Emily Dinan, V. Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur D. Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, A. Black, Alexander I. Rudnicky, J. Williams, Joelle Pineau, M. Burtsev, and J. Weston. 2019. The second conversational intelligence challenge (convai2). *ArXiv*, abs/1902.00098.

Angela Fan, Claire Gardent, Chloe Braud, and Antoine Bordes. 2020. Augmenting transformers with knn-based composite memory for dialogue. *arXiv preprint arXiv:2004.12744*.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

J. Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and M. Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ArXiv*, abs/2003.11080.

Yimin Jing, Deyi Xiong, and Zhen Yan. 2019. BiPaR: A bilingual parallel dataset for multilingual and

cross-lingual reading comprehension on novels. 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 2452–2462, Hong Kong, China. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017a. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

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

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020a. Nearest neighbor machine translation.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020b. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

A. Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. 5:361–397.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, S. Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. *ArXiv*, abs/1910.07475.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pages 110–119.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. Xpersona: Evaluating multilingual personalized chatbot. *ArXiv*, abs/2003.07568.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020b. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

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

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael

Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and et al. 2021. Recipes for building an open-domain chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Sebastian Schuster, S. Gupta, Rushin Shah, and M. Lewis. 2019a. Cross-lingual transfer learning for multilingual task oriented dialog. *ArXiv*, abs/1810.13327.

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

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. 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 1577–1586, Beijing, China. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. *ArXiv*, abs/2008.03946.

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

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

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, Online. Association for Computational Linguistics.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

A Appendix

A.1 Training and Inference Parameters of kNN-Chat

We train and inference all our experiments on NVIDIA-SMI Quadro RTX 8000 GPU. The CUDA version is 11.4. Experiments are conducted with fairseq and *faiss* tool. The parameters we used are shown in table 8.

module	Parameter	value
mBART	Pre-trained model	mbart.cc25
	Learning Rate (Lr)	3e-5
	Lr Scheduler	Polynomial
	Warmup Update	2500
	Max Tokens	1024
	Optimizer	Adam
kNN-Chat Datastore	Monolingual Size	2,000,000
	Cross-lingual Size	2,000,000
	Multilingual Size	2,000,000
	Lambda	0.2
	Temperature	10
kNN-Chat Faiss Search	Probe Number	32
	Ncentroids	4096
	Quantizer	IndexflatL2
	Index	IndexIvFPQ
	Use Float16	True

Table 8: Model parameter settings.

A.2 Turn-level Human Evaluation Guideline

Fluency measures fluency of each response:

- score 0 (bad): unfluent and difficult to understand.
- score 1 (fair): there are some errors in the response text but still can be understood.
- score 2 (good): fluent and easy to understand.

Appropriateness examines relevancy of each response when given the context:

- score 0 (bad): not relevant to the current context.
- score 1 (fair): relevant to the current context, but using some irrelevant knowledge.
- score 2 (good): otherwise.

Informativeness

- score 0 (bad): safe response / universal response, and not relevant to the current context.
- score 1 (fair): safe response / universal response, but relevant to the current context.
- score 2 (good): otherwise.

A.3 Dialogue-level Human Evaluation Guideline

Coherence measures fluency, relevancy, and logical consistency of each response when given the current context:

- score 0 (bad): more than two-thirds responses irrelevant or logical contradictory to the given context.
- score 1 (fair): more than one-third of responses are irrelevant or logical contradictory to the given current context.
- score 2 (good): otherwise.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

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

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1 and 2

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

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

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

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

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

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

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3 and 6

C ☒ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.1

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.1

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

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

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Ethics Statement

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics Statement

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.