Pchatbot: A Large-Scale Dataset for Personalized Chatbot

Hongjin Qian[1], Xiaohe Li[2], Hanxun Zhong[2], Yu Guo[2], Yueyuan Ma[2], Yutao Zhu[3]
Zhanliang Liu[2], Zhicheng Dou[1*], Ji-Rong Wen[1,2]
[1] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2] School of Information, Renmin University of China, Beijing, China
[3] Université de Montréal, Québec, Canada
{ian,*dou}@ruc.edu.cn

ABSTRACT

Natural language dialogue systems raise great attention recently. As many dialogue models are data-driven, high-quality datasets are essential to these systems. In this paper, we introduce Pchatbot, a large-scale dialogue dataset that contains two subsets collected from Weibo and Judicial forums respectively. To adapt the raw dataset to dialogue systems, we elaborately normalize the raw dataset via processes such as anonymization, deduplication, segmentation, and filtering. The scale of Pchatbot is significantly larger than existing Chinese datasets, which might benefit the data-driven models. Besides, current dialogue datasets for personalized chatbot usually contain several persona sentences or attributes. Different from existing datasets, Pchatbot provides anonymized user IDs and timestamps for both posts and responses. This enables the development of personalized dialogue models that directly learn implicit user personality from the user's dialogue history. Our preliminary experimental study benchmarks several state-of-the-art dialogue models to provide a comparison for future work. The dataset can be publicly accessed at Github: https://github.com/qhjqhj00/Pchatbot.

CCS CONCEPTS

• Computing methodologies → Language resources; **Discourse, dialogue and pragmatics**; **• Information systems**;

KEYWORDS

dataset, personalization, dialogue, chatbot

ACM Reference Format:
Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A Large-Scale Dataset for Personalized Chatbot. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3404835.3463239

1 INTRODUCTION

The dialogue system is a longstanding challenge in Artificial Intelligence. Intelligent dialogue agents have been rapidly developed but their effectiveness is still far behind general expectations. The reasons for the lag are multi-dimensional in which the lack of datasets is a fundamental constraint. Training a chatbot usually requires a large-scale dataset, but collecting real conversations between people requires tremendous human labor. Hence, most existing studies mainly leverage publicly available post-comments to simulate conversations between users. Example datasets are Ubuntu Dialogue Corpus [14] and Douban Corpus [30], which are sourced from online forums. As discussed by Gao et al. [8], there are still many challenges for dialogue chatbot. Personality consistency is one of these challenges. Regarding personalized dialogue dataset, previous works depict personality using either personality sentences (PERSONA-CHAT [32]) or persona attributes (Personality Assignment Dataset [19]). With the availability of these datasets, lots of dialogue models have been proposed.

Many recent neural dialogue models demonstrate substantial gains on chatbots' performances using large-scale dataset [3, 16, 33]. Thus, the scale of most current dialogue datasets becomes a major barrier that limits the development of neural dialogue models. The limitation of the data scale is exacerbated regarding the personalized dialogue dataset. The main reasons include: (1) Most current personalized datasets are based on the explicit user profiles. Such profiles are often manually annotated (*e.g.*, user's persona descriptions), thus being very costly. (2) Explicit user profiles can only provide relatively limited personalized information as explicit user profiles are usually comprised of static personality descriptions. For example, in PERSONA-CHAT [32], each interlocutor is described by five short sentences. Such static property also makes the explicit user profiles hard to update.

To tackle the aforementioned problems, in this work, we introduce Pchatbot, a large-scale Chinese conversation dataset dedicated to the development of personalized dialogue models. Pchatbot has two subsets, named PchatbotW and PchatbotL, built from open-domain Weibo and several judicial forums. They respectively have 130 million and 59 million high-quality conversations. To the best of our knowledge, Pchatbot is the largest Chinese dialogue dataset. As shown in Table 1, PchatbotW is 14 times larger than the current largest persona-based open-domain corpus in Chinese, *i.e.*, Personality Assignment Dataset [19]. The detailed statistics of Pchatbot are shown in Table 5.

An example of PchatbotW is illustrated in Figure 1. In addition to the content of the post-response pairs, the ID of the corresponding user and the publication timestamp are also provided. We believe such data have the potential to support several kinds of research questions, at least the following three: (1) *Single-turn dialogue.* This

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463239

Figure 1: Examples of PchatbotW.

task only considers the interaction between the post and the response within a single turn [11, 23, 27, 31, 36], thus being naturally supported by our dataset. As our dataset is significantly larger than other Chinese datasets, it is promising to learn a model with better performance [3, 16, 33]. (2) *Multi-referenced dialogue.* In natural language dialogue, a post can have multiple appropriate responses. So, recent studies start to explore generating diverse responses for a given input [10, 20, 34]. It is evident to see that our dataset can also support this kind of research since several responses corresponding to one post are collected. (3) *Personalized dialogue.* With the help of user IDs and timestamps, we can aggregate user-wise data to obtain the user's dialogue history. As tremendous personalized information (such as speaking style, vocabulary, and interests) is hidden under the user's dialogue history, intuitively we can design models that directly learn implicit user profiles from the dialogue history, which enlightens a new research way for personalized chatbots. Besides, the Pchatbot dataset has user ID for both sides of interlocutors, which expands the application scenarios where we can model the personalities of both the source user and the target user. In this work, we mainly focus on investigating the application of our dataset on personalized dialogue.

Dataset construction is usually confronted with two major challenges: (1) how to adapt the raw data to a specific task, and (2) how to protect user privacy. The most intuitive idea addressing the two challenges is to directly remove data noise or private information. However, we need to consider the scope of data noise or private information and how to recognize them. Furthermore, whether removing such texts undermines semantics also remains unsolved. After fine-grained data analysis and human evaluation, we conclude that the quality of the dataset can be effectively improved via removing meaningless conversations including hashtags, URLs, emoticons, text duplication, and multi-languages. Besides, yet the raw data in the Pchatbot dataset can be publicly accessed on the original websites, the raw datasets of Pchatbot are vulnerable to privacy leakage. The reason is that social platforms have ubiquitous private information such as phone numbers, emails, and social media accounts. To protect privacy, these texts are either replaced by indistinguishable marks or deleted depending on whether semantics would be undermined.

Pchatbot is a ready-to-use and well-documented dataset. It is licensed under CC BY-NC[1]. Researchers are required to fill in an application form to obtain the dataset files[2]. For both subsets PchatbotL and PchatbotW, we provide three release versions: (1) the cleaned dataset; (2) the standard dataset for generation-based chatbot; (3) the standard dataset for retrieval-based chatbot. Specifically, the cleaned dataset contains all information of the raw datasets after cleaning processes (*e.g.*, deduplication, anonymization, etc.). The standard datasets are constructed from the cleaned dataset and can be directly used for the corresponding task. For example, we prepare response candidates to each post in the retrieval-based standard dataset. Along with the dataset, we provide tools and tutorials that are used to load, clean, aggregate, and construct the datasets.

We experiment with both generation-based and retrieval-based dialogue models on the corresponding standard datasets to provide benchmark that can be used for comparison in the further study. Experimental results also verify the advantages of the availability of user IDs and the large-scale data can improve the performance.

In summary, the advantages of Pchatbot are as follows:

(1) To the best of our knowledge, the Pchatbot dataset is the largest Chinese dialogue dataset. Neural dialogue models might gain substantial improvements using such a large-scale dataset;

(2) Pchatbot dataset contains two subsets, namely PchatbotW and PchatbotL. The two subsets are dedicated to the open-domain and professional domain (judicial domain), respectively. Such diversity could broaden the application domains of dialogue chatbots.

(3) We include anonymized user IDs and timestamps in Pchatbot. This will greatly enlarge the potentiality for developing personalized dialogue agents that learn implicit user profiles from the user's dialogue history.

(4) We benchmark several state-of-the-art dialogue models for both generation-based and retrieval-based. Experimental results can be used for comparison in future study.

2 RELATED WORK

High-quality dialogue systems need large-scale dialogue datasets for training. However, it takes a lot of manual labor and time to collect real human conversation data. In recent years, scholars use post-comments to simulate human dialogue and have published a series of datasets. These datasets can be divided into domain-specific datasets [2, 14, 29] and open-domain datasets [7, 21, 24, 28, 30]. Lowe et al. [14] used Ubuntu Chat Logs to build the Ubuntu dialogue corpus with 930,000 dialogues. Chen et al. [5] constructed the JD Customer Service Corpus including 435,005 dialogues based on customer service dialogues from JD.com. These domain-specific datasets can be used to build task-oriented dialogue systems [2, 29].

Open-domain datasets contain conversation data for open topics. Due to the characteristics of social media, the text published by users on social networks is close to real human conversation. In recent years, researchers have constructed some open-domain datasets from social media, such as Twitter [21, 24], Weibo [23, 28],

[1]https://creativecommons.org/licenses/by-nc/2.0/
[2]https://github.com/qhjqhj00/Pchatbot/blob/main/application.pdf

Table 1: Statistics of existing dialogue corpora and Pchatbot. '-' means not being mentioned in corresponding papers. Dialogues means sessions in multi-turn conversations or pairs in single-turn conversations. Utterances means sentences in the dataset. PchatbotW and PchatbotL are the subsets of Pchatbot which we introduce in this paper.

Dataset	# Dialogues	# Utterances	# Words	Language	Source	Personalized Info.
Twitter Corpus [21]	1,300,000	3,000,000	-	English	Twitter	None
PERSONA-CHAT [32]	10,981	164,356	-	English	Crowdsourcing	Persona descriptions
Reddit Corpus [16]	700,000,000	1,400,000,000	-	English	Reddit	User ID & Timestamp
STC Data [28]	38,016	618,104	15,592,143	Chinese	Weibo	None
Noah NRM Data [23]	4,435,959	8,871,918	-	Chinese	Weibo	None
Douban Conversation Corpus [30]	1,060,000	7,092,000	131,747,880	Chinese	Douban	None
Personality Assignment Dataset [19]	9,697,651	19,395,302	166,598,270	Chinese	Weibo	Key-value pair profile
PchatbotW (Our)	139,448,339	278,896,678	8,512,945,238	Chinese	Weibo	User ID & Timestamp
PchatbotL (Our)	59,427,457	118,854,914	3,031,617,497	Chinese	Judicial Forums	User ID & Timestamp

Table 2: Statistics of the standard PchatbotW datasets.

	PchatbotW-R	PchatbotW-G
# Users	420,000	300,000
Avg. history length	32.3	11.4
Avg. # words of post	24.9	22.9
Avg. # words of response	10.1	9.6
# Response candidates	10	N/A
# Training samples	3,000,000	2,707,880
# Validation samples	600,000	600,000
# Testing samples	600,000	600,000

and Douban [30]. However, we argue that the scopes of these datasets are not enough to train data-driven dialogue systems.

As discussed by Vinyals and Le [27], it is still difficult for current dialogue systems to pass the Turing test, a major reason is the lack of a coherent personality. In order to train a coherent dialogue system, Li et al. [12] first attempts to model persona by utilizing user IDs to learn latent variables for representing each user in the Twitter dataset. But as far as we know, this dataset has not been made publicly available. To make chatbots maintain a coherent personality, other classic strategies mainly focus on how to endow dialogue systems with a coherent persona by pre-defined attributes or profiles [16, 19, 32]. These works restrict persona in a collection of attributes or texts, which ignore the language behavior and interaction style of a person.

In this work, we construct Pchatbot, a large-scale dataset with personalized information, to solve the mentioned issues. Pchatbot has two subsets from open domain and specific domain respectively.

Table 1 shows the data scales of Pchatbot and other datasets. In addition, all posts and responses of Pchatbot are attached with user IDs and timestamps, which can be used to learn not only persona profiles but interaction styles from the users' dialogue histories.

3 PCHATBOT DIALOGUE DATASET

Pchatbot dataset is sourced from public websites. It has two subsets from Weibo and judicial forums respectively. Raw data are normalized by removing invalid texts and duplications. Privacy information in raw data is also anonymized using indistinguishable placeholders.

3.1 Dataset Construction

Each item of raw data in the dataset is started by a post made by one user and multiple responses then follow. We extract the post-response pairs from the original threads.

3.1.1 General Preprocessing Pipeline. Since the Pchatbot dataset is collected from social media and forums, private information such as homepage, telephone, email, ID card number, and social media account, is ubiquitous. Besides, there are also many sensitive words such as pornography, abuse, and political words. Therefore, we design a preprocessing pipeline to deal with the raw data, which contains the following four steps:

(1) **Anonymization.** We replace private information in the data with placeholders using either rule-based methods or information extraction models. Specifically, we use regex expressions to recognize texts such as email, phone numbers, and account numbers. We use NER models to extract entities like names and addresses.

(2) **Filtering Sensitive Words.** The sensitive words are detected by the matching method with a refined sensitive word list[3]. As sensitive words are also very important in terms of semantics, simply replacing them with placeholders will undermine the completeness of the sentences. Therefore, we directly filter out all (post, response) pairs with sensitive words.

(3) **Filtering Utterances by Length.** We clean the utterance whose length is less than 5 or more than 200 because the short utterances tend to contain limited information, while the long utterances usually have noise.

(4) **Word Segmentation.** For Chinese word segmentation, we use Jieba toolkit[4]. Since Jieba is implemented for general Chinese word segmentation, we introduce a law terminology list[3] as an extra dictionary for enhancement in PchatbotL.

As the two subsets of Pchatbot are from different sources, in addition to the general preprocessing pipeline, the detailed preprocessing strategies and descriptions are introduced as follows.

3.1.2 PchatbotW. In China, Weibo is one of the most popular social media platforms for users to discuss various topics and express their opinions. The basic function of this platform is very similar to Twitter. We crawl the publicly available Weibo posts and their comments across one year (from September 10, 2018 to September

[3]https://github.com/fighting41love/funNLP
[4]https://github.com/fxsjy/jieba

10, 2019). Then, we randomly select about 23M users, and get their conversation histories. As a result, we obtain 341 million (post, response) pairs in total.

Due to the nature of social media, posts and responses published by Weibo users are extremely diverse that cover various aspects of daily life. Therefore, interactions between users can be considered as daily casual conversations in the open domain. Since Weibo text is in a casual manner containing a lot of noise, to improve the quality of the data, we do the following data cleaning operations in addition to the general preprocessing pipeline:

(1) **Removing Hashtags.** Users like to tag their contents with related topics by hashtags, which usually consist of several independent words or summaries wrapped by '#'. Since splicing hashtag text into contents will affect the semantic coherence, we remove the hashtags from the content. (2) **Removing URLs.** Users' posts and responses sometimes contain multimedia content, images, videos, and other web pages. They are converted to URLs in Weibo. We also remove these URLs from the content. (3) **Removing Emoticons.** On the Weibo platform, users can use emoticons (including emoji and kaomoji) to convey their emotions. Emoticons consist of various symbols, which introduce noises to dialogues. Therefore, we remove these emoticons by regex and dictionary. (4) **Handling Mentions.** On the platform, users can also use '@nickname' to mention or notice other users. When users comment or repost others, 'Reply @nickname:' and '//@nickname:' will be automatically added into the users' contents. These mentions serve as reminders and often have little relevance to the users' content. So, we remove them to guarantee the consistency of utterances. (5) **Handling Duplicate Texts.** Duplicate texts appear in different granularities. For word-level duplication, the duplicated Chinese characters are normalized to two characters. For example, "太好笑了，哈哈哈哈哈" ("That's so funny. hahahahaha") will be normalized as "太好笑了，哈哈" ("That's so funny. haha"). As for response-level duplication under a post, they are usually caused by different users sending the same responses. Duplicate responses under a post reduce the varieties of interactions, so we remove those occurring more than three times. Furthermore, we also find utterance-level duplication in the dataset. Duplicated utterances in the entire dataset may affect the balance of the dataset. Models are reported to generate general responses when being trained on a large number of duplicate utterances [12]. Specifically, in PchatbotW, we limit the frequency of the same utterances as 10,000. For utterances that have a frequency over 10,000, we randomly remove the over part. (6) **Multi-languages.** Due to the diversity of Weibo, some users' content contains multiple languages. We remove samples containing more than 70% content that is in other languages.

For fair comparison in future work, we construct two standard datasets from PchatbotW to evaluate dialogue models, namely PchatbotW-R and PchatbotW-G, which are used for retrieval-based and generation-based tasks, respectively. Concretely, following previous works that construct retrieval-based datasets, in PchatbotW-R, we retrieve 10 response candidates for each data sample. The generation-based dataset PchatbotW-G is directly derived from the PchatbotW. Statistics for the two datasets are shown in Table 2.

3.1.3 PchatbotL. Judicial forums are professional platforms that open to users for consultation and discussion in the judicial domain.

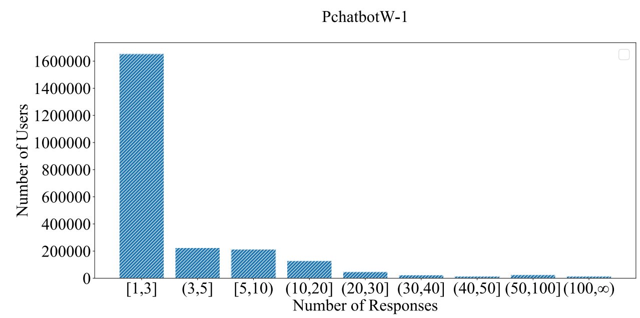

Figure 2: Distribution of users' number with different scopes of responses on PchatbotW-1

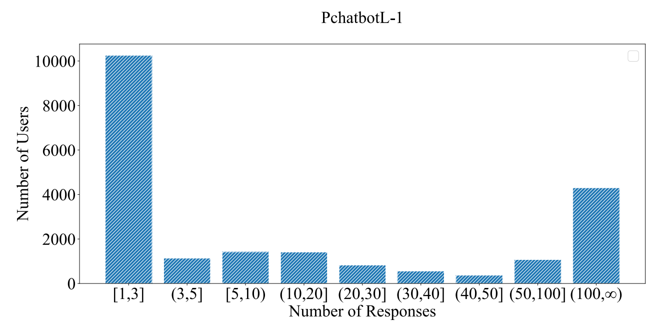

Figure 3: Distribution of users' number with different scopes of responses on PchatbotL-1

Table 3: An example of data in Pchatbot

Field	Content
Post	下冰雹了!真刺激! (Hailing! It's really exciting!)
Post user ID	5821954
Post timestamp	634760927
Response	出去感受更刺激 (It's more exciting to go out.)
Response user ID	592445
Response timestamp	634812525
Partition index (1-10)	1
Train/Dev/Test (0/1/2)	0

People can seek legal aid from lawyers or solve the legal problems of other users in the judicial forums.

We crawl around 59 million post-response pairs from 5 judicial websites, including 66law.cn, findlaw.cn, lawtime.cn, 110.com, and 9ask.cn, from October 2003 to February 2017. Since the data of the judicial forums are questions from users and answers from lawyers, topics mainly focus on the legal domain. Post-response pairs are almost of high quality so that only basic preprocesses are needed. More information for the dataset can be found on the release page.

3.2 Data Partition

The original Pchatbot dataset is very large. For convenient use, we divide Pchatbot into 10 partitions evenly according to the user IDs in responses. Each partition has a similar size of (post, response) pairs. PchatbotL-1 and PchatbotW-1 are the first partitions of PchatbotL and PchatbotW, respectively. Figure 2 and Figure 3 show the

Table 4: Experimental results on the PchatbotW-R and PchatbotW-G.

PchatbotW-R	R_{10}@1	R_{10}@2	R_{10}@5	MRR	PchatbotW-G	BLEU-1	ROUGE-L	Dist-1	Dist-2	P-F1
(1) Conv-KNRM	0.323	0.520	0.893	0.538	(2) Seq2Seq	4.889	7.594	0.299	3.404	0.771
(1) DAM	0.438	0.644	0.966	0.635	(2) Speaker	3.958	5.580	0.951	29.780	1.534
(1) IOI	**0.442**	**0.651**	**0.969**	**0.639**	(2) PersonaWAE	1.945	**9.064**	0.523	8.549	**6.408**
(1) RSM-DCK	0.428	0.627	0.947	0.623	(2) DialoGPT	**5.038**	7.458	**13.995**	**52.674**	3.562

Table 5: Detailed statistics. PchatbotW-1 and PchatbotL-1 are the 10% partitions of the corresponding subsets.

	PchatbotW	PchatbotL	PchatbotW-1	PchatbotL-1
# Posts	5,319,596	20,145,956	3,597,407	4,662,911
# Responses	139,448,339	59,427,457	13,992,870	5,523,160
# Users in posts	772,002	5,203,345	417,294	1,107,989
# Users in resp.	23,408,367	203,636	2,340,837	20,364
Avg. # resp. / post	26.214	2.950	3.890	1.184
Max. # resp. / post	525	120	136	26
Vocabulary Size	9,148,532	1,329,930	3,447,433	551,071
Avg. # words of post	49.37	36.88	49.40	37.26
Avg. # words of resp.	11.68	14.08	11.70	14.13

distribution of the length of users' dialogue history of PchatbotW-1 and PchatbotL-1, respectively. We find that the users in the PchatbotL tend to have more responses than users in the PchatbotW. The reason is that repliers in PchatbotL are usually professional lawyers who routinely provide judicial help in the forums. Other partitions have similar distributions. We also divide the datasets into train/dev/test sets. In the division of train/dev/test set, given a user, we ensure that the time of its records in the dev-set and test-set are behind the records in the train-set by using timestamps.

3.3 Data Format and Statistics

Pchatbot's schema is shown in Table 3. Each record of Pchatbot includes 8 fields: post, post user ID, post timestamp, response, response user ID, response timestamp, partition, and train/dev/test identity. User IDs and timestamps are attached in the Pchatbot dataset for each post or response. User IDs can be used to distinguish the publisher of each post or response. Timestamps provide temporal information that can be used to build a historical response sequence for each user. The historical sequence could help to train dialogue models that imitate the speaking style of specific users.

In Table 5, we show the statistics of the Pchatbot dataset. We find that the number of users who comment (23,408,367) is significantly larger than those who post (772,002) in PchatbotW. However, in PchatbotL, the number of users who comment (203,636) is much smaller than the number of users who post (5,203,345). We attribute this to the differences between the two platforms. Social media users are more willing to engage in interactions, while users from judicial forums tend to ask legal questions. Besides, the number of lawyers who answer legal questions in judicial forums is limited.

The scale of the Pchatbot dataset significantly outperforms previous Chinese datasets for dialogue generation. To be concrete, PchatbotW contains 5,319,596 posts and more than 139 million (post, response) pairs. PchatbotL contains 20,145,956 posts and more than 59 million (post, response) pairs. The largest dataset before has only less than 10 million (post, response) pairs. With such scales, performance improvement for data-driven neural dialogue models can be almost guaranteed. Pchatbot dataset provides sufficient valid responses as ground-truth for a post. On average, each post has 26 responses in PchatbotW. This helps to establish dialogue assessment indicators at the discourse level.

4 APPLICATION AND ANALYSIS

The Pchatbot dataset can be used in a wide range of applications of the dialogue system. In this section, we conduct preliminary studies on the effectiveness of the dataset. We first benchmark state-of-the-art models over Pchatbot for comparison in future work. We also investigate the effectiveness of the scale of training data and the length of dialogue history, respectively.

4.1 Settings and Evaluation Metrics

In our benchmark experiments, we keep the parameter setting the same as described in the corresponding papers except that we replace the word embeddings with ours. We will release codes for benchmark models on the release page. We also provide pre-trained language models including GloVe [18], BPE [22], Fasttext [1] which are trained on the dataset. These pre-trained models can be downloaded on the release page.

For retrieval-based dialogue models, we use **R_n@k** (recall at position k in n candidates) and **MRR** (Mean Reciprocal Rank) to measure the model's ability to select a personalized response from all candidates. For generation-based dialogue models, we use the **BLEU** [17] metric which is widely used to evaluate the model-generated responses. Besides, we use **Distinct-1/2** proposed in [11] to evaluate the diversity of responses generated by the model. To measure the personality consistency of generation-based models, we use **P-F1** as an evaluation metric [13].

4.2 Benchmark Models

We experiment with the following models[5]:

(1) Retrieval-based models: **Conv-KNRM** [6]: Single-turn dialogue model that uses CNN to capture n-gram features; **DAM** [35]: Multi-turn dialogue model that takes the user's dialogue history as context to construct multi-level text segment representations with stacked self-attention; **IOI** [26]: Multi-turn dialogue model that captures deep interactive matching features between response and utterance in the conversation context; **RSM-DCK** [9]: Knowledge enhanced multi-turn dialogue model that considers persona descriptions as external knowledge.

[5]We will continue updating other benchmark results on the project page.

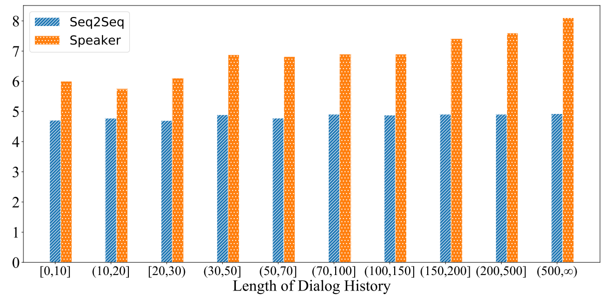

Figure 4: Effect of the length of dialogue history.

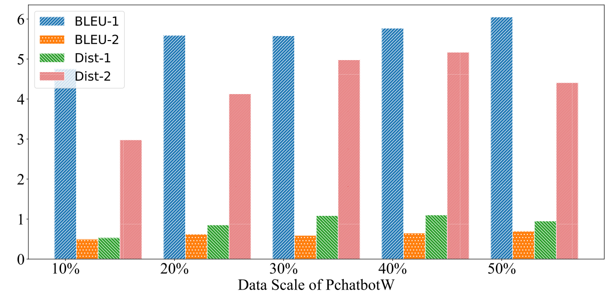

Figure 5: Effect of data scale.

(2) Generation-based models: **Seq2Seq** [15, 25]: Single-turn dialogue model that uses RNN-based encoder-decoder to generate responses; **Speaker Model** [12]: Single-turn personalized dialogue model that utilizes user IDs to learn user embeddings; **PersonaWAE** [4]: Single-turn personalized dialogue model that samples the personalized vector from a personalized Gaussian mixture distribution and uses it to guide the response generation; **DialoGPT** [33]: Large-scale dialogue model that concatenates the utterances as a long sequence and learns to generate the response.

4.3 Benchmark Experiments

4.3.1 Results. Table 4 shows the experimental results. For the retrieval-based models, we have the following findings: (1) single-turn model (Conv-KNRM) performs worse than others. The potential reason is the lack of conversational context which refers to the user's dialogue history in our scenario; (2) multi-turn model (DAM and IOI) performs better than others illustrating that using user's dialogue history as context can effectively booster the performance of ad-hoc matching; (3) for the knowledge-enhanced model (RSM-DCK), following Zhang et al. [33], we apply heuristic methods to generate external knowledge. But the effectiveness of such external knowledge is limited, which enlightens the need to directly model the user's personality from the user's dialogue history.

Generation-based models lead to similar findings: (1) the personalized models (Speaker and PersonaWAE) significantly outperforms the Seq2Seq model regarding Dist-1, Dist-2, and P-F1, which demonstrates that modeling user indeed leads to generating informative and personalized responses; (2) large-scale model (DialoGPT) obtains the best results for its great generalization ability brought by large-scale parameters. However, it performs worse than PersonaWAE regarding P-F1.

In general, using the user's dialogue history as context indeed improves dialogue models' performances especially for personalized metrics. Besides, personalized models that use either explicit user profiles or user embeddings fail to fully explore the personalized information that is hidden under the user's dialogue history. Pchatbot dataset enables designing models that can directly learn implicit user profiles. Such models are more useful regarding personalized chatbots, especially in a practical scenario.

4.3.2 Impact of the Length of Dialogue History. We evaluate the quality of responses generated for users with different lengths of dialogue history. For personalized chatbots, we expect the chatbots generate responses that are close to the original responses. Thus, we choose BLEU as the key indicator. We conduct experiments on a personalized dialogue model (Speaker) and a non-personalized chatbot model (Seq2Seq). The results are shown in Figure 4. From the figure we can find that: (1) with more dialogue history, the two models' performances continue to increase; (2) the discrepancy of BLEU scores between the two models gradually increases. In other words, personalized dialogue models can benefit more than non-personalized models.

4.3.3 Impact of the Scale of Data. We evaluate the effectiveness of dataset scale by conducting experiments on five subsets of different sizes using the Seq2Seq model. We construct these subsets by merging the partitions. Specifically, we use partition-1 as the smallest dataset and add partition-2 to partition-5 successively to construct bigger datasets.

Experimental results of incremental scale are shown in Figure 5. The results show that with the increasing of the training data size, the model's performance has a growing trend across all metrics. It confirms that using more training data helps to improve the model's effectiveness. Besides, we find that the diversity metrics(Distinct-K) turn goes down when using 50% of data. We attribute the phenomenon to that with more training data, the Seq2Seq model prefers to generate similar and generic responses [12].

5 CONCLUSION AND FUTURE WORK

In this paper, we introduce the Pchatbot dataset that has two subsets from the open domain and judicial domain respectively, namely PchatbotW and PchatbotL. All posts and responses in Pchatbot are attached with anonymous user IDs as well as timestamps, which greatly broadens the potentialities of a personalized chatbot. Besides, the scale of the Pchatbot dataset is significantly larger than previous datasets and this further enhances the capacity of intelligent dialogue agents. We evaluate the Pchatbot dataset with several baseline models and experimental results demonstrate the great advantages triggered by user IDs and large scale. The Pchatbot dataset and corresponding codes can be publicly viewed at Github.

ACKNOWLEDGEMENTS

Zhicheng Dou is the corresponding author. This work was supported by National Natural Science Foundation of China No. 61872370 and No. 61832017, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Shandong Provincial Natural Science Foundation under Grant ZR2019ZD06.

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguistics* 5 (2017), 135–146. https://transacl.org/ojs/index.php/tacl/article/view/999

[2] Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning End-to-End Goal-Oriented Dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.* https://openreview.net/forum?id=S1Bb3D5gg

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[4] Zhangming Chan, Juntao Li, Xiaopeng Yang, Xiuying Chen, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Modeling Personalization in Continuous Space for Response Generation via Augmented Wasserstein Autoencoders. 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 1931–1940. https://doi.org/10.18653/v1/D19-1201

[5] Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical Variational Memory Network for Dialogue Generation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018.* 1653–1662. https://doi.org/10.1145/3178876.3186077

[6] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (Eds.). ACM, 126–134. https://doi.org/10.1145/3159652.3159659

[7] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, CMCL@ACL 2011, Portland, Oregon, USA, June 23, 2011.* 76–87. https://www.aclweb.org/anthology/W11-0609/

[8] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, Tutorial Abstracts*, Yoav Artzi and Jacob Eisenstein (Eds.). Association for Computational Linguistics, 2–7. https://doi.org/10.18653/v1/P18-5002

[9] Kai Hua, Zhiyuan Feng, Chongyang Tao, Rui Yan, and Lu Zhang. 2020. Learning to Detect Relevant Contexts and Knowledge for Response Selection in Retrieval-based Dialogue Systems. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 525–534. https://doi.org/10.1145/3340531.3411967

[10] Marie-Anne Lachaux, Armand Joulin, and Guillaume Lample. 2020. Target Conditioning for One-to-Many Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 2853–2862. https://doi.org/10.18653/v1/2020.findings-emnlp.256

[11] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016.* 110–119. https://www.aclweb.org/anthology/N16-1014/

[12] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.* https://www.aclweb.org/anthology/P16-1094/

[13] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to Select Knowledge for Response Generation in Dialog Systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 5081–5087. https://doi.org/10.24963/ijcai.2019/706

[14] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic.* 285–294. https://www.aclweb.org/anthology/W15-4640/

[15] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015.* 1412–1421. https://www.aclweb.org/anthology/D15-1166/

[16] 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, Brussels, Belgium, October 31 - November 4, 2018.* 2775–2779. https://www.aclweb.org/anthology/D18-1298/

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.* 311–318. https://www.aclweb.org/anthology/P02-1040/

[18] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL.* 1532–1543. https://www.aclweb.org/anthology/D14-1162/

[19] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.* 4279–4285. https://doi.org/10.24963/ijcai.2018/595

[20] Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. 2019. Are Training Samples Correlated? Learning to Generate Dialogue Responses with Multiple References. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 3826–3835. https://doi.org/10.18653/v1/p19-1372

[21] Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA.* 172–180. https://www.aclweb.org/anthology/N10-1020/

[22] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.* https://www.aclweb.org/anthology/P16-1162/

[23] 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers.* 1577–1586. https://www.aclweb.org/anthology/P15-1152/

[24] 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 *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015.* 196–205. https://www.aclweb.org/anthology/N15-1020/

[25] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada.* 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks

[26] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 1–11. https://doi.org/10.18653/v1/p19-1001

[27] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *CoRR* abs/1506.05869 (2015). arXiv:1506.05869 http://arxiv.org/abs/1506.05869

[28] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A Dataset for Research on Short-Text Conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL.* 935–945. https://www.aclweb.org/anthology/D13-

1096/

[29] Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan W. Black. 2013. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference, The 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 22-24 August 2013, SUPELEC, Metz, France*. 404–413. https://www.aclweb.org/anthology/W13-4065/

[30] 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 496–505. https://doi.org/10.18653/v1/P17-1046

[31] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder P. Singh and Shaul Markovitch (Eds.). AAAI Press, 3351–3357. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14563

[32] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 2204–2213. https://doi.org/10.18653/v1/P18-1205

[33] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, Asli Çelikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, 270–278. https://doi.org/10.18653/v1/2020.acl-demos.30

[34] Tianyu Zhao and Tatsuya Kawahara. 2020. Multi-Referenced Training for Dialogue Response Generation. *CoRR* abs/2009.07117 (2020). arXiv:2009.07117 https://arxiv.org/abs/2009.07117

[35] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 1118–1127. https://doi.org/10.18653/v1/P18-1103

[36] Yutao Zhu, Zhicheng Dou, Jian-Yun Nie, and Ji-Rong Wen. 2020. ReBoost: a retrieval-boosted sequence-to-sequence model for neural response generation. *Inf. Retr. J.* 23, 1 (2020), 27–48. https://doi.org/10.1007/s10791-019-09364-x