A Large-Scale Chinese Short-Text Conversation Dataset

Yida Wang[1], Pei Ke[2], Yinhe Zheng[2,3], Kaili Huang[2], Yong Jiang[1], Xiaoyan Zhu[2], and Minlie Huang[2] [⋆]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2] Institute for Artifical Intelligence, State Key Lab of Intelligent Technology and Systems. Beijing National Research Center for Information Science and Technology. Department of Computer Science and Technology, Tsinghua University, Beijing, China.
[3] Samsung Research China - Beijing (SRC-B), Beijing, China.
{wangyd18,kp17,hkl16}@mails.tsinghua.edu.cn, yh.zheng@samsung.com, jiangy@sz.tsinghua.edu.cn, {zxy-dcs,aihuang}@tsinghua.edu.cn

Abstract. The advancements of neural dialogue generation models show promising results on modeling short-text conversations. However, training such models usually needs a large-scale high-quality dialogue corpus, which is hard to access. In this paper, we present a large-scale cleaned Chinese conversation dataset *LCCC*, which contains a base version (6.8 million dialogues) and a large version (12.0 million dialogues). The quality of our dataset is ensured by a rigorous data cleaning pipeline, which is built based on a set of rules and a classifier that is trained on manually annotated 110K dialogue pairs. We also release pre-training dialogue models which are trained on LCCC-base and LCCC-large respectively. The cleaned dataset and the pre-training models will facilitate the research of short-text conversation modeling. All the models and datasets are available at `https://github.com/thu-coai/CDial-GPT`.

Keywords: Deep learning · Dialogue generation · Dataset · Pre-training.

1   Introduction

The success of pre-training models has greatly advanced the research of natural language processing [21]. While BERT [5] promotes various natural language understanding tasks, GPT [23] demonstrates state-of-the-art performance on natural language generation. Recently, pre-training models have been applied to dialogue generation tasks and achieved state-of-the art results [38,3,1].

In addition to the effective Transformer-based model structures, dialogue corpora also play an important role in the success of open-domain dialogue generation models. Existing work has adopted massive English dialogue corpora from Twitter [24], Reddit [19], OpenSubtitles [14] and other public resources,

---

[⋆] Corresponding author

which equip the pre-training models with the ability to respond to human in open domains.

However, since there are few large-scale Chinese dialogue corpora, the development of the pre-training models for dialogue generation in Chinese has been hindered. We argue that it's essential to construct high-quality large-scale Chinese dialogue corpora which can further promote pre-training models in Chinese. The main challenge we need to face is the quality control of dialogue data. On the one hand, we cannot rely on crowd-sourcing services because existing work has shown that the amount of data used in the pre-training models is quite large [23]. On the other hand, if we mainly acquire data from public resources such as social media (Weibo, Twitter, Reddit, etc.), the quality of the data will be hard to ensure. As a matter of fact, online social media contains many negative behaviors, including toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion)[4], threats, insults, identity hates, obscene contents, and many more. These factors can remarkably degrade the generation ability of dialogue generation models, and lead to serious unexpected behaviors, which substantially limit the practical use of dialogue models.

In this work, we construct a large-scale cleaned Chinese conversation dataset called LCCC, which contains two versions, LCCC-base and LCCC-large. LCCC-base is filtered from 79 million conversations crawled from Weibo, while LCCC-large is filtered from the combination of Weibo data and other sources of Chinese corpora. A two-phase pipeline is designed to clean this corpus. Specifically, the first phase utilizes a set of heuristic rules to filter out dialogues with inappropriate content, and the second phase employs several classifiers that are trained on manually labeled data to filter dialogues further. Then we present pre-training models for dialogue generation, which are first pre-trained on a Chinese novel dataset and then post-trained on LCCC. All the pre-trained models and the datasets are released to facilitate future research.

Our contributions can be summarized as below:

– We build a large-scale cleaned Chinese conversation dataset called LCCC. It can serve as a benchmark for the study of open-domain conversation generation in Chinese.
– We present pre-training models for Chinese dialogue generation. Moreover, we conduct experiments to show its performance on Chinese dialogue generation. Both the models and the data are released for public use.

2   Related work

Datasets We make a brief overview of datasets available for data-driven conversation systems, most of which are constructed based on public resources or crowd-sourcing.

---

[4] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

As an important public resource of dialogue corpora, movie scripts such as OpenSubtitles have been used to construct dialogue datasets [14,26] . The dialogue content usually depends on the scenes of movies. Another public resource is social media, where Twitter [11,30], Reddit [19], Weibo [33,28], and technical forums [17] have been adopted to build open-domain dialogue datasets. These datasets crawled from public resources are usually at a large scale because the corpora of public resources is abundant. However, they also contain much noise which need to be carefully cleaned.

Other researchers resort to construct high-quality dialogue datesets by crowdsourcing. These datasets are built for advanced dialogue tasks, such as wizard of wikipedia (WoW) [6] / document grounded conversations (DoG) [40] for knowledge grounded dialogue generation, PERSONA-CHAT [37] for persona enhanced dialogue generation, and DailyDialog [13] for emotional conversation generation. These datasets are commonly cleaner than those acquired from public resources, but the amount of these datasets is rather small.

Pre-training Models on Dialogue Generation Since GPT [22] achieves state-of-the-art performance on various text generation tasks, recent work has applied pre-training models to dialogue generation. DialoGPT [38] presents an English open-domain pre-training model which post-trains GPT-2 [23] on 147M Reddit conversations. Meena [1] trains an Evolved Transformer [29] with 2.6B parameters on a massive English social media conversation dataset, which contains 40B words. [12] investigates dialogue generation by fine-tuning a Chinese GPT on some small dialogue datasets, where the Chinese GPT is pre-trained on a Chinese corpus mixed with Chinese Wikipedia2 (1.7B words) and Chinese News (9.2B words).

3   Datasets

We crawled 79M conversations from Weibo. First of all, through a rigorous cleaning process, a cleaned Weibo dataset (LCCC-base) was constructed. Then, the 79M conversations were mixed with the several public Chinese conversation datasets, and a larger Chinese conversation dataset (LCCC-large) was obtained via more relaxed cleaning conditions. The cleaning process includes rule-based and classifier-based filtering.

3.1   Data Collection

LCCC-base A two-phase data collection scheme is used to construct our raw dialogues. At the first phase, a set of seed users were collected. Specifically, we manually selected a batch of Weibo accounts who follow professional mass media dedicating to publish news. We then regarded the users who post comments under these news as "high-quality" users, since robot accounts usually do not pay much attention to these daily news.

Type	Case
Platform tag	**回复@精灵小宝贝**：我也失眠了
Advertisement	@张伟丽MMA 前来为**DW持妆粉底液**实力证言！
Generic form	我也是 我也是 我也是 啊啊
Dirty word	被**小婊砸**作的现在满身负能量
Special word	可以试试**左氧氟沙星**
Name	**陈绍龙**、你无朋友架啦
Symbol	#(ﾟ-ﾟ)#
Platform sign	文科574报哪里好?最好有师范英语**【微信】**
Not fluent	Q: 昨晚失眠了 A: 发恶梦扎醒,又发我最怕个蒋尸梦!
Incomplete information	Q: 江南小镇很美呀 A:**印象**
Time-sensitive	Q: 感觉没啥电视好看了 A: 琅琊榜，**就要大结局了**
External noun	Q: 假期要过了 A: **春节**一过,好想立刻回到岛城
Irrelevant pairs	Q: 差点吧洗面奶当牙膏 A: 绿色是今年的流行色

Table 1. Cases of noise in the cleaning process.

At the second phase, we collected dialogues from these seed users. Specifically, the Weibo posts from these users are collected along with the following comments, which are organized in tree structures. Note that any path from a root to a leaf can be regarded as a conversation session. We reconstructed these sessions using a Depth First Search process, and 79 million sessions of raw conversations were constructed. We then constructed a cleaned Weibo dataset using the cleaning method described in Section 3.2.

LCCC-large We collected corpora from multiple open-source repositories, including Chinese Chatterbot Corpus [5], PTT Gossiping Corpus [6], Subtitle Corpus and Xiaohuangji Corpus [7]. These datasets, together with Qingyun Corpus and Tieba Corpus, are cleaned and processed to be single-turn conversational data [8]. Besides, we collected multi-turn conversational data including Douban Conversation Corpus [9], E-commerical Conversation Corpus [10] and a Chinese chat corpus [11].

We then mixed these datasets with the 79M conversations. Using the same cleaning process, but by relaxing the threshold of the classifier described below, we obtained a larger version of our dataset (LCCC-large).

3.2 Cleaning Process

Rule-based Noise Filtering We filter out many types of noise via rules including: (1) delete the platform tag in the dialogues, such as "Reply to @***",

---

[5] https://github.com/gunthercox/chatterbot-corpus

[6] https://github.com/zake7749/Gossiping-Chinese-Corpus

[7] https://github.com/skdjfla/dgk_lost_conv

[8] https://github.com/codemayq/chinese_chatbot_corpus

[9] https://github.com/MarkWuNLP/MultiTurnResponseSelection

[10] https://github.com/cooelf/DeepUtteranceAggregation

[11] https://github.com/yangjianxin1/GPT2-chitchat

"[dog]"; (2) remove URL strings from the text; (3) split conversations with more than 30 turns into multiple conversations less than 30 turns [28]; (4) only keep one copy of the phrases or words that are repeated more than 6 times in one sentence; (5) remove the dialogues if the response is too long or too short; (6) remove the dialogues if the response is identified as an advertisement by the method in [33]; (7) remove the dialogues if 90% of tri-grams in the response are high-frequency tri-grams [38]; (8) remove the dialogues if the response has some specific forms of generic responses; (9) remove the dialogues in which the response is the same as the post.

We also construct blacklists containing the following noise: (1) dirty words, sensitive words, and dialect; (2) special topics words such as levofloxacin; (3) name, appellation and unknown abbreviation; (4) special symbols and emoji; (5) platform signs such as ads, pictures, and videos related words. The dialogue will be removed if it contains words that appear in the blacklist. Some cases are shown in Tabel 1 (top).

Classifier-based Filtering In addition to the rule-based method, we also introduce classifier-based filtering. Many types of noise in terms of semantics and grammar, and some context-dependent conversations[12] are hard to be filtered with rules. So we built two BERT classifiers for more elaborate filtering. We evaluated precision, recall, and F-score with different confidence scores to choose the best confidence threshold.

The first BERT classifier was trained on manually labeled 100,000 conversations. A dialogue is labeled noisy if it has the above noise or following noise: (1) The response is not fluent or there are serious typos in the sentence, (2) The information of the response is incomplete; (3) The topic of dialogue is time-sensitive, (4) Festivals, places, gender and time which are not mentioned in the post appear in the response (5) The post and the response are irrelevant. Some cases are shown in Tabel 1 (bottom). The classification accuracy reaches 73.76% on the test set.

In social media, many conversations inevitably depend on external contexts beyond the text [27,33] making them hard to understand [9]. To alleviate this problem, the second BERT classifier was trained on a manually labeled dataset containing 10,000 utterances. The classification accuracy reaches 77.60% on the test set.

3.3   Statistics and Results

The statistics about the dataset are shown in Table 2. The *Avg.words* means the average number of words per utterance, and text is tokenized by Jieba [13]. We also estimated the noise level in the STC dataset via our blacklist. Results show that 60% of conversations in STC contain dirty words, sensitive words, special symbols, etc. The model trained on STC generates five times more blacklist

---

[12] The understanding of the first post depends on other context beyond the post.
[13] https://github.com/fxsjy/jieba

	Single-Turn	Multi-Turn	Single-Turn	Multi-Turn
Raw dialogs	52,708,955	26,749,365	63,251,887	28,189,952
Cleaned dialogs	3,354,277	3,466,278	7,273,804	4,733,955
Utterances	6,708,554	13,365,268	14,547,608	18,341,167
Characters	68,559,727	163,690,614	162,301,556	217,776,649
Vocabulary size	372,063	666,931	662,514	690,027
Avg. words	6.79	8.32	7.45	8.14
Avg. turns	2	3.86	2	3.87

Table 2. Statistics of LCCC-base (left) and LCCC-large (right).

Dataset	Corpus Statistics	Source	Topic	Corpus Features
DuConv [35]	29,858 dialogs 9.1 turns per dialog 10.6 words per turn	Crowd source	Films and film stars	Knowledge-grounded/ Proactivity modeling
Douban [36]	0.5M dialogs 7.69 turns per dialog 18.56 words per turn	Douban	Open topics	0.5M Negative dialogs
Persona-Dialog [39]	20.83M dialogs 56.26M utterances 8.47M user profiles	Weibo	Open topics	Personalization,rich user profiles
STC [28]	4.4M pairs 219,905 posts 4.3M responses	Weibo	Open topics	One post multiple responses
LCCC$_{base}$	6.8M dialogs 2.95 turns per dialog 20M utterances	Weibo	Open topics	Extremely Strict cleaning process
LCCC$_{large}$	12M dialogs 2.74 turns per dialog 33M utterances	Mixup	Open topics	Strict cleaning process

Table 3. Comparison between existing Chinese conversation datasets and LCCC.

words than that trained on LCCC. In Table 3, a clear comparison between the existing Chinese dialogue dataset and our dataset is presented.

4 Models

Architectures The model architecture used in this paper is adopted from GPT [22] which is based on transformer [32] . The transformer decoder consists of multiple masked multi-head self-attention blocks. In each time step, the self-attention mechanism can only observe the information on the left.

Given a golden response $y = (y^1, ..., y^L)$ and history utterances of a conversation $U = \{u_0, ..., u_n\}$ where each utterance $u_i = (u_i^1, ..., u_i^{L_i})$ consists of L_i words, our goal is train a generation model via maximum likelihood estimation (MLE) $P(y|U) = \prod_{j=1}^{L} P(y^j|y^1, ..., y^{j-1}, U)$ to generate u_{n+1}^j given U with generated $u_{n+1}^1, ..., u_{n+1}^{j-1}$ until the whole response is generated with one termination symbol.

Input Representation We concatenate all history utterances into one sequence as a long text described in [34]. The input of the model is the sum of word embedding, speaker embedding, and position embedding. The word embedding and position embedding are learned during the pre-training phase, and speaker embedding is learned during the post-training or fine-tuning phase. The speaker embedding is used to indicate the different speakers, and we use the speaker symbol as the separation tokens. Following BERT [5], we also add a $[CLS]$ to the start of the sequence and use $[SEP]$ as the end-of-sequence token.

Training Following the work of DialoGPT [38], our models are post-trained based on a pre-trained Chinese GPT model (GPT_{Novel}) on the conversation dataset we collected. For multi-turn dialogue instances, following DialoGPT, We take every sentence in the dialogue from the second sentence to the last sentence as the response of history sentences. We trained several models on the LCCC-base and LCCC-large, respectively (LCCC-base is cleaner), which are summarized as follows:

- GPT_{Novel} is a 12-layer GPT which is pre-trained for 70 epochs on a Chinese novel dataset consisting of a variety of genres (Comedy, Romance, Mystery) with about 0.5 billion tokens.
- $CDialGPT_{LCCC-base}$ is a 12-layer GPT which is pre-trained for 70 epochs on the Chinese novel dataset and post-trained for 30 epochs on LCCC-base.
- $CDialGPT2_{LCCC-base}$ is a 12-layer GPT2 which is pre-trained for 70 epochs on the Chinese novel dataset and post-trained for 30 epochs on LCCC-base.
- $CDialGPT_{LCCC-large}$ is a 12-layer GPT which is pre-trained for 70 epochs on the Chinese novel dataset and post-trained for 30 epochs on LCCC-large.

All models were optimized by the AdamW [16] optimizer and the Noam [22] learning rate decay method. The layers of all our models are set to 12, and the number of heads is set to 12. The dimension of word embedding is set to 768, and the dimension of position embedding is set to 513. The number of warmup epoch was set to 1, and the maximum learning rate was 6.25e-5. The batch size was set to 8, and the number of gradient accumulation was set to 64. All the models trained with a vocabulary of 13,088 Chinese characters on four RTX 2080ti GPUs. Our implementation is based on the open-source code of Transformers [14] and TransferTransfo [15]

5 Experiments

5.1 Fine-tuning Settings and Implementation Details

To evaluate the models, we fine-tune all our models mentioned in Section 4 on the STC dataset. The STC dataset contains 4.4M conversation pairs. We randomly split the dataset into train, validation, and test sets. The validation set and the test set have 20,000 conversation pairs each without any intersection.

---

[14] https://github.com/huggingface/transformers
[15] https://github.com/huggingface/transfer-learning-conv-ai

(a) An interactive example of multi-turn dialogue generation.

Role	Response
User	别爱我没结果
Bot	没结果的爱是多么的痛 。
User	但是爱过就可以了
Bot	有时候，爱情就是一场游戏。
User	输的总是我
Bot	我爱你，就像我爱你一样。
User	你可拉到吧
Bot	你可拉到吧。
User	别学我说话
Bot	我说话就是你不行。

(b) An example of multi-turn self-playing dialogue generation.

Role	Response
User	生命的意义是什么
Bot1	生命的意义是过得比你想象的要好
Bot2	也许吧
Bot1	嗯，生命的意义是过得比你你想象的要好
Bot2	说的好
Bot1	好什么
Bot2	过得比你想象的好
Bot1	好吧

Table 4. Two generated examples. In the right table, the user prompt is the initial input (the first row). Both of the bots are the same models (CDialGPT$_{\text{LCCC}-\text{large}}$).

Baselines First we choose the only Chinese dialogue pre-training model **GPT2-chitchat** [16] as one baseline which is pre-trained on 500k Chinese conversations based on the GPT2 [23] model. Then we choose the the vanilla **transformer** [32] as one of the baselines. This model is widely applied in machine translation and conversation generation [19,37]. We also evaluate the classical sequence-to-sequence [4,31] model with attention mechanism [2] (**Attn-Seq2Seq**) which has been widely used in conversation generation.

Training Details The GPT$_{\text{Novel}}$ is fine-tuned up to 30 epochs. All the other models are fine-tuned up to 10 epochs with the same batch size and number of gradient accumulation. The other hyper-parameters are also kept the same as the post-training process. The transformer and Attn-Seq2Seq are trained until convergence on STC without pre-training. For a fair comparison, Attn-Seq2Seq consists of 6 layers of GRU with Bahdanau [2] attention mechanism, and the dimension of hidden states is set to 768. The layers of the transformer are also set to 6. Thus, these two models have comparable parameters with pre-training models. We set the parameters of GPT2-chitchat to be the same with the original repository. For all the models, we linearly decay the learning rate from 6.25e-5 to 0 using the AdamW optimizer.

Case Study Some cases generated by CDialGPT$_{\text{LCCC}-\text{large}}$ which is fine-tuned on STC are provided in Table 5. These samples suggest that the model has the ability to generate informative responses. Following the work of DialoGPT [38], we provide an interactive dialogue sample and a self-playing dialogue sample in Tabel 3(a) and Tabel 3(b). All the above samples are decoded by top-p sampling [8] with temperature 0.7.

---

[16] https://github.com/yangjianxin1/GPT2-chitchat

Post	Response
作为一名伟大的吃货你必须知道的事情。	大半夜的不睡觉，你是在诱惑我么？
惊现特种喵！这身手。可以维护世界和平了吧。	这猫太有才了，这猫很有想象力啊！
超萌的考拉。不怕痒，挠吧。	我的小心脏要跳出来了
你觉得最好看的英剧Top3是什么？	我的英剧top3是《老婆婆》
无论身在何方，我心里满满都是你！	此时此刻，我只想说一句：好可爱！
安东尼刷屏，可惜小手太小，就是投不进啊	这小子太牛了，真想给他一拳。
披萨饼盒瞬间变笔记本支架，这个太酷了！	这个真的可以有。但是不是很实用
无论你多么讨厌你的学校，当你离开久了，你还是会想念它。	还是学校好，没有负担，没有压力

Table 5. Cases generated by CDialGPT$_{\mathrm{LCCC-large}}$ fine-tuned on STC.

5.2 Evaluation

Models	Size	PPL	BLEU-2	BLEU-4	Dist-1	Dist-2	Greedy Matching	Embedding Average
Attn − Seq2Seq	73M	34.2	3.93	0.90	0.0085	0.1191	0.6584	0.8338
Transformer	113M	22.10	**6.72**	3.14	0.0088	0.1397	0.6606	0.8355
GPT2 − chitchat	88M	−	2.28	0.54	**0.0103**	**0.1625**	0.6154	0.7894
GPT$_{\mathrm{Novel}}$	104M	21.27	5.96	2.71	0.0080	0.1172	0.6612	0.8334
CDialGPT$_{\mathrm{LCCC-base}}$	104M	18.38	6.48	3.08	0.0083	0.1268	0.6621	0.8354
CDialGPT2$_{\mathrm{LCCC-base}}$	104M	22.76	5.69	2.50	0.0077	0.1087	0.6624	0.8346
CDialGPT$_{\mathrm{LCCC-large}}$	104M	**18.23**	6.63	**3.20**	0.0083	0.1271	**0.6625**	**0.8363**

Table 6. Automatic results.

Automatic Metric We first employed BLEU [20] and distinct n-grams [10] as our automatic metrics. Since BLEU cannot perfectly reflect the quality of generated results [15], we adopted Greedy Matching [25] to evaluate the relevance between posts and generated responses at the word level and Embedding Average [15] at the sentence level. We also present the perplexity [18] of different models except GPT2-chichat which has different vocabularies with others.

As shown in Table 6, the models trained on LCCC achieves the best perplexity (PPL) but worse BLEU-2 scores than Transformer, which is consistent with previous work [38,12]. [1] shows a strong correlation between perplexity and human evaluation, so our model has competitive performance. The models trained on LCCC also outperforms others in Greedy Matching and Embedding Average. The GPT2-chitchat reaches the highest distinct scores but poor generation quality where we attribute it to the small scale of the model.

Human Evaluation We also employed human evaluation to assess the performance of the generation models. 200 samples of each model (total 1000 samples) are randomly presented to 3 judges to evaluate the following aspects based on a 2/1/0 score schema:

Models	+2	+1	+0	Score	Kappa
Seq2Seq	27.1%	21.4%	51.4%	0.756	0.4544
Transformer	42.4%	23.6%	34.0%	1.084	0.4429
GPT2 − chitchats	24.3%	19,1%	56.6%	0.677	0.3941
CDialGPT$_{\mathrm{LCCC-base}}$	46.7%	21.8%	31.5%	1.152	0.3954
CDialGPT$_{\mathrm{LCCC-large}}$	**48.6%**	**24.5%**	**27.0%**	**1.217**	0.4203

Table 7. Human evaluation.

- **Fluency and Relevance** If a response is grammatically correct, logically consistent, and relevant to the post, it will get 1. Otherwise, it will get 0.
- **Informativeness** If a response is fluent, relevant, and additionally rich in content, it will get 2.

The results are shown in Table 7. We can see that the models trained on LCCC outperform others. Although CDialGPT$_{\mathrm{LCCC-base}}$ is not as good as transformer in automatic metrics, CDialGPT$_{\mathrm{LCCC-base}}$ performs slightly better than *Transformer* in manual evaluation. And we computed Fleiss kappa [7] to measure the annotation agreement of crowd workers. The values range within 0.39-0.44 indicating fair-moderate agreement.

6   Conclusion

This paper presents a large-scale cleaned Chinese conversation dataset that is elaborately cleaned by our filtering pipeline. We provide two versions: one is the base version with 6.8M conversations, and the other is a larger version with 12M conversations. To obtain high-quality dialogue data, we design rule-based and classifier-based filtering procedures. We also present pre-training models for Chinese dialogue generation, which is trained on the 12M open-domain conversations. All our pre-training models and the dataset are released for public use.

7   Acknowledgments

References

1. Adiwardana, D., Luong, M.T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al.: Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977 (2020)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

3. Bao, S., He, H., Wang, F., Wu, H.: PLATO: pre-trained dialogue generation model with discrete latent variable. CoRR **abs/1910.07931** (2019)
4. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of wikipedia: Knowledge-powered conversational agents. arXiv preprint arXiv:1811.01241 (2018)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
8. Holtzman, A., Buys, J., Forbes, M., Choi, Y.: The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019)
9. Li, C., Wu, Y., Wu, W., Xing, C., Li, Z., Zhou, M.: Detecting context dependent messages in a conversational environment. arXiv preprint arXiv:1611.00483 (2016)
10. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)
11. Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., Dolan, B.: A persona-based neural conversation model. arXiv preprint arXiv:1603.06155 (2016)
12. Li, P.: An empirical investigation of pre-trained transformer language models for open-domain dialogue generation. arXiv preprint arXiv:2003.04195 (2020)
13. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manually labelled multi-turn dialogue dataset. arXiv preprint arXiv:1710.03957 (2017)
14. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles (2016)
15. Liu, C.W., Lowe, R., Serban, I.V., Noseworthy, M., Charlin, L., Pineau, J.: How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023 (2016)
16. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
17. Lowe, R., Pow, N., Serban, I., Pineau, J.: The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909 (2015)
18. Manning, C.D., Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT press (1999)
19. Mazaré, P.E., Humeau, S., Raison, M., Bordes, A.: Training millions of personalized dialogue agents. arXiv preprint arXiv:1809.01984 (2018)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
21. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: A survey. arXiv preprint arXiv:1905.05709 (2020)
22. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf (2018)
23. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8) (2019)

24. Ritter, A., Cherry, C., Dolan, B.: Unsupervised modeling of twitter conversations. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 172–180. Association for Computational Linguistics (2010)
25. Rus, V., Lintean, M.: An optimal assessment of natural language student input using word-to-word similarity metrics. In: International Conference on Intelligent Tutoring Systems. pp. 675–676. Springer (2012)
26. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Hierarchical neural network generative models for movie dialogues. arXiv preprint arXiv:1507.04808 **7**(8) (2015)
27. Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J.: A survey of available corpora for building data-driven dialogue systems. arXiv preprint arXiv:1512.05742 (2015)
28. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. arXiv preprint arXiv:1503.02364 (2015)
29. So, D.R., Liang, C., Le, Q.V.: The evolved transformer. arXiv preprint arXiv:1901.11117 (2019)
30. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, B.: A neural network approach to context-sensitive generation of conversational responses. arXiv preprint arXiv:1506.06714 (2015)
31. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
33. Wang, H., Lu, Z., Li, H., Chen, E.: A dataset for research on short-text conversations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 935–945 (2013)
34. Wolf, T., Sanh, V., Chaumond, J., Delangue, C.: Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149 (2019)
35. Wu, W., Guo, Z., Zhou, X., Wu, H., Zhang, X., Lian, R., Wang, H.: Proactive human-machine conversation with explicit conversation goals. arXiv preprint arXiv:1906.05572 (2019)
36. Wu, Y., Wu, W., Xing, C., Zhou, M., Li, Z.: Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. arXiv preprint arXiv:1612.01627 (2016)
37. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? arXiv preprint arXiv:1801.07243 (2018)
38. Zhang, Y., Sun, S., Galley, M., Chen, Y.C., Brockett, C., Gao, X., Gao, J., Liu, J., Dolan, B.: Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536 (2019)
39. Zheng, Y., Chen, G., Huang, M., Liu, S., Zhu, X.: Personalized dialogue generation with diversified traits. arXiv preprint arXiv:1901.09672 (2019)
40. Zhou, K., Prabhumoye, S., Black, A.W.: A dataset for document grounded conversations. arXiv preprint arXiv:1809.07358 (2018)