symmetry

MDPI

Article

An Empirical Study on Deep Neural Network Models for Chinese Dialogue Generation

Zhe Li [1,†], Mieradilijiang Maimaiti [2,†], Jiabao Sheng [3], Zunwang Ke [3], Wushour Silamu [4,*], Qinyong Wang [5] and Xiuhong Li [4]

[1] Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang Multilingual Information Technology Research Center, College of Software, Xinjiang University, Urumqi 830046, China; lizhe@stu.xju.edu.cn

[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China; meadljmm15@mails.tsinghua.edu.cn

[3] College of Software, Xinjiang University, Urumqi 830046, China; jiabao@stu.xju.edu.cn (J.S.); kzwang@xju.edu.cn (Z.K.)

[4] College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China; xjulxh@xju.edu.cn

[5] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4702, Australia; qinyong.wang@uq.edu.au

* Correspondence: wushour@xju.edu.cn

† Zhe Li and Mieradilijiang Maimaiti are contributed equally to this research.

Received: 27 September 2020; Accepted: 19 October 2020; Published: 23 October 2020

Abstract: The task of dialogue generation has attracted increasing attention due to its diverse downstream applications, such as question-answering systems and chatbots. Recently, the deep neural network (DNN)-based dialogue generation models have achieved superior performance against conventional models utilizing statistical machine learning methods. However, despite that an enormous number of state-of-the-art DNN-based models have been proposed, there lacks detailed empirical comparative analysis for them on the open Chinese corpus. As a result, relevant researchers and engineers might find it hard to get an intuitive understanding of the current research progress. To address this challenge, we conducted an empirical study for state-of-the-art DNN-based dialogue generation models in various Chinese corpora. Specifically, extensive experiments were performed on several well-known single-turn and multi-turn dialogue corpora, including KdConv, Weibo, and Douban, to evaluate a wide range of dialogue generation models that are based on the symmetrical architecture of Seq2Seq, RNNSearch, transformer, generative adversarial nets, and reinforcement learning respectively. Moreover, we paid special attention to the prevalent pre-trained model for the quality of dialogue generation. Their performances were evaluated by four widely-used metrics in this area: BLEU, pseudo, distinct, and rouge. Finally, we report a case study to show example responses generated by these models separately.

Keywords: natural language processing; dialogue generation; deep learning; network architecture; empirical investigation

1. Introduction

Text generation is a core task in artificial intelligence and natural language processing, which is essential for many popular downstream applications, such as the chitchat style dialogue system for human–machine conversations. Compared with traditional statistical machine learning methods [1,2], the methods based on deep neural networks (DNNs) [3–5] have achieved superior performance because they can effectively leverage massive data to capture high-level feature representations and automatically learn strategies for response generations while requiring minimum supervision. A variety of deep learning architectures, such as Seq2Seq, RNNSearch, transformer, generative adversarial nets, reinforcement learning, and pre-trained language models, have been utilized to achieve the state-of-the-art response generation performances on different datasets.

Specifically, the Seq2Seq-based models tend to generate highly generic responses, but they are frequently stuck in an infinite loop of repetitive responses. To increase the diversity of the generated responses and enable the conversation context-aware, the reinforcement learning (RL)-based models foster a more sustained dialogue and manage to produce more interactive responses; however, the stability of the quality of sentences generated by them is relatively low. On the other hand, GAN-based models outperform Seq2Seq models in the chatbot problem in many perspectives, but the results are still sub-optimal because they suffer from the gradient block problem in the discrete text space, one of the bottlenecks in all models employing GAN. Recently, utilizing pre-training language models to address all NLP-related problems, including the response generation task, emerged as the new state-of-the-art. They have been shown to produce responses that are closer to human beings [5,6]. However, they suffer from the problem that the final results are highly dependent on the pre-training corpus and decoding methods. Figure 1 shows different responses generated by the aforementioned DNN-based dialogue generation systems given an example question.

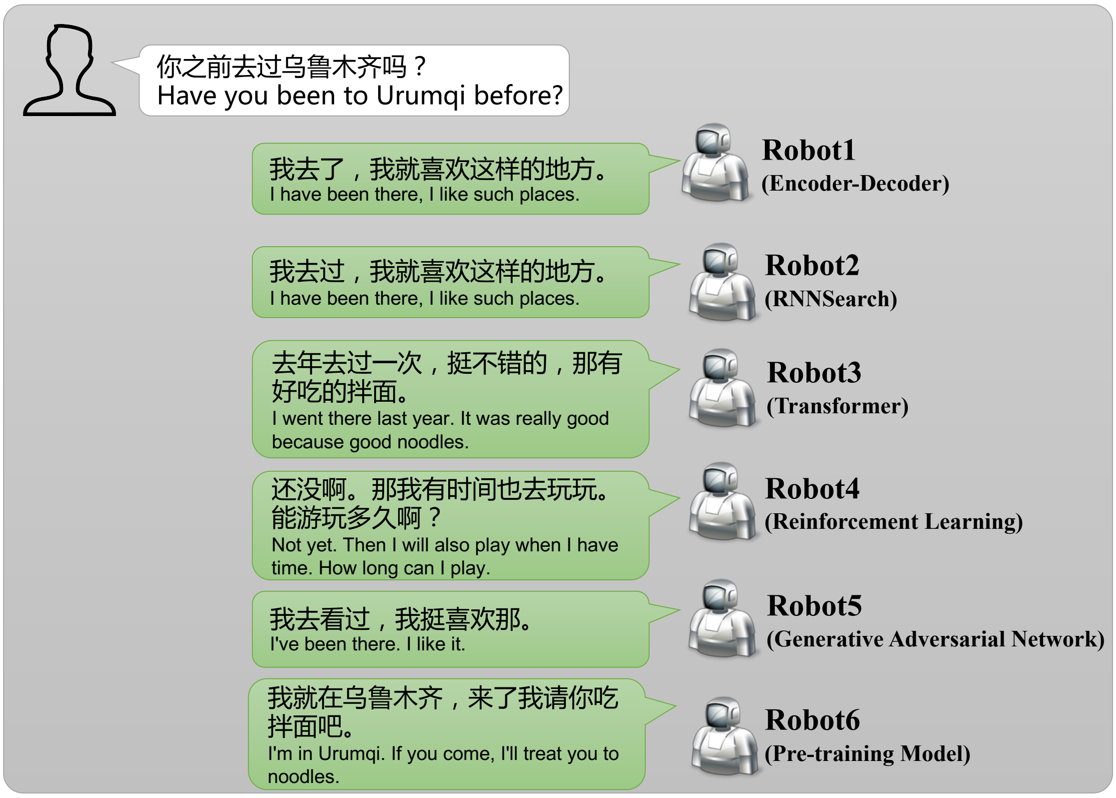
你之前去过乌鲁木齐吗？
Have you been to Urumqi before?
我去了，我就喜欢这样的地方。
I have been there, I like such places.
Robot1
(Encoder-Decoder)
我去过，我就喜欢这样的地方。
I have been there, I like such places.
Robot2
(RNNSearch)
去年去过一次，挺不错的，那有好吃的拌面。
I went there last year. It was really good because good noodles.
Robot3
(Transformer)
还没啊。那我有时间也去玩玩。能游玩多久啊？
Not yet. Then I will also play when I have time. How long can I play.
Robot4
(Reinforcement Learning)
我去看过，我挺喜欢那。
I've been there. I like it.
Robot5
(Generative Adversarial Network)
我就在乌鲁木齐，来了我请你吃拌面吧。
I'm in Urumqi. If you come, I'll treat you to noodles.
Robot6
(Pre-training Model)

Figure 1. The comparison between six systems with various architectures for same query.

Despite the dominant success of DNN-based dialogue generation systems shown in the literature, it is interesting to further compare the performances of the DNN structures on the real-world corpora under the same evaluation metrics and study the differences, which has been seldom explored. It is also important to both related researchers and software engineers because they need practical and detailed guidelines for the performances of different deep learning architectures and pre-trained language models on the task of Chinese dialogue generation. To fill this gap in open-domain dialogue generation, reference [7] did an empirical investigation on pre-trained transformer language models. However, they only evaluated the architectures of the pre-trained language model and Seq2Seq and did not comprehensively evaluate the effects of different architectures. Moreover, to the best of our knowledge, there is no empirical comparative study of different DNN structures on open Chinese corpora for the dialogue generation task.

To that end, in this paper, we present an empirical study on the performances and impacts of symmetrical DNN architectures (i.e., Seq2Seq, RNNSearch, transformer, generative adversarial nets, reinforcement learning, and pre-trained Language model) for the task of Chinese dialogue generation on three typical single-turn and multi-turn Chinese dialogue corpora: KdConv, Weibo, and Douban. We chose six specific and typical models using the aforementioned techniques, and they also have been verified to achieve superior performance in the literature. Detailed evaluation results based on four metrics, BLEU, rouge, perplexity, and distinct of the generated results, are reported. Furthermore, we analyze the results and try to disclose the mechanisms that cause the differences.

The contributions of this paper can be summarized as follows:

- We conducted extensive experiments to compare the performances of six typical DNN architectures for the dialogue generation task on three open Chinese corpora.
- We performed a case study to help understand their performances in an intuitive manner.
- We analyzed the mechanisms behind the different performances by these models and provide practical recommendations for model selection.

2. Background

In this section, we first introduce different DNN architectures dominating the dialog generation task recently.

2.1. Seq2Seq

T words and a target sequence $Y = (y_1, y_2, ..., y_{T'})$ whose length is T' consist of a source sequence $X = (x_1, x_2, ..., x_T)$, which make the generation probability of the model to be the maximized Y under the $X : p(y_1, y_2, ..., y_{T'}|x_1, x_2, ..., x_T)$ condition. In detail, the encoder-decoder framework is a structure of Seq2Seq [8,9]. The encoder word by word reads X, and uses a context vector c produced by RNN to speak with it. The decoder input estimates c to generate a probability of Y. The context vector c by the encoder RNN as follows

$$h_t = f(x_t, h_{t_1}) \tag{1}$$

In there, h_t is the hidden state and t (as f) is a non-linear function, such as instant LSTM and GRU. The hidden state counterpart to last word h_T is c. The decoder is an RNN model that has an additional conditional context vector c. The calculated that candidate words p_t probability distribution at every time is

$$s_t = f(y_{t-1}, s_{t-1}, c) \tag{2}$$

$$p_t = softmax(s_t, y_{t-1}) \tag{3}$$

where s_t is the hidden state of the decoder RNN at time t and y_{t1} is the word at time t_1 in the response sequence. The Seq2Seq objective function is defined as:

$$p(y_1, ..., y_{T'}|x_1, ..., x_T) = \prod_{t=1}^{T'} p(y_t|v, y_1, ..., y_{t-1}) \tag{4}$$

2.2. RNNSearch

To improve Seq2Seq performance, RNNSearch introduces the attention mechanism. Each word in y depends on a different context vector c, and it is observed that each word in y may be related to a different part of x. In particular, y_i corresponds to the context vector c_i, and c_i is the weighted average of the hidden state of the encoder $h_1, ..., h_t$:

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \tag{5}$$

where $a_{i,j}$ is computed by:

$$\alpha = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})} \tag{6}$$

$$e_{ij} = g(s_{t-1}, h_j) \tag{7}$$

where g is a multilayer perceptron.

2.3. Transformer

Transformer abandons the recurrent network structure of RNN and models a piece of text entirely based on attention mechanisms. The most important module of the coding unit is the self-attention module, which can be described as:

$$Attention(Q, K, V) = Softmax\left(\frac{Q\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{8}$$

To extend the ability of the model to focus on different locations and to increase the representation learning capacity of subspaces for attention units, transformer adopts the "multi-head" mode that can be expressed as:

$$MultiHead(Q, K, V\ ) = Concat(head_1, ..., head_h)W^O \tag{9}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^K) \tag{10}$$

2.4. Generative Adversarial Nets

Generative adversarial nets [8,10,11] aim to train two competing networks, one of which is to train a generator making the generated results as close to the real ones as possible, and the other aim is to train a discriminator which determines whether the data are from the distribution of the real or the distribution learned by the generator.

Suppose that a generation model is $G(z|\theta_g)$, where z is a random noise; G converts this random noise into data type x. Taking the text problem as an example, we can treat the output of G as a text and D as a discriminant model. For any input x, the output of $D(x|\theta_d)$ is a real number in the range of $[0.1]$, which is used to determine the probability of this text being a real text. Let p_r and p_g represent the distribution of real text and the distribution of generated text respectively. The objective function of the discriminant model is as follows:

$$max(V, D) = \mathbb{E}_{x \sim p_{data(x)}}[\log D(x)] + \mathbb{E}_{z \sim p_{(z)}}[\log(1 - D(G(z)))] \tag{11}$$

The goal of a similar generation model is to make the discrimination model unable to distinguish the real text from the generated text. The objective function of the generation model is expressed as:

$$min = \mathbb{E}_{z \sim p_{(z)}}[\log(1 - D(G(z)))] \tag{12}$$

Then the whole optimization objective function is described:

$$minmax(V, D) = \mathbb{E}_{x \sim p_{data(x)}}[\log D(x)] + \mathbb{E}_{z \sim p_{(z)}}[\log(1 - D(G(z)))] \tag{13}$$

2.5. Reinforcement Learning

We regard the generated text as actions that are taken through a policy defined by an encoder-decoder model. Using a policy search optimizes network parameters to maximize the expected future of the reward. In dialogue generation, the main methods of reinforcement learning are policy gradient methods and Q-learning.

Q-learning is the basic value-based algorithm. Assume Q to be the value function; then the optimal value of the Q-learning algorithm using the Bellman equation can be expressed:

$$Q^*(s,a) = (BQ^*)(s.a) \tag{14}$$

where the Bellman operator (B) can be described as:

$$(BK)(s,a) = \sum_{\acute{s}\in\mathcal{S}} T(s,a,\acute{s})(R(s,a,\acute{s})) + \gamma \max_{\acute{a}\in\mathcal{A}} K(\acute{s},\acute{a}) \tag{15}$$

On the other hand, policy gradient is meant to find a neural network parameterized policy in order to maximize the expected cumulative reward. The easiest approach to obtain the policy gradient estimator could be to utilize an algorithm, and one of the methods could be defined as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) Q_\pi(s,a)] \tag{16}$$

2.6. Pre-Training Language Model

A text sequence $X = (x_1, x_2, ..., x_T)$ and a sequence $Y = (y_1, y_2, ..., y_{T'})$ to be given to denote the sequence of input context and target response, respectively. The conditional probability $p(x_t|x_{0:t-1})$ modeled by a probability distribution over the vocabulary given linguistic context $x_{0:t-1}$. The context $x_{0:t-1}$ is modeled by neural encoder $f_{enc}(\cdot)$, and the conditional probability:

$$p(x_t|x_{0:t-1}) = g_{LM}\Big(f_{enc}(x_{0:t-1})\Big) \tag{17}$$

where $gLM(\cdot)$ is the prediction layer.

For a large corpus, we can use maximum likelihood estimation (MLE) to train the whole network. Firstly, X and Y are connected, and then the predicted loss of the whole target response sequence is obtained as the loss function. The loss term for predicting the dialogue context X is:

$$L_{LM} = -\sum_{t=1}^{T} \log p(x_t|x < t) \tag{18}$$

3. Related Work

In the field of natural language processing, dialogue generation has attracted more and more attention from researchers. The Meta-dialog system (MDS) worked on by [12] combines the advantages of both meta-learning approaches and human–machine collaboration; existing end-to-end dialog systems perform less effectively when data are lacking. Reference [13] proposes a novel dataset named multi-turn to facilitate the conversation reasoning research. We briefly reviewed three related approaches: generative adversarial nets, reinforcement learning, and pre-training language model. Existing techniques for building an open-domain dialogue system can be categorized into three groups.

It has been proven that the end-to-end task-oriented dialog system can have remarkable success in recent studies. Reference [14,15], based on the end-to-end framework, have achieved a productive performance in dialogue generation; Reference [16] improves the quality of the selected response for a retrieval-based dialog system based on end-to-end architecture. We introduce the dialogue of the end-to-end architecture from three structures: seq2seq, RnnSearch, and transfomer.

- **Seq2Seq**: The first group learns the response generation model under the simple Seq2Seq architecture. Reference [17] used multi-layer LSTM to map the input sequence to a fixed-dimensional vector and then used another deep-layer LSTM to decode the target sequence from the vector, with high automation and flexibility. Reference [18] introduced HRED, which uses a hierarchical codec architecture to model all contextual sentences. Reference [19] used an extended encoder-decoder that provides encoding of context and external knowledge.
- **RNNSearch**: The second group is based on the basic sequence-to-sequence with attention frame [20,21]. Reference [22] proposed a new multi-round dialogue generation model, which uses a self-attention mechanism to capture long-distance dependence—a weighted sequence(wise) attention model which uses cosine similarity to measure the degree of correlation proposed by [23] for HEED. Reference [24] introduced an attention mechanism into HRED, and proposed a new hierarchical recursive attention network (HRAN).
- **Transformer**: The last group at the peak of attention architectures with the transformer framework. Reference [25] proposed a NMT model via multi-head attention; others were inspired by this paper. Reference [26] proposed an incremental transformer with the deliberation decoder to solve the task of document grounded conversations. Reference [27] proposed a transformer-based model to address multi-turn unstructured text facts open-domain dialogue.

Generative adversarial nets: Reference [28] proposed a sequence generation method, SeqGAN, to effectively train generative adversarial nets for structured sequences generation via policy gradient. For poems and speech-language generation, SeqGAN showed excellent performance in generating the creative sequences. Reference [29] proposed a new generative adversarial nets (GAN) variant for dialog generation, introduced an approximate embedding layer in GAN, and added a discriminator with anti-function to improve the diversity of answers. Reference [30] solved the problems of either inconsistent personality across conversations or average personality of users by generating a controlling agent's persona upon, instead of conditioning on prior conversations of a target actor.

Reinforcement Learning: Reference [4] showed how to introduce deep reinforcement learning in chatbot dialogue to model future reward. Reference [31] presents a new approach to best utilize a fixed budget-conscious scheduling (BCS) small amount of user interactions (budget) for learning task-oriented dialogue agents to extend deep dyna-Q (DDQ). Reference [32] proposed a new model-based reinforcement learning approach, discriminative deep dyna-Q (D3Q), for task-completion dialogue policy learning. Reference [33] presented a new reinforcement learning framework, switch-based active deep dyna-Q (Switch-DDQ), for task-completion dialogue policy learning.

Pre-training language models: Recently, pre-training language models including GPT-2 [34] or BERT [35] in various tasks of NLP have achieved enormous success, including machine translation, text classification, summarization, question answering, etc. Academics are also working to assemble the language models for the task of dialogue generation. Typically, Reference [5] conducted some experimental analyses about the appearance of language models on dialogue generation. Reference [6] proposed a relevant promoting language model by incorporating a topic inference component into the language model to conduct diverse and informative dialogue generation. Reference [36] presented an end-to-end monolithic neural model for goal-oriented dialogues using GPT-2. Reference [37] proposed a new dialogue generation framework, which uses pre-training to support various conversations, such as chit-chat, knowledge-based dialogues, and conversational questions and answers.

4. Empirical Study

4.1. Datasets

The corpora used in our study were collected from openly available resources such as KdConv (https://github.com/thu-coai/KdConv), Weibo (https://ai.tencent.com/ailab/nlp/dialogue/#datasets), and Douban (https://ai.tencent.com/ailab/nlp/dialogue/#datasets) respectively.

- **KdConv:** KdConv is a dataset of Chinese multi-domain knowledge-driven transformation, which establishes topics in multiple rounds of dialogues on the knowledge graph. KdConv contains 4.5 K dialogues and 86 K utterances from three fields (film, music, and tourism), with an average number of rounds of 19.0. These dialogues on related topics include in-depth discussions and natural transitions between multiple topics. At the same time, the corpus can also be used for transfer learning and exploration of domain adaptation.
- **Weibo:** A single-turn open-domain Chinese dialogue dataset, which is originally collected and released by reference [38] from a Chinese social media platform Weibo https://www.weibo.com/. Here, we used a refined version released by reference [39] (https://ai.tencent.com/ailab/nlp/dialogue/#datasets). The number of samples for training, validation, and testing are 400 M, 19,357, and 3200 respectively. Character-based vocabulary size is 10,231.
- **Douban:** A multi-turn open-domain Chinese dialogue corpus collected from Douban group, a well-known Chinese online community which is also a common data source for dialogue systems [40]. Here, we utilize a version called Restoration-200 K dataset released by reference [41]. There are 193,769 samples for training, 5095 for validation, and 5104 for testing. Vocabulary size is 5800.

More detailed statistics about the dataset are summarized in Table 1.

Table 1. Statistics of the dialogue datasets.

Corpus	Type	Train	Dev	Test	Vocab	Avg Length
KdConv	Multi-Turn	3600	450	450	63,134	21.41
Weibo	Single-Turn	4,244,093	19,357	3200	10,231	34.21
Douban	Multi-Turn	193,769	5095	5104	5800	31.44

4.2. Evaluation Metrics

We adopt BLEU [42], distinct [43], rouge [44], and perplexity [3] as the evaluation metrics to measure the quality of the generated response. For BLEU, we employ the values of BLEU 1–4 and show the value of rouge-1/2/L. Intuitively, the higher the BLEU and rouge scores, the more n-gram overlaps between the generated responses, and thereby the better the performance. To be more specific, BLEU-N is formally defined as:

$$BLEU\text{-}N = \exp(min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n) \tag{19}$$

where PN represents the modified $N\text{-}gram$ precision, WN equals N, and R and C represent the lengths of the reference response and the prediction response respectively. Intuitively, a higher BLEU score means more n-gram overlaps between the comparative responses, thereby indicating better performance. At the same time, with the continuous change from unigram to 4-gram, BLEU-4 is applied more and more in machine translation and dialogue systems.

Rouge-N is formally defined as:

$$Rouge\text{-}N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)} \tag{20}$$

where n represents the length of $n\text{-}gram$, and $gram_n$ and $Count_{match}(gram_n)$ are the maximum numbers of $n\text{-}gram$ that occur simultaneously in both candidate summary and a set of reference summaries.

We used the values of 1–4 to evaluate the diversity of generated responses. Distinct-N is defined as:

$$Distinct(n) = \frac{Count(uniquengram)}{Count(word)} \tag{21}$$

$Count(uniquengram)$ represents the number of $ngrams$that are not repeated in the reply, and $Count(word)$ represents the total number of $ngram$ words in the reply. The larger the *Distinct-n*, the higher the diversity of the generation. Nevertheless, perplexity is a well-established performance metric for generative dialogue models.

On the other hand, perplexity explicitly measures the ability of the model to account for the syntactic structure of the dialogue, and the syntactic structure of each utterance and lower perplexity is indicative of a better model. We define word perplexity:

$$\exp(-\frac{1}{N_W}\sum_{n=1}^{N}\log P_\theta(U_1^n, U_2^n, U_3^n)) \tag{22}$$

For the model with the parameter θ, a dataset containing n triplets, $u_1^n, u_2^n, u_{3\,n=1}^{n\,n}$, and n_w refers to the number of tokens in the whole dataset. Lower confusion represents a better model. Confusion clearly measures the model's ability to explain the syntactic structure of dialogue and each discourse. In dialogue, the distribution of words in the next utterance is highly multi-modal; i.e., there are many possible answers, which makes confusion especially suitable because it always measures the probability of regenerating accurate reference utterances.

4.3. Comparison Models

In this section, we first briefly introduce the models that we aim to study.

- **HRED (the code implementations can be found on https://github.com/julianser/hed-dlg):** HRED [18] is a hierarchical RNN-based encoder-decoder constructed for the multi-turn dialogue generation tasks. In the dialogue, the encoder RNN maps each utterance to a discourse vector. High-level context neural networks track past utterances by iteratively processing each discourse vector. The next utterance prediction is implemented by RNN decoder, which obtains the hidden state of context RNN and generates probability distribution on the token of the next utterance.
- **ReCoSa (the code and data are available at https://github.com/zhanghainan/ReCoSa):** ReCoSa [22] is a model based on attention mechanism, Firstly, the word-level LSTM encoder is executed to get an initial representation of each context. The self-focus mechanism is then used to update both the context and the mask response representation. Finally, the weight of attention between each context and response representation is calculated and used for further decoding.
- **Guyu (the code and models are available at https://github.com/lipiji/Guyu):**Guyu [26] is a transformer-based auto-regressive model for the task of open-domain dialogue generation. Guyu conducts representation learning utilize masking multi-head self-attention as the core technical operation. Various decoding strategies are employed to conduct the response-text generation. Adam with Noam learning-rate decay strategy is employed to optimize the model parameters.
- **RL (the code and models are available at https://github.com/liuyuemaicha/Deep-Reinforcement-Learning-for-Dialogue-Generation-in-tensorflow):**This is a model that uses reinforcement learning to simulate the future returns in chat robot dialogues. The model simulates the conversation between two virtual agents and uses the policy gradient method to reward a sequence, which shows three useful conversation characteristics: large amount of information, strong consistency, and easy answer (related to forward-looking function).
- **GAN-AEL (the code and models are available at https://github.com/lan2720/GAN-AEL):** GAN-AEL [29] is a GAN framework to model single-turn short-text conversations. GAN-AEL

trains the Seq2Seq network to simultaneously perform a discriminant classifier, which measures the difference between the human-generated response and the machine-generated response and introduces an approximate embedding layer to solve the non-differentiable problem caused by sampling-based output decoding in the Seq2Seq generation model steps.

- **BigLM-24: (the code and models are available at https://github.com/lipiji/Guyu)** This is a language model with both the pre-training and fine-tuning procedures [26]. BigLM-24 is the typical GPT-2 model with 345 million parameters (1024 dimensions, 24 layers, 16 heads). During training, we employ maximum likelihood estimation (MLE) to conduct the parameter learning. In the inference stage, various decoding strategies such as greedy search, beam search, truncated top-k sampling, and nucleus sampling are employed to conduct the response-text generation.

4.4. Setup

We are interested in a model that performs robustly across a diverse set of tasks. To this end, we used the same hyperparameters as those in the original paper. We ran these typical models on 4 Tesla K80 GPUs and saved the best model on the validation set for testing.

4.5. Results and Analysis

Tables 2–4 and Figures 2 and 3 describe the detailed evaluation results for all the models on the KdConv, Weibo, and Douban datasets, separately.

Which models generate results with better relevance?

• As shown in Table 2, on the KdConv dataset, according to the BLEU and perplexity, the performance prioritizes for pre-training model-based models, transformer-based models, RNNSearch-based models, RL-based models, Seq2Seq-based models, and GAN-based models. As given in Table 3, but on the Weibo dataset, from the results, we can observe that the performances of RL-based models are better than those of RNNSearch-based models. The performances of GAN-based models were better than those of the Seq2Seq-based models. As given in Table 4, on the Douban dataset, according to the performances of those six architectures, the performances are ranked as pre-training model-based models, transformer-based models, RNNSearch-based models, RL-based models, Seq2Seq-based models, and GAN-based models. Among them, Seq2Seq-based models and GAN-based models performed similarly.

How about diversity?

• As given in Figure 2, pre-training can benefit performance on diversity (distinct). Figure 3 also shows pre-training model-based models obtain better rouge scores than other models on all the three datasets. Transformer-based models also perform well. This is attributed to that the pre-trained model-based models are flexible enough to handle various down-stream tasks of dialogue generation.

Are larger models helpful?

• As shown in Tables 2–4, the pre-training model-based model obtained a lower perplexity. Compared with the Seq2Seq framework, attention mechanisms can improve the performance. Transformer-based models also obtained better performances on most of the datasets. GAN-based models had worse performances in dialogue generation. RL-based models are necessary for further research in dialogue generation. Obviously, pre-training model-based models obtained better performances on automatic evaluation metrics than other models on three datasets. This phenomenon waws more evident on the multi-turn datasets of Douban and KdConv.

How about high-quality training data?

• Models with high-quality conversation datasets can also improve the performance. From the results, we see that models on KdConv are usually better than Douban and Weibo on the metrics of BLEU, perplexity, distinct, and rouge. The reason may be that KdConv can provide more useful information than the other datasets.

Table 2. Experimental results of BLEU and perplexity on KdConv dataset. The best and second-best results in each metric are bold and underlined, respectively.

Model	Relevance				Perplexity
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	
HRED [18]	36.87	26.68	21.31	17.96	41.07
ReCoSa [22]	27.69	14.13	7.35	14.31	36.71
Guyu [26]	37.15	28.27	23.13	18.89	35.56
RL [4]	30.29	21.79	16.15	13.02	37.96
GAN-AEL [29]	25.08	13.73	8.71	5.75	46.09
BigLM-24 [26]	**61.82**	**57.71**	**54.96**	**52.79**	**26.68**

Table 3. Experimental results of BLEU and perplexity on Weibo dataset. The best and second-best results in each metric are bold and underlined, respectively.

Model	Relevance				Perplexity
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	
HRED [18]	15.17	2.41	0.27	0.05	90.47
ReCoSa [22]	32.14	10.34	4.27	2.19	89.24
Guyu [26]	38.11	12.34	5.38	4.57	53.10
RL [4]	35.04	11.45	5.39	3.29	58.67
GAN-AEL [29]	28.34	8.45	4.38	3.03	91.31
BigLM-24 [26]	**38.93**	**14.72**	**7.59**	**5.07**	**40.23**

Table 4. Experimental results of BLEU and perplexity on Douban dataset. The best and second-best results in each metric are bold and underlined, respectively.

Model	Relevance				Perplexity
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	
HRED [18]	13.13	2.67	0.36	0.07	93.94
ReCoSa [22]	14.58	1.38	0.28	0.09	89.84
Guyu [26]	**17.19**	**5.66**	1.06	0.16	77.24
RL [4]	14.43	1.48	0.19	0.01	79.28
GAN-AEL [29]	13.30	0.90	0.05	0.01	85.44
BigLM-24 [26]	14.39	4.26	**1.26**	**0.59**	**41.46**

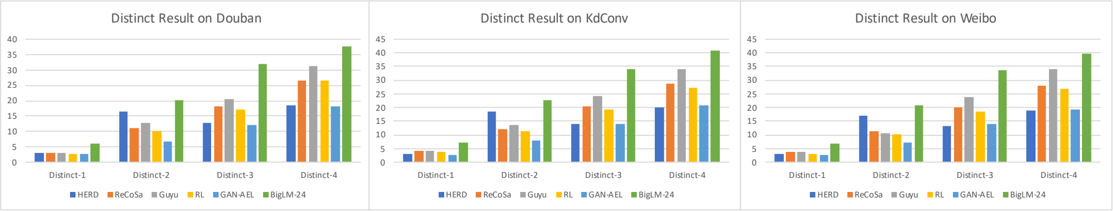

Figure 2. Experimental results with automatic evaluation distinct-1/2/3/4.

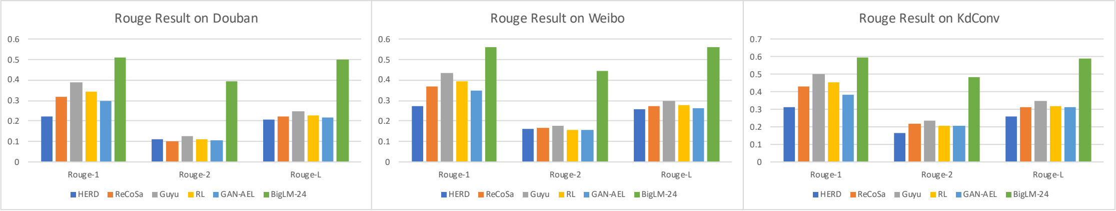

Figure 3. Experimental results with automatic evaluation rouge-1/2/L.

4.6. Case Study

We conducted a case study to demonstrate how those models respond given a query. As shown in Table 5, the red keywords indicate the relevant context to the response and the blue contexts represent the generated diversity and consistency of the responses. Seq2Seq-based models tend to generate responses that are slightly relevant to the context. After introducing the attention mechanism, RNNSearch-based models and transformer-based models can generate context-aware responses such as "天气," "生活," and "新疆." Nevertheless, generating responses with RL-based models and generative adversarial nets can generate keywords related to queries, but the semantics are often inconsistent with the query and are still difficult for the stability of the generated statement quality. In contrast, responses generated by the pre-training language model-based models show better quality, achieving both high relevance and diversity. This demonstrates the ability of large-scale language models in the decoding phase.

Table 5. Examples of response generation. The red keywords indicate the relevant context to the response and the blue contexts represent the generated diversity and consistency of the responses.

	Corpus		
Architecture	**Weibo**	**KdConv**	**Douban**
Query	今天天气不错 It's a nice day today	新疆是个美丽的地方 Xinjiang is a beautiful place	什么是有意义的生活 What is a meaningful life？
Ground Truth	确实是个好天气。 It's really nice weather.	新疆确实是个美丽的地方。 Xinjiang is really a beautiful place.	生命的意义在于奋斗。 The meaning of life lies in fighting.
Seq2Seq [18]	你也要注意身体呀. You should take care of your health, too.	你要去看看吗? Would you like to see it?	存在的意义在于生活 The meaning of existence lies in life.
RNNSearch [22]	我想看到你 I want to see you.	你知道有什么景点么? Do you know any scenic spots?	生活就是生活 Life is life.
Transformer [26]	北京天气不错的说 The weather in Beijing is good.	周边景点有哪些? What are the surrounding scenic spots?	生活毫无意义,重要的是什么让你快乐 Life is meaningless. What matters is what makes you happy.
RL [4]	好天气 The good weather.	新疆的美食还是有的 Xinjiang cuisine is still available.	这是为了你的生活 It's for your life.
GAN [29]	天气不错吗? Is the weather nice?	地址在哪呀 Where is the address?	我还是快乐的过着自己的生活 I'm still happy with my life.
PTM [26]	早安!早上好!美丽的北京! Good morning! Beautiful Beijing!	需要门票吗?能玩多久? Do I need tickets? How long will it last?	个人的生活其实就是妥协,找到平衡点 Personal life is all about compromise and finding balance.

We also conducted bot-to-bot interaction experiments on KdConv-bot via the pre-trained language model-based model, and the sample results are shown in Table 6. We set up the four-round so that the two robots could interact for four rounds. We extract the demonstration. It should be noted that there were no two robot scene settings in our work; we just managed the context memory to generate the next sound. Therefore, we can observe the topic will drift over time.

Table 6. Example of interactions between two multi-turn KdConv-bots.

Seed Query	你的梦想是什么? What is your dream?
	Robot A:你的梦想是什么? What is your dream?
Results	Robot B:你猜呀? Can you guess? Robot A:成为最好的人工智能机器人。 Be an excellent artificial intelligence robot. Robot B:跟我多聊聊你会知道的。 Well, talk to me a lot and you'll know. Robot A:调侃你么? Making fun of you ? Robot B:我知道你知道的。 I know you know. Robot A:你知道北京石刻艺术博物馆吗? Do you know Beijing Stone Carving Art Museum ? Robot B:我不知道。 I don't know ?

4.7. Discussion

We present an empirical study on the performances and impacts of six different dialogue generation DNN architectures. Encoder–decoder-based models are strong baseline methods and always produce better outcomes in automatic evaluation metrics. We argue that introducing the a pre-trained language model into the encoder–decoder framework may further enhance the performance significantly. If the problem that the gradient from the generative model is difficult to pass from the discriminative model to the generative model cannot be effectively solved, GAN-based models will suffer from the dialogue generation performance. The RL-based models can create the long-term influence of a generated response in an ongoing dialogue; however, the design of the reward function depends on experience, and the training process is difficult; the stability of generating sentences is not good enough.

We found that generated text in Weibo and Douban datasets was more random, with topics often drifting. The reason comes down to the distribution of the data in the dataset; they are chat datasets, so for the same question, the datasets contained multiple topics. In the KdConv dataset, the dialogue performance in the fields of music, travel, and film are better, because the KdConv dataset mainly contains the data distribution of the three fields. This indicates that the quality of dialogue generation is strongly related to content and data distribution. Notwithstanding, some grave issues also live in the results generated—the grammatical issue and the topic drift problem, to name a few. More seriously, sometimes the chatbots will generate contra-factual or offensive content. Hence, each architecture is required better model structures, training paradigms, and decoding strategies that need to be investigated and built in the future.

5. Conclusions and Future Work

In this paper, we show an empirical exploration regarding the performances of different architectures in the task of open-domain Chinese dialogue generation. A wide range of experiments were carried out on typical single-turn and multi-turn conversational datasets. We reported the detailed values of automatic evaluation metrics BLEU, perplexity, distinct, and rouge for the generated dialogue. We found that the text generated on the large-scale pre-training model is superior to other models in terms of evaluation metrics. An attention mechanism can significantly improve the performance of dialogue generation tasks. We also reported a case study to show example responses generated by these models separately and analyzed the reasons behind the different performances by these models and provided practical recommendations for model selection.

In future work, we will compare more architectures, such as VAE [45,46], and we will select 3–5 models for evaluation of each architecture. Meanwhile, more automatic metrics—embedding matching, METEOR, etc., as well as the human evaluation performance, will be completed in the future.

Author Contributions: Writing–original draft preparation, Z.L.; writing–review and editing, M.M.; methodology, Z.K. and J.S.; supervision, W.S.; software, Q.W.; resources, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by supported by Xinjiang University Ph.D. Foundation Initiated Project grant number 620312343, Xinjiang Uygur Autonomous Region Graduate Research and Innovation Project grant number XJ2020G071, Dark Web Intelligence Analysis and User Identification Technology grant number 2017YFC0820702-3, National Language Commission Research Project grant number ZDI135-96, and funded by National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC).

Acknowledgments: We thank the anonymous reviewers for their valuable feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 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, Washington, DC, USA, 18–21 October 2013; pp. 935–945.
2. Hu, B.; Lu, Z.; Li, H.; Chen, Q. Convolutional neural network architectures for matching natural language sentences. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2042–2050.
3. Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Hierarchical neural network generative models for movie dialogues. *arXiv* **2015**, arXiv:1507.04808.
4. Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; Gao, J. Deep Reinforcement Learning for Dialogue Generation. *arXiv* **2016**, arXiv:1606.01541.
5. Olabiyi, O.; Mueller, E.T. Multi-turn dialogue response generation with autoregressive transformer models. *arXiv* **2019**, arXiv:1908.01841.
6. Li, X.; Li, P.; Bi, W.; Liu, X.; Lam, W. Relevance-Promoting Language Model for Short-Text Conversation. *arXiv* **2019**, arXiv:1911.11489.
7. Li, P. An Empirical Investigation of Pre-Trained Transformer Language Models for Open-Domain Dialogue Generation. *arXiv* **2020**, arXiv:2003.04195.
8. Wang, Q.; Yin, H.; Wang, H.; Hung, N.Q.V.; Huang, Z.; Cui, L. Enhancing Collaborative Filtering with Generative Augmentation. In *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 548–556.
9. Wang, Q.; Yin, H.; Chen, T.; Huang, Z.; Wang, H.; Zhao, Y.; Viet Hung, N.Q. Next Point-of-Interest Recommendation on Resource-Constrained Mobile Devices. In Proceedings of the Web Conference, Taipei, Taiwan, 20–24 April 2020; pp. 906–916.

10. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
11. Wang, Q.; Yin, H.; Hu, Z.; Lian, D.; Wang, H.; Huang, Z. Neural Memory Streaming Recommender Networks with Adversarial Training. In *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 2467–2475.
12. Dai, Y.; Li, H.; Tang, C.; Li, Y.; Sun, J.; Zhu, X. Learning Low-Resource End-To-End Goal-Oriented Dialog for Fast and Reliable System Deployment. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 9–10 July 2020; pp. 609–618.
13. Cui, L.; Wu, Y.; Liu, S.; Zhang, Y.; Zhou, M. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1406–1416. [CrossRef]
14. Qin, L.; Xu, X.; Che, W.; Zhang, Y.; Liu, T. Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6344–6354. [CrossRef]
15. Cho, H.; May, J. Grounding Conversations with Improvised Dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 2398–2413. [CrossRef]
16. Ma, W.; Cui, Y.; Liu, T.; Wang, D.; Wang, S.; Hu, G. Conversational Word Embedding for Retrieval-Based Dialog System. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1375–1380. [CrossRef]
17. Sutskever, I.; Vinyals, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
18. Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, AZ, USA, 12–17 February 2016.
19. Ghazvininejad, M.; Brockett, C.; Chang, M.W.; Dolan, B.; Gao, J.; Yih, W.t.; Galley, M. A knowledge-grounded neural conversation model. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
20. Vinyals, O.; Le, Q. A neural conversational model. *arXiv* **2015**, arXiv:1506.05869.
21. Shang, L.; Lu, Z.; Li, H. Neural Responding Machine for Short-Text Conversation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, China, 16–21 August 2015; pp. 1577–1586.
22. Zhang, H.; Lan, Y.; Cheng, X. ReCoSa: Detecting the Relevant Contexts with Self-Attention for Multi-Turn Dialogue Generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3721–3730.
23. Tian, Z.; Yan, R.; Mou, L.; Song, Y.; Feng, Y.; Zhao, D. How to make context more useful? An empirical study on context-aware neural conversational models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 231–236.
24. Xing, C.; Wu, W.; Wu, Y.; Zhou, M.; Huang, Y.; Ma, W.Y. Hierarchical recurrent attention network for response generation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
26. Li, Z.; Niu, C.; Meng, F.; Feng, Y.; Li, Q.; Zhou, J. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 12–21.

27. Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; Weston, J. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April– 3 May 2018.
28. Yu, L.; Zhang, W.; Wang, J.; Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, USA, 4–9 February 2017.
29. Xu, Z.; Liu, B.; Wang, B.; Sun, C.J.; Wang, X.; Wang, Z.; Qi, C. Neural response generation via gan with an approximate embedding layer. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 617–626.
30. Boyd, A.; Puri, R.; Shoeybi, M.; Patwary, M.; Catanzaro, B. Large Scale Multi-Actor Generative Dialog Modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 66–84. [CrossRef]
31. Zhang, Z.; Li, X.; Gao, J.; Chen, E. Budgeted Policy Learning for Task-Oriented Dialogue Systems. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3742–3751.
32. Su, S.Y.; Li, X.; Gao, J.; Liu, J.; Chen, Y.N. Discriminative Deep Dyna-Q: Robust Planning for Dialogue Policy Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3813–3823.
33. Wu, Y.; Li, X.; Liu, J.; Gao, J.; Yang, Y. Switch-based active deep dyna-q: Efficient adaptive planning for task-completion dialogue policy learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7289–7296.
34. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
35. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
36. Ham, D.; Lee, J.G.; Jang, Y.; Kim, K.E. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 583–592. [CrossRef]
37. Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 85–96. [CrossRef]
38. Shang, L.; Lu, Z.; Li, H. 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)*; Association for Computational Linguistics: Beijing, China, 2015; pp. 1577–1586. [CrossRef]
39. Gao, J.; Bi, W.; Liu, X.; Li, J.; Shi, S. Generating multiple diverse responses for short-text conversation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6383–6390.
40. 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. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 496–505.
41. Pan, Z.; Bai, K.; Wang, Y.; Zhou, L.; Liu, X. Improving Open-Domain Dialogue Systems via Multi-Turn Incomplete Utterance Restoration. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 1824–1833.
42. 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 of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
43. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A Diversity-Promoting Objective Function for Neural Conversation Models. In Proceedings of the NAACL-HLT 2016, San Diego, CA, USA, 12–17 June 2016; pp. 110–119.

44. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
45. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2014**, arXiv:1312.6114.
46. Zhao, T.; Zhao, R.; Eskenazi, M. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 654–664.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.