electronics

MDPI

Article

Keyword-Aware Transformers Network for Chinese Open-Domain Conversation Generation

Yang Zhou [1], Chenjiao Zhi [1], Feng Xu [2], Weiwei Cui [3], Huaqiong Wang [4], Aihong Qin [4], Xiaodiao Chen [2,4], Yaqi Wang [4,*] and Xingru Huang [3,*]

[1] Alibaba Group, Hangzhou 311121, China
[2] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310005, China
[3] School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK
[4] College of Media Engineering, Communication University of Zhejiang, Hangzhou 310042, China
* Correspondence: wangyaqi@cuz.edu.cn (Y.W.); xingru.huang@qmul.ac.uk (X.H.)

Abstract: The open-domain conversation generation task aims to generate contextually relevant and informative responses based on a given conversation history. A critical challenge in open-domain dialogs is the tendency of models to generate safe responses. Existing work has often incorporated keyword information in the conversation history for response generation to relieve this problem. However, these approaches interact weakly between responses and keywords or ignore the association between keyword extraction and conversation generation. In this paper, we propose a method based on a Keyword-Aware Transformers Network (KAT) that can fuse contextual keywords. Specifically, the model enables keywords and contexts to fully interact with responses for keyword semantic enhancement. We jointly model the keyword extraction task and the dialog generation task in a multi-task learning fashion. Experimental results of two Chinese open-domain dialogue datasets showed that our proposed model outperformed the methods in both semantic and non-semantic evaluation metrics, improving Coherence, Fluency, and Informativeness in manual evaluation.

Keywords: dialog generation; keyword extraction; open-domain dialog; natural language processing

check for updates

Citation: Zhou, Y.; Zhi, C.; Xu, F.; Cui, W.; Wang, H.; Qin, A.; Chen, X.; Wang, Y.; Huang, X. Keyword-Aware Transformers Network for Chinese Open-Domain Conversation Generation. *Electronics* **2023**, *12*, 1228. https://doi.org/10.3390/electronics12051228

Academic Editors: Shangsong Liang and Zaiqiao Meng

Received: 7 February 2023
Revised: 23 February 2023
Accepted: 28 February 2023
Published: 4 March 2023

1. Introduction

Open-domain dialog generation aims to generate a fluent, coherent, and informative response given the dialog history [1–5], which is a long-term goal of artificial intelligence [3]. There is a boom in developing intelligent open-domain dialog systems due to the availability of a large conversational corpus [6] and the recent success of pre-training models in natural language processing [7]. For instance, Generative Pre-trained Transformer (GPT) [8] demonstrates the state-of-the-art performance in natural language generation [9,10]. GPT-based methods are proposed in open-domain dialog systems to achieve decent performances on the metrics of quality, relevance, engagement, and diversity [11,12]. An open-domain dialog agent is critical for various applications spanning from entertainment and knowledge sharing to customer services [13].

Despite the success of multi-turn dialog systems that leverage context information to make a response, they still suffer from suboptimal performance due to the following reasons [14–16]. First, the previous approach was to fully utilize all words in the context, leading to the introduction of redundant information that was not relevant to the response. Second, the system's response is not natural. In actual replies, people tend to emphasize the key content/keywords mentioned by both parties, allowing them to be more coherent. Finally, some pre-trained language models focus on contextual representation and ignore the process of keyword extraction [17,18], resulting in a large number of meaningless responses [13]. In addition to these challenges, in actual conversational situations, people usually connect the keywords in the conversation with some concepts in their brains and choose the appropriate words to generate the corresponding responses based on the concepts. For example, given the context of another conversation "I have a headache and a cough", the respondent often associates the headache and cough with a health problem, and then associates the concept of health with taking medication and rest. Finally, the appropriate response from these associations is "I think you should go to the doctor and get some rest" rather than just saying "I'm sorry to hear that".

To solve the above problem and simulate the flow of human conversations naturally, this paper proposes a novel Keyword-Aware Transformers Network (KAT) approach that can fuse keyword information in the context and jointly train keyword extraction and conversation generation tasks. In addition, the KAT model can be very easily ported to pre-trained language models. Our model has three critical modules: Contextual Encoder, keyword information enhancement module, and dialog generation module. Specifically, we use a Contextual Encoder to obtain an encoded representation of the context and then capture the core information by performing keyword predicting on the sentences in the context with the keyword predicting module. Then, the keyword information enhancement module enhances the extracted keyword information and enhances its semantic information. The final dialog generation module generates relevant and fluent responses based on the global contextual information and the extracted keyword information. We validated the model on two Chinese conversation corpora: a large short-text conversation dataset, STC, collected from Weibo and the Chinese DailyDialog dataset (DD). Our experiments demonstrated that our model not only achieved better results in automatic metrics compared to the basic model but also made some progress in manual evaluation and converged significantly faster. In addition, we verified the robustness of our model by adding noise to the test set. The contributions of the paper are as follows:

1. We propose a novel architecture KAT to ensure a deep fusion of keyword information as an aid to the dialog generation task.
2. We propose a joint modeling approach for both keyword predicting and dialog generation tasks to improve the dialog-generating performances.
3. We experimentally demonstrate that our KAT approach not only has significantly higher automatic and manual evaluation metrics than the baseline model but also has improved robustness and convergence speed.

2. Related Work

2.1. Chinese Conversation Dataset

In recent years, open-domain dialog systems have attracted a lot of interest with the released Chinese conversational corpus. Some open-access large-scale datasets are collected from social media such as Weibo, Twitter and Reddit [19]. However, several datasets contain many negative behaviors, including toxic comments, threats, insults, identity hates, obscene content, etc., making it hard to ensure the quality of the data and this markedly degrades the generation ability of dialog generation models. In this study, we utilize a high-quality Chinese version of the multi-round dialog dataset, or DailyDialog Dataset with 13,118 dialogs [20], containing rich daily scenarios and many emotional changes. In addition, to further validate the generalization performance of the model, we exploit the large dialog dataset STC [21] to verify the validity of the model. All the data in STC is collected from Weibo and has relatively few types of scenes.

2.2. Chinese Pre-Training Model

Pre-trained language models have made a great contribution to open-domain conversational generation tasks and rely on large-scale conversation datasets [22]. The Chinese open-domain dialog pre-trained language model includes DialogGPT [19], PLATO [11], EVA [23] and EVA2.0 [24]. These models cast open-domain dialog modeling as a seq2seq learning problem [25]. However, to truly understand human conversations, a dialog agent should be able to explore the semantic concepts or keywords from the context to make an explainable response. We focused on injecting the keywords extraction into pre-trained language models for open-domain dialog systems. To our knowledge, DialoGPT achieves advanced performance in Chinese open-domain dialog datasets. Thus, in this paper, we considered using the DialoGPT network pre-trained on the base version of the large-scale cleaned Chinese conversation dataset (LCCC-base) as our backbone [19].

2.3. Keyword-Aware Dialog System

Keyword information is a very critical feature in dialog generation tasks. In the course of human dialog, the responses to the dialog are mainly expanded based on keywords in the context [13]. Keyword information can not only bring out the core points in the context but also provide prior knowledge and direction to the responder, facilitating deepening of the topic of discussion and making the conversation more attractive. In previous work, information extraction and information enhancement of keywords in context were processed separately by two independent networks [13]. For example, LDA [26] is used to obtain the subject words, and the information of relevant subject words is enhanced by the attention mechanism to achieve the purpose of information enhancement [27]. Then, probability distributions of subject terms in the dialog generation process were introduced to generate relevant responses [28]. Previous work studied the affective and behavioral information to predict action, emotion, and topic words. There are, however, several shortcomings to these approaches, including the lack of a deep understanding of keywords and the reliance on unsupervised models to extract keywords. In this paper, we propose a generative model to incorporate keyword information at a deep level and to jointly train keyword predicting (uniformly called prediction) and dialog generation tasks.

3. Approach

The architecture of the KAT network is shown in Figure 1. It contains three main modules: Contextual Encoder, keyword information enhancement module, and dialog generation module. Specifically, the Contextual Encoder aims to obtain a representation of the context. The keyword information enhancement module contains two operations, extraction and enhancement, to enhance the model's keyword extraction capability so as to improve the quality of the response. Finally, the dialog generation module performs keyword prediction and generates dialog responses. Detailed information for each module and corresponding output equations are provided in Figure 2. Next, we describe our task formulation and architecture in detail.

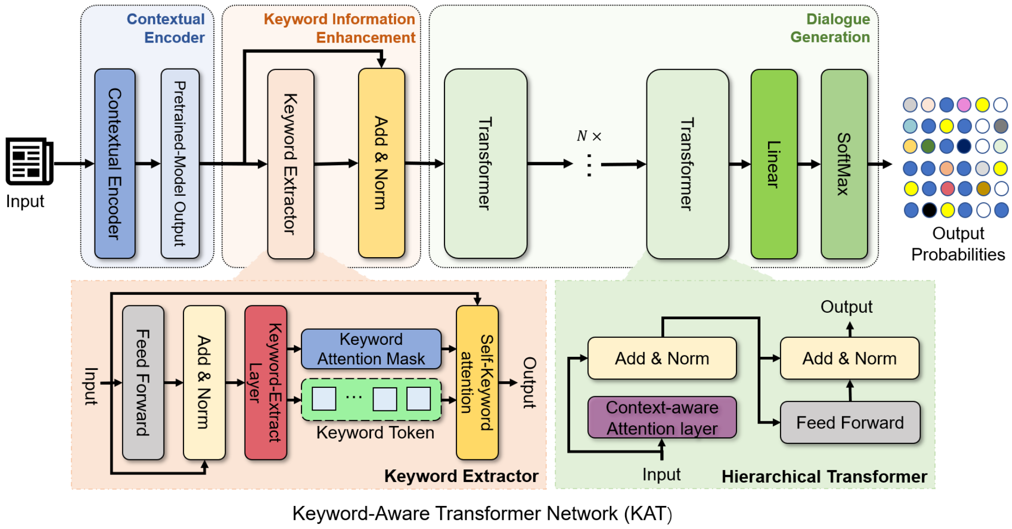

Figure 1. The overview of the proposed Keyword-Aware Transformers Network including Contextual Encoder, keyword information enhancement module, and dialog generation module. The Keyword Extractor is on a salmon background, while the Hierarchical Transformer layer is on a light green background.

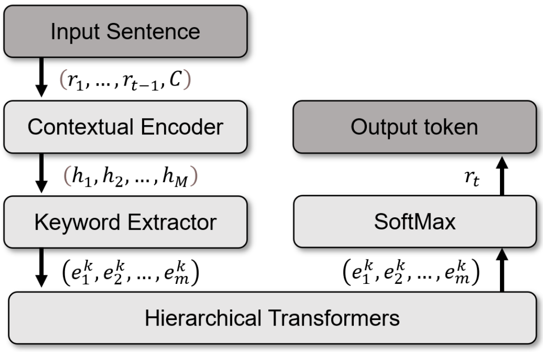
Input Sentence
$(r_1, \ldots, r_{t-1}, C)$
Contextual Encoder
$(h_1, h_2, \ldots, h_M)$
Keyword Extractor
$(e_1^k, e_2^k, \ldots, e_m^k)$
Hierarchical Transformers
$(e_1^k, e_2^k, \ldots, e_m^k)$
SoftMax
r_t
Output token

Figure 2. Overview of the proposed approach and output for keyword-aware dialog generation. The outputs of each module are provided in the boxes below their corresponding equations.

3.1. Task Formulation

The problem formulation of our task was as follows. For general dialog response generation, given a dialog history $C = (u^1, u^2 \cdots, u^{N-1})$ consisting of $N-1$ historical utterances, where n^{th} utterance is defined as $u^n = (x_1^n, x_2^n \cdots, x_M^n)$ that consists of M words, and the target response $r = (r_1, r_2 \cdots, r_T)$ with T words, our goal was to estimate the probability distribution $P(r|C) = \prod_{t=1}^{T} P(r_t|r_1, \cdots, r_{t-1}, C)$ and to generate the token r_t with the given dialog history C and the generated tokens $(r_1, \cdots, r_{t-1})$ until the whole response was generated with one termination symbol.

The purpose of the Contextual Encoder is to encode the token of the context and $t-1$ part of the response into a vector. For a given context and the $t-1$ part of the response, we first need to encode this information. We can use pre-trained models or structures, such as RNNs, to do so.

To obtain representations of the dialog history, we followed previous work [1] and separated each utterance with special tokens $[speaker1]$ and $[speaker2]$. We prepared a special token $[CLS]$ at the beginning of the dialog history to reconstruct the context input, and added the special token $[SEP]$ at the end. The input is represented as follows:

$$Input = ([CLS], [speaker1], u^1, [speaker2], u^2, ..., [speaker1], u^{N-1}, [speaker2], r_1 ..., r_{t-1}, [SEP]) \quad (1)$$

In addition to the dialog history, when generating the tth token in the response, the input also includes the tokens generated from time step 0 to $t-1$.

3.2. Contextual Encoder

To obtain contextual representations, we utilized a contextual encoder, denoted as $Enc_{context}$, to encode each token in the input, which leads to a series of context-aware hidden states $(h_1, h_2, ..., h_M)$, represented as:

$$h_m = Enc_{context}([CLS], [speaker1], u_1^1, u_2^1, ..., [speaker1], u_1^{N-1}, u_2^{N-1}, ..., [speaker2], r_1 ..., r_{t-1}, [SEP]) \quad (2)$$

Here, u^i represents the ith utterance in the dialog history, u_k^i represents the kth token in the i^{th} sentence in the dialog history, where k denotes the length of the sentence, $r_1, ..r_{t-1}$ represents the tokens generated by the decoder from time step 1 to $t-1$, $h_m \in R^{d_{model}}$ represents the contextual representation of the m^{th} token in the input, and d_{model} is the hidden dimension of the contextual encoder.

3.3. Keyword Information Enhancement Module

The keyword information enhancement module is proposed. It extracts the dominant keyword features of each sentence to generate keyword-aware responses. We believe that keywords do not need to be strictly limited, as long as the relevant words contain certain content or patterns. For example, in the sentence "It's a nice day", "nice" can also be used as a keyword, because we can generate logically correct replies based on it, such as "Yes, let's go play basketball together!".

3.3.1. Self-Keyword Attention

To allow for effective interaction between the generated tokens and the keyword tokens in the context, and to enable the model to focus on relevant contextual information, we propose a novel self-keyword attention mechanism that enhances the original self-attention mechanism by operating on the keys in an element-wise manner. We apply the attention function to a batch of queries simultaneously by organizing them in a matrix called Q. Similarly, we group the keys and values into separate matrices K and V, respectively, with Q, K, and V all sharing the same dimensions as the output vectors of the Contextual Encoder, denoted as H.

To obtain a deeper representation of the keys, we perform an element-wise interaction between the key vectors K and the query vector Q to obtain an updated key vector representation, K''. This is computed as:

$$K' = [K; Q + K; Q - K; Q \odot K] \tag{3}$$

$$K'' = W^{KW}(Tanh(K')) \tag{4}$$

where [;] denotes concatenation and $-$, $+$, $\odot$ represents element-wise addition, subtraction, and multiplication, respectively. Tanh is the hyperbolic tangent function.

After computing the updated key vector representation K'', we obtain the matrix of self-keyword attention outputs as:

$$SelfKeywordAttention(Q, K'', V) = softmax(\frac{Q(K'')^T}{\sqrt{d_k}})V \tag{5}$$

We used multiple heads in the self-keyword attention mechanism, which allows the model to capture information from different positions and feature subspaces. This is computed as:

$$KeywordMultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^z \tag{6}$$

where $W^{KW} \in R^{4d_h \times d_h}$ and $W^z \in R^{hd_h \times d_h}$. In this work, we employed 8 parallel attention layers, or heads, and for each of these, we used $d_h = d_{model}/h = 96$, where d_{model} represents the dimensionality of the output vectors of the Contextual Encoder.

3.3.2. Keyword Attention Mask

Keyword Attention Masks (KA-masks) force the keyword extractor to focus on keyword tokens, enhancing keyword extraction efficiency. In the GPT model, based on the autoregressive modeling approach, due to the nature of the one-way language model, the token at each time step can only interact with the token on the left side when modeling the context. However, this modeling approach limits the model to extracting contextual information. In a conversation, all the contextual tokens are known, and a two-way language modeling approach can fully exploit the contextual semantics and provide more valid information for the generated conversation. So, we improved the masked approach so that the words above and below could fully interact with each other, as shown in the upper left part of Figure 3. In addition, we wanted the model to allow the tokens in the responses to interact only with the keywords in the context. In this way, the generated responses could focus on the keywords in context, providing more guidance for the conversation generation task (as shown in the bottom left part of Figure 1; the part with mask one is the keyword token). In the response generation phase, the lower triangular matrix can only interact with the generated response token.

3.4. Context-Aware Attention Layer

In the Hierarchical Transformer layer, we proposed a new context-aware Attention layer to fully utilize context information from each neighborhood layer. In the previous section, we used the self-keyword attention mechanism to allow the token in the response to interact with the keyword token, which, to some extent, guides response generation.

However, this approach does not fully utilize context information. Therefore, we further modified the mask matrix in the second self-attention operation of the module, as shown in the bottom half of Figure 3, where the reply part can interact with all the tokens in the context, but the token in the reply can only interact with the token in the t − 1 part on its left side. The transformer layer can be stacked with multiple layers. Different layers can learn different levels of semantic information.

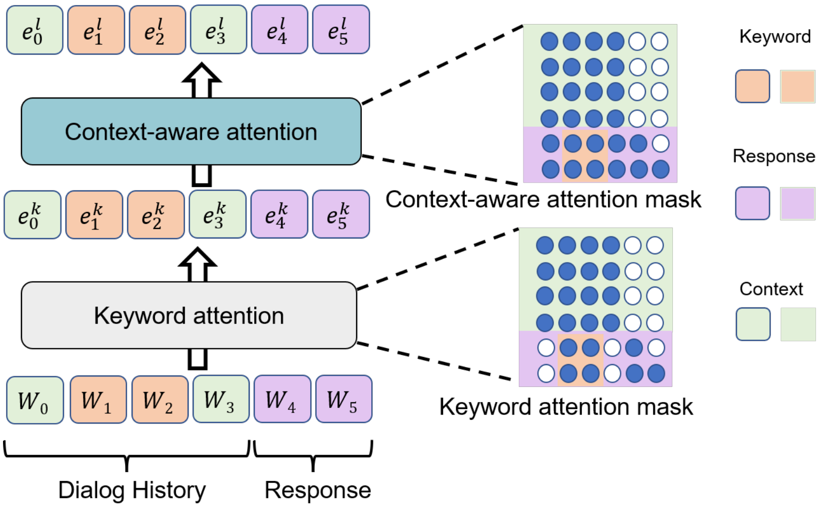
e_0^l e_1^l e_2^l e_3^l e_4^l e_5^l
Context-aware attention
e_0^k e_1^k e_2^k e_3^k e_4^k e_5^k
Keyword attention
W_0 W_1 W_2 W_3 W_4 W_5
Dialog History
Response
Context-aware attention mask
Keyword attention mask
Keyword
Response
Context

Figure 3. Two types of the proposed attention mechanism, including context-aware attention and keyword attention. The upper right dashed extension shows the specific details of the context-aware attention mask, while the lower right shows the keyword attention mask in detail.

3.5. Joint Dialog Generation and Keyword Prediction

Dialog responses are usually generated by a single decoder. In this research, we proposed a keyword-guided dialog generator, while the predicted keyword can constrain the output dialog. Keyword information is crucial for conversation generation and contains important semantic information in the context. When generating responses, this can make the generated sentences more contextually consistent, semantically relevant, and logical by focusing on keyword information. In addition, with the inclusion of keyword information, the model can learn a keyword-related generation pattern, have faster convergence speed and stronger resistance to interference and robustness. When the non-keyword part changes, it can also generate reasonable dialog responses by combining parts of the keywords. Therefore, we believe that the keyword predicting task and the dialog generation task can help each other, and the ablation experiments also demonstrated that both tasks were effective for each other.

Specifically, our keyword predicting task was a binary classification task that predicted whether each token was part of the keyword. The optimization goal was:

$$L_1 = -\frac{1}{M}\sum_{i=1}^{M}\tilde{y}_i log(p(y_i)) \tag{7}$$

where $\tilde{y}_i$ is the keyword token true label, $p(y_i)$ is the probability value of the predicted key token. For the conversation generation task, we fused the contextual information and the predicted token from the previous time step to predict the correct token for the current time step. The standard negative log-likelihood loss is as follows:

$$L_2 = -\frac{1}{T}\sum_{t=1}^{T} log(p(r_t|r_1, \cdots, r_{t-1}, C)) \tag{8}$$

where, r_t is the predicted token for the current time step, $r_1, \cdots, r_{t-1}$ are predicted tokens, C is the contextual part of the conversation, T is the length of the responses. The final optimization objective is:

$$L = \lambda_1 L_1 + \lambda_2 L_2 \tag{9}$$

4. Experiments

4.1. Dataset

r^j To verify the validity of the model, we evaluated it on the Chinese DailyDialog dataset (DD) and the STC dataset. For each sentence in the two datasets, we performed normalization, i.e., removed emojis and other expressions. For consecutive occurrences of symbols, we kept only one and performed lowercase conversion on all corpora. After that, all data was randomly sliced.

4.1.1. Chinese Dailydialog

Chinese DailyDialog dataset (DD) contains 11,943 multi-round dialogs with an average of 8 tokens per dialog and 15 tokens per sentence [29]. All sentences in the DD are translated from the English version of the daily dialog, which contains a wide variety of topics and rich content. During pre-processing, we de-duplicated the Chinese version of the DailyDialog dataset to ensure that the training and test sets did not contain identical dialogs. Following this, we randomly selected the training set, validation set, and test set in the ratio 8:1:1.

4.1.2. STC Dataset

A total of 4.3M conversations are included in STC, derived from the Weibo dataset corpus. These conversations are similar to forum Q&As, slightly different from daily conversations, but larger. In this study, we randomly divided the dataset into a training set validation set and a test set, where the validation set contained 20,000 conversations and the test set contained 15,000 conversations. In the testing process, the best-performing model was used after the experimental results were validated on the validation set.

4.2. Evaluation Metrics

4.2.1. Automatic Evaluation

An automatic evaluation metric and a manual evaluation metric were used to evaluate the model's effectiveness. The metrics for automatic evaluation included blue value distinct and rouge metrics. BLEU [30] and rouge [31] metrics measure the quality of model response generation. BLEU(Bilingual evaluation understudy) is a metric used to evaluate the difference between the sentences generated by the model (candidate) and the reference sentences (reference), and is based on the Precision of n words. The higher its value, the closer the generated sentences are to the reference sentences. The formula is shown below:

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C' \in \{\text{ Candidates }\}} \sum_{n\text{-gram} \in C'} \text{Count}(n\text{-gram })} \tag{10}$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{11}$$

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{12}$$

where the n-gram takes the content inside the sentence as the basic unit, sets the sliding window size as n, and divides the sentence to get multiple word fragment sequences of length n; p_n is the geometric average of the modified n-gram precisions. c represents the length of the generated sentence, r represents the length of the reference. w_n is the weight of n-gram, and we take it as $\frac{1}{n}$. In this paper, we chose n as 1 to judge the degree of similarity between the generated sentences and the original text, and n as 2 and 4 to judge the fluency of the generated sentences. ROGUE-2 (Recall-Oriented Understudy for Gisting Evaluation) was used to calculate the recall on the basis of 2-gram. The formula is shown below:

$$ROUGE-2 = \frac{\sum_{S\in\{\text{ReferenceSummaries}\}}\sum_{\text{n-gram} \in S} \text{Count}_{\text{match}}\,(\text{2-gram})}{\sum_{S\in\{\text{ ReferenceSummaries}\}}\sum_{n\text{-gram} \in S} \text{Count}\,(\text{2-gram })} \tag{13}$$

where the denominator of the formula counts the number of 2-grams in the reference sentence. The numerator counts the number of 2-grams shared by the reference sentence and the generated sentence.

Distinct [32] metrics measure the variety of generated responses by calculating the proportion of each unigram and bigram. Distinct was used to determine the diversity of the generated sentences. The higher its value, the higher the diversity of the generated sentences.

$$\text{Distinct}(n) = \frac{\text{Count }(\text{unique } n\text{-gram})}{\text{Count}(\text{word})} \tag{14}$$

Count(unique n-gram) denotes the number of non-repeated n-grams in a reply, and Count(word) denotes the total number of n-gram words in a reply. In the text, we take n to be 1 and 2, respectively.

4.2.2. Manual Evaluation

Our manual evaluation involved three crowdsourcers judging the results we generated, and each model was evaluated by scoring 150 samples randomly selected. Final evaluation results were calculated from the average scores of the crowdsourcers. In the performance evaluation, we considered three metrics: fluency, relevance, and information richness.

4.3. Experimental Setup

In the training stage, we used the cross-entropy as the training loss for the dialog response generation task. In the prediction phase, we used Nucleus Sampling to enhance the diversity of generated responses [33]. In addition, we put a constraint on the minimum length generated, which was set to 4. The epochs of DD and STC datasets were both set to 10 and the learning rates of DD and STC were set to 7×10^{-5} and 6×10^{-5}. To coordinate the joint training of the two training objectives, $\lambda 1 = 1, \lambda 2 = 1$. For the keyword prediction task, we introduced the label smoothing technique. Module learning rates are decayed linearly using the Adam optimizer for all modules. Transformer decoders and transformer encoders are trained directly on the training set, not pre-trained. Transformer–decoder has a hidden layer size of 1024 and a layer count of 1. Transformer–encoder has a hidden layer size of 1024 and a layer count of 2. Its training consists of 30 rounds. GPT–LCCC-based training consists of 17 rounds, while GPT2–LCCC-based training consists of 27 rounds.

4.4. Baselines

We selected the following baseline models for comparison, as shown in Table 1, in which we list the name and features of baselines. In the table, we focus on the Chinese pre-trained language model because it is more powerful than light weighted models. We put more relevant light weighted models in Section 2.2.

1. Transformer–ED contains a transformer–encoder and a transformer–decoder, without pre-training [34].
2. Transformer–Dec contains only a transformer–decoder, which uses a left-to-right attention mask in both the context and response sections.
3. CDialGPT-LCCC-base contains the model architecture GPT1, trained on the Large Chinese Conversation base dataset (LCCC-base) [19].
4. CDialGPT2-LCCC-base contains model architecture GPT2, trained on LCCC-base [19].
5. CDialGPT-LCCC-large contains the model architecture GPT1, trained on the LCCC-large dataset.

Table 1. Comparison of the previous works.

Model	Architecture	If Fine-Tuned	Fine-Tuned Dataset
Transformer-ED	Transformer	No	-
Transformer-Dec	Transformer Decoder	No	-
CDialGPT-LCCC-base	GPT	Yes	LCCC-base dataset
CDialGPT2-LCCC-base	GPT2	Yes	LCCC-base dataset
CDialGPT-LCCC-large	GPT	Yes	LCCC-large dataset

4.5. Automatic Metrics on Chinese Daily-Dialogue Dataset

4.5.1. Automated Evaluation Metrics

The comparison of the main experiment is shown in Table 2. After introducing keyword information, the keyword-award-GPT1 (NO.5) and keyword-award-GPT2 (NO.7) achieved better results in BLEU, ROUGE, and D-2 compared with their baselines (NO.4 and NO.6). In particular, keyword-award-GPT1 improved BLEU-4 and ROUGE-2 metrics by 1.52 and 1.59, respectively, compared to the baseline. It proved the validity of the model. The model also reached convergence at epoch 10, indicating that it converged faster than the baseline method. Furthermore, the proposed method only gained a slight improvement on Distinct-1 and Distinct-2, since there were no restrictions on the generation of the baseline model, resulting in a larger search space. According to the human evaluation and case study, we found that the introduction of keyword information improved the model's keyword acquisition ability, limiting the space for token generation and reducing the possibility of raw words being generated. Furthermore, we also observed that, despite no significant increase in the Distinct automatic evaluation metric for our model, the model effectively reduced safe responses. Rich keyword information led to more contextually relevant, consistent, and richer word combinations.

Table 2. Comparison with state-of-the-art results on Chinese Daily-Dialog Dataset. For each evaluation metric in every column, the best result is highlighted in boldface.

NO.	Model	BLEU-1	BLEU-2	BLEU-4	ROUGE-2	D-1	D-2	Epoch
1	Transformer-Decoder	22.63	15.01	10.62	9.62	**27.27**	24.29	30
2	Transformer-ED	22.97	14.24	9.24	8.96	21.46	20.54	30
3	GPT-LCCC-Large	28.13	20.50	15.69	14.32	26.66	14.32	17
4	GPT-LCCC-Base	26.64	18.32	14.82	13.12	26.15	26.38	17
5	GPT-LCCC-Base+KAT	**28.78**	**21.17**	**16.34**	**14.71**	26.28	26.45	10
6	GPT2-LCCC-Base	25.40	17.15	11.92	11.04	26.70	25.11	27
7	GPT2-LCCC-Base+KAT	27.20	19.90	15.48	12.06	26.94	**26.99**	20

With more dialog rounds, longer sentences, and more complex scenarios in the DD dataset, the number of converged rounds was significantly smaller than in the baseline dataset. According to our analysis, the distribution of the DD dataset was not quite consistent with the distribution of the pre-trained data. DDs are typically multi-round conversations, which are rich in topics and have obvious non-Chinese characteristics. The pre-training data for the base model was mostly derived from Chinese forums and microblogs, which are significantly different from daily conversations. Consequently, CDialGPT and CDialGPT-2 require more time to fit the new data distribution. Our model enhanced contextual information by incorporating keyword information, which allowed it to converge faster and generate smoother and more informative conversational responses.

4.5.2. Manual Evaluation Metrics

Table 3 shows the results of the manual experimental comparisons. In comparison with their respective baselines, the proposed models showed significant improvements in the manual measures. Based on feedback from the annotators, our improved model improved fluency and contextual semantic relevance by introducing keywords and improving keyword predicting accuracy. In the evaluation of the base, we found that the base model had a higher probability of generating contextual inconsistencies, disfluencies, and logical inconsistencies in the responses, particularly in long conversations (multiple conversation rounds or long contexts).

Table 3. Human Evaluation Results on Chinese Daily-Dialog Dataset. For each evaluation metric in every column, the best result is highlighted in boldface.

Model	+2	+1	+0	Mean-Score
GPT-LCCC-Base	41.38%	30.46%	28.16%	1.1322
GPT-LCCC-Base+KAT	62.43%	23.7%	13.87%	**1.4855**
GPT2-LCCC-Base	39.33%	30.0%	30.67%	1.0867
GPT2-LCCC-Base+KAT	53.95%	27.63%	18.42%	**1.3553**

4.6. Ablation Experiments

Separate experiments were conducted to test the effect of the keyword information enhancement module and the keyword predicting module for joint modeling. The results can be seen in Table 4.

Table 4. Ablation Analysis on Chinese Daily-Dialog Dataset. For each evaluation metric in every column, the best result is highlighted in boldface.

Model	BLEU-1	BLEU-2	BLEU-4	ROUGE-2	D-1	D-2
GPT-LCCC-Base	26.64	18.32	14.82	13.12	26.15	26.38
GPT-LCCC-Base+Key-Word Enhance	28.04	20.44	15.64	14.46	26.17	**26.77**
GPT-LCCC-Base+KAT	**28.77**	**21.17**	**16.34**	**14.71**	**26.28**	26.45
GPT2-LCCC-Base	26.49	18.59	13.63	13.29	26.70	25.11
GPT2-LCCC-Base+Key-Word Enhance	26.52	19.04	14.43	11.61	26.79	25.34
GPT2-LCCC-Base+KAT	**27.20**	**19.90**	**15.48**	**12.06**	**26.94**	**26.99**

1. Pre-trained-model + keyword-enhance: We used the keywords extracted by jieba and hanlp as the input of the model directly, and made the model pay more attention to the keyword token by the keyword information enhancement module;
2. Pre-trained-model + keyword-enhance + keyword-extractor: We used the keywords extracted by jieba and hanlp as tags and added keyword-extractor to make the model have keyword extraction capability by itself. The model's ability to extract key information was enhanced by joint multi-task learning.

The BLEU and ROUGE values of the BASE models (CDialGPT and CDialGPT-2) were not as good as ours. The inclusion of keywords improved the information and proved its effectiveness. The experimental results improved further after we added a keyword extractor and performed joint multi-task training on this basis. The effectiveness of each component of the model was demonstrated through ablation experiments.

4.7. Evaluation on the STC Dataset

We also evaluated the proposed method on the STC dataset collected from Weibo, which is larger than the Chinese DD dataset. As the STC conversation corpus is relatively cluttered, and the original dataset contains much redundant punctuation, emojis, and other clutter symbols, we set it to two settings:

1. The training set was STC-clean after pre-processing and the test set was test-clean;

2. The training set was STC-clean after pre-processing and the test set was test-original.

A test set in different settings further demonstrated the model's performance in different scenarios. Setting 1: As shown in Table 5, our model achieved good results compared to the baseline model, although the improvement was smaller. In our analysis, we considered the following factors:

1. The STC data volume is very large and the scenario is single and simple. The model extracted enough information from the dataset species to complete the conversation task on that dataset.
2. The microblogging dataset of STC is rather heterogeneous, and there are a large number of identical contexts corresponding to different target responses in the training set, which means that context and response appear to be many-to-one.

Table 5. Results on Original STC. For each evaluation metric in every column, the best result is highlighted in boldface.

Model	**BLEU-1**	**BLEU-2**	**BLEU-4**	**ROUGE-2**	**D-1**	**D-2**
GPT-LCCC-Base-Oridata	11.88	6.46	3.21	3.61	15.26	22.44
GPT-LCCC-Base+KAT-Oridata	**12.80**	**7.29**	**3.79**	**4.33**	**15.33**	**23.81**

These effects lead the model to learn some generative patterns by keywords but lacked additional information about the context, such as the speaker's interest in the topic, the response time, and the environment in which they are located. It is difficult to correctly identify which response to use in the current environment. Overall, further enhancements were also achieved on this dataset by our approach, demonstrating the effectiveness of our module on large datasets as well as confirming its generalization performance.

Setting 2: From Table 6, it can be observed that we could still achieve a good improvement of the model with lower quality data, which verified the robustness of the model.

Table 6. Results on Clean STC. For each evaluation metric in every column, the best result is highlighted in boldface.

Model	**BLEU-1**	**BLEU-2**	**BLEU-4**	**ROUGE-2**	**D-1**	**D-2**
GPT-LCCC-Base	14.11	7.46	3.96	3.93	15.20	23.07
GPT-LCCC-Base+Key Word	14.78	8.01	4.25	4.41	15.43	**24.09**
GPT-LCCC-Base+Key Word+Multask-Ori	**14.87**	**8.08**	**4.29**	**4.45**	**15.56**	23.78

4.8. Case Study

Our KAT model produced a more empathetic and fluent response to the conversations. According to the base model, when the user expressed an interest in a conversation, the base model responded, "I'm not really interested". In Table 7, the model produced logical role disfluencies; "In my current job, you should remain patient." In the long dialog in Table 7,the model used keywords "Academic achievement, scholarships, being, student leaders", understood the current situation and produced contextually relevant responses, such as "Yes, I used to be class president", in line with the ground truth. The two base models produced responses that did not match the context at all; "Not only was I a student but I was also assigned to the School of Management." In Table 7, the model captured the key point in the context and produced a response that was the same as the ground truth, while the base model generated the safe-response problem of "Where is he going?".

Table 7. Generated sentences by our proposed method.

Context	speaker:你不介意的话，我想和你谈谈。 speaker: I would like to talk to you if you don't mind.
GPT-Base	我一点也不介意。 I don't mind at all.
GPT-Base+KAT	当然, 鲍勃。我很乐意。 Sure, Bob. I'd love to.
GPT2-Base	不太感兴趣。 Not really interested.
GPT2-Base+KAT	好吧, 当然。 Well, of course.
Keyword	不介意，想和你，谈谈 Don't mind. I want to talk with you.
Gold	你想跟我谈什么？ What do you want to talk to me about?
Context	speaker: 现在，奇尔兹小姐把你的简历交给了我，我有机会看了一遍，我必须说我印象深刻。 speaker: Now, Miss Childs handed me your CV and I had the opportunity to look it over and I must say I was impressed.
GPT-Base	非常感谢你。我会的。我能先告诉你一些情况吗? Thank you very much. I will. Can I tell you something first?
GPT-Base+KAT	非常感谢你。我试着保持简洁明了。如果有任何问题,请随时问我。 Thank you very much. I'll try to keep it simple and clear. If you have any questions, please feel free to ask me.
GPT2-Base	非常感谢你。我试着保持简洁明了。在我现在的工作中,你应该保持耐心。 Thank you very much. I try to keep it simple and clear. In my current job, you should be patient.
GPT2-Base+KAT	非常常感谢你。我试着保持简洁明了。如果有任何问题,请随时问我。 Thank you very often. I'll try to keep it simple and clear. If you have any questions, please feel free to ask me.

Table 7. *Cont.*

Keyword	简历，交给了我，看了，一遍，必须，印象深刻 CV, handed to me, read, over, must, impressed
Gold	非常感谢你。我试着保持简洁明了。如果有任何问题，请随时问我。 Thank you very much. I’ll try to keep it simple and clear. If you have any questions, please feel free to ask me.
Context	speaker:你的学习成绩如何？;listener:我一直做得很好。我曾两次获得奖学金。 ... speaker:你当过学生干部吗? speaker: How have you been doing in your studies?; listener: I have always done well. I have won scholarships twice. … speaker: Have you ever been a student leader?
GPT-Base	不仅我是学生，我还被分配到管理学院。 Not only was I a student but I was also assigned to the School of Management.
GPT-Base+KAT	是的，我以前是班上的班长。 Yes,I used to be class president.
GPT2-Base	是的，我是个消防员，我以前是个工程师。 Yes, I’m a firefighter and I used to be an engineer.
GPT2-Base+KAT	是的，我以前是班上的班长。 Yes, I used to be class president.
Keyword	学习成绩，奖学金，自学，经济学，工商管理，会计学，学生干部 Academic achievement, scholarship, self-study, economics, business administration, accounting, student leader
Gold	对，我曾经是我们班的班长。 Yes, I used to be the class president of my class.
Context	speaker:基思多大了？;listener:他21岁。詹姆斯多大了？ ... speaker:他很好。他上周退休了。这是他人生的转折点。现在他可以放松享受退休生活了。 speaker:How old is Keith? ; listener:He’s 21. How old is James? ... speaker: He’s fine. He retired last week. It was a turning point in his life. Now he can relax and enjoy his retirement.

Table 7. *Cont.*

GPT-Base	你和他一起工作多久了？ How long have you been working with him?
GPT-Base+KAT	他可以花更多的时间和孙子孙女在一起了。 He can now spend more time with his grandchildren.
GPT2-Base	他要去哪里? Where is he going?
GPT2-Base+KAT	他可以花更多的时间和孙子孙女在待在一起了。 He can now spend more time with his grandchildren.
Keyword	多大，大一岁，更年轻，父亲，人生，转折点，放松，享受，退休，生活 how old, one year older, younger, father, life, turning point, relax, enjoy, retirement, life
Gold	他可以花更多的时间和孙子孙女在一起 He can spend more time with his grandchildren

4.9. Discussion

In this section, we provide an analysis and discussion of our proposed model, the Keyword-Aware Transformers Network (KAT), from two perspectives: (1) the possible reasons why our model can achieve better performance, and (2) future improvements to the model.

Our proposed model jointly models the keyword extraction task and the dialog generation task in a multi-task learning fashion, fusing contextual keywords during the dialog generation process. Through our experiments on two Chinese open-domain dialog datasets, we observed that our KAT model outperformed the baseline in both semantic and non-semantic evaluation metrics, including Coherence, Fluency, and Informativeness, with faster convergence speed. We attribute this improved performance to the fact that our model more quickly and accurately extracts key information from the context through the incorporation of keyword information, resulting in more meaningful and informative responses. Furthermore, similar combinations of keywords often possess similar semantics, providing the model with useful cues for generating responses.

By analyzing specific cases, such as those shown in Table 4 and Table 7,, we determined that the keyword extraction task enabled our model to acquire key information from the context and to generate responses that were closely related to the context. For instance, by identifying the contextual keywords of “scholarship,” “academic record,” and “student leader,” the response “Yes, I used to be the class monitor” was generated, while the baseline model generated a grammatically correct, but largely unrelated, response. These results further demonstrated the effectiveness of our KAT model.

Although our experiments demonstrated the effectiveness of incorporating keyword information in open-domain conversation generation, we recognize that there is still much room for improvement. We believe that there is a wealth of information within dialogs that can be further integrated into our model to improve its performance. For example, dialogs often contain information about personality traits, educational backgrounds, and emotional states of the speakers. By further segmenting and categorizing the contextual information, we can enhance our model’s ability to capture relevant information and generate more diverse responses. How to effectively leverage this additional information and integrate it into our model is a promising direction for future research.

In summary, our proposed KAT model demonstrated its effectiveness in improving the quality of open-domain conversation generation through the incorporation of contextual keywords. However, we recognize that there is still much to be done to further enhance our model’s capabilities, and we believe that the exploration of additional contextual information is a promising direction for future research in this field.

5. Conclusions

We propose a model called KAT in this paper that uses context keywords to generate context-relevant, fluent, and confident responses for open-domain conversations. Keyword predicting and conversation generation are also modeled jointly. Experiments with the model showed competitive results on a variety of dataset types. Our analysis also proved the effectiveness of the proposed modules. In the future, we will explore how to generate open-domain conversations that better utilize sentiment, as well as emotional states and personal information.

Author Contributions: Conceptualization, Y.Z., X.C. and C.Z.; data curation, Y.Z. and C.Z.; writing—original draft preparation, Y.Z., A.Q. and H.W.; writing—review and editing, F.X., X.H. and W.C.; visualization, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported, in part, by the National Natural Science Foundation of China (No. 62206242), and the Zhejiang Basic Public Welfare Research Project (Grant No. LGF22F020015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Zhang, D.; Chen, X.; Xu, S.; Xu, B. Knowledge Aware Emotion Recognition in Textual Conversations via Multi-Task Incremental Transformer. In Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain, (Online), 8–13 December 2020; Scott, D., Bel, N., Zong, C., Eds.; International Committee on Computational Linguistics: Barcelona, Spain, 2020; pp. 4429–4440. https://doi.org/10.18653/v1/2020.coling-main.392.
2. Peng, W.; Hu, Y.; Xie, Y.; Xing, L.; Sun, Y. CogIntAc: Modeling the Relationships between Intention, Emotion and Action in Interactive Process from Cognitive Perspective. In Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2022, Padua, Italy, 18–23 July 2022; pp. 1–8. https://doi.org/10.1109/CEC55065.2022.9870410.
3. Huang, M.; Zhu, X.; Gao, J. Challenges in Building Intelligent Open-domain Dialog Systems. *ACM Trans. Inf. Syst.* **2020**, *38*, 21:1–21:32. https://doi.org/10.1145/3383123.
4. Chen, F.; Meng, F.; Chen, X.; Li, P.; Zhou, J. Multimodal Incremental Transformer with Visual Grounding for Visual Dialogue Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online Event, 1–6 August 2021; pp. 436–446.
5. Peng, W.; Hu, Y.; Xing, L.; Xie, Y.; Sun, Y.; Li, Y. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23–29 July 2022; pp. 4324–4330. https://doi.org/10.24963/ijcai.2022/600.
6. Mazaré, P.E.; Humeau, S.; Raison, M.; Bordes, A. Training Millions of Personalized Dialogue Agents. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Brussels, Belgium, 31 October–4 November 2018; pp. 2775–2779. https://doi.org/10.18653/v1/D18-1298.
7. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. https://doi.org/10.1007/s11431-020-16473.
8. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
9. 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, Online, 5–10 July 2020; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 85–96. https://doi.org/10.18653/v1/2020.acl-main.9.
10. 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. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 270–278. https://doi.org/10.18653/v1/2020.acl-demos.30.
11. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E.M.; Boureau, Y.L.; et al. Recipes for Building an Open-Domain Chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Online, 19–23 April 2021; Main Volume; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24.
12. Adiwardana, D.; Luong, M.; 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* **2020**, arXiv:2001.09977.
13. Wang, W.; Huang, M.; Xu, X.S.; Shen, F.; Nie, L. Chat More: Deepening and Widening the Chatting Topic via A Deep Model. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR'18, Ann Arbor, MI, USA, 8–12 July 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 255–264. https://doi.org/10.1145/3209978.3210061.
14. Zhang, Z.; Han, X.; Zhou, H.; Ke, P.; Gu, Y.; Ye, D.; Qin, Y.; Su, Y.; Ji, H.; Guan, J.; et al. CPM: A large-scale generative Chinese Pre-trained language model. *AI Open* **2021**, *2*, 93–99. https://doi.org/10.1016/j.aiopen.2021.07.001.
15. Zhang, Z.; Gu, Y.; Han, X.; Chen, S.; Xiao, C.; Sun, Z.; Yao, Y.; Qi, F.; Guan, J.; Ke, P.; et al. CPM-2: Large-scale cost-effective pre-trained language models. *AI Open* **2021**, *2*, 216–224. https://doi.org/10.1016/j.aiopen.2021.12.003.
16. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. Pre-trained models: Past, present and future. *AI Open* **2021**, *2*, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002.
17. Wang, H.; Li, J.; Wu, H.; Hovy, E.; Sun, Y. Pre-Trained Language Models and Their Applications. *Engineering* 2022, *in press* https://doi.org/10.1016/j.eng.2022.04.024.
18. Kim, Y.; Kim, J.H.; Lee, J.M.; Jang, M.J.; Yum, Y.J.; Kim, S.; Shin, U.; Kim, Y.M.; Joo, H.J.; Song, S. A pre-trained BERT for Korean medical natural language processing. *Sci. Rep.* **2022**, *12*, 1–10.
19. Wang, Y.; Ke, P.; Zheng, Y.; Huang, K.; Jiang, Y.; Zhu, X.; Huang, M. A Large-Scale Chinese Short-Text Conversation Dataset. In Proceedings of the Natural Language Processing and Chinese Computing, Zhengzhou, China, 14–18 October 2020; Zhu, X., Zhang, M., Hong, Y., He, R., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 91–103.
20. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv* **2017**, arXiv:1710.03957.

21. Shang, L.; Lu, Z.; Li, H. Neural responding machine for short-text conversation. *arXiv* **2015**, arXiv:1503.02364.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
23. Zhou, H.; Ke, P.; Zhang, Z.; Gu, Y.; Zheng, Y.; Zheng, C.; Wang, Y.; Wu, C.H.; Sun, H.; Yang, X.; et al. EVA: An Open-Domain Chinese Dialogue System with Large-Scale Generative Pre-Training. *arXiv* **2021**, arXiv:2108.01547
24. Gu, Y.; Wen, J.; Sun, H.; Song, Y.; Ke, P.; Zheng, C.; Zhang, Z.; Yao, J.; Zhu, X.; Tang, J.; et al. EVA2.0: Investigating Open-Domain Chinese Dialogue Systems with Large-Scale Pre-Training. *arXiv* **2022**, arXiv:2203.09313.
25. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds. Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
26. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
27. Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; Ma, W.Y. Topic Aware Neural Response Generation. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31. https://doi.org/10.1609/aaai.v31i1.10981.
28. Wu, C.H.; Zheng, Y.; Wang, Y.; Yang, Z.; Huang, M. Semantic-Enhanced Explainable Finetuning for Open-Domain Dialogues. *arXiv* **2021**, arXiv:2106.03065
29. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 27 November–1 December 2017; Asian Federation of Natural Language Processing: Taipei, Taiwan, 2017; Volume 1, pp. 986–995.
30. 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; Association for Computational Linguistics: Philadelphia, PA, USA, 2002; pp. 311–318. https://doi.org/10.3115/1073083.1073135.
31. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; Association for Computational Linguistics: Cedarville, OH, USA, 2004; pp. 74–81.
32. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A Diversity-Promoting Objective Function for Neural Conversation Models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; Association for Computational Linguistics: Cedarville, OH, USA, 2016; pp. 110–119. https://doi.org/10.18653/v1/N16-1014.
33. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The Curious Case of Neural Text Degeneration. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
34. Zheng, Y.; Zhang, R.; Huang, M.; Mao, X. A Pre-Training Based Personalized Dialogue Generation Model with Persona-Sparse Data. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA. 7–12 February 2020; Volume 34, pp. 9693–9700. https://doi.org/10.1609/aaai.v34i05.6518.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.