PK-Chat: Pointer Network Guided Knowledge Driven Generative Dialogue Model

Cheng Deng[1], Bo Tong[1], Jiaxin Ding[1], Luoyi Fu[1], Xinde Cao[1], Xinbing Wang[1], Chenghu Zhou[1,2]
[1]Shanghai Jiao Tong University,
[2]Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences
davendw@sjtu.edu.cn
Corresponding authors: Jiaxin Ding and Luoyi Fu

ABSTRACT

In the realm of end-to-end dialogue systems research, leveraging real-world knowledge to produce natural, fluent, and human-like responses with accurate information is paramount. However, domain-specific conversational dialogue systems often grapple with coherence issues and may inadvertently incorporate incorrect external information while responding to queries. This challenge stems from out-of-vocabulary problems and misinformation arising from the parameters of the neural network. In this study, we present **PK-Chat**, a groundbreaking **P**ointer network guided **K**nowledge-driven generative dialogue model. PK-Chat integrates a unified pre-trained language model with a pointer network for knowledge graphs. The words generated by PK-Chat within the dialogue are derived from predictions made using word lists and direct information from the external knowledge graph. Furthermore, we employ PK-Chat to develop a dialogue system tailored for academic contexts, specifically within the field of geosciences. To assess the quality of dialogue systems in academic scenarios, we establish an academic dialogue benchmark. The source code for our work is publicly available online [1].

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

Dialogue System, Pointer Network, Academic Knowledge Graph, Natural language Generation

ACM Reference Format:
Cheng Deng[1], Bo Tong[1], Jiaxin Ding[1], Luoyi Fu[1], Xinde Cao[1], Xinbing Wang[1], Chenghu Zhou[1,2]. 2023. PK-Chat: Pointer Network Guided Knowledge Driven Generative Dialogue Model. In *Proceedings of The 12th International Joint Conference on Knowledge Graphs (IJCKG '23).* ACM, Tokyo, Japan, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Developing dialogue systems using advanced language models like PLATO and GPT [1, 15] has become a prominent research direction. Fine-tuning on such models enables the generation of human-like conversational responses. However, existing generative dialogue systems often produce general-purpose responses, leading to a lack of domain expertise and semantic coherence in the replies [22].

To address this issue, researchers have integrated knowledge graphs such as Freebase [8] and Yago [6] into dialogue systems [24]. These knowledge graphs are embedded into vectors within latent semantic spaces, and these embeddings are utilized to generate relevant text candidate sets [20]. However, there is a challenge in maintaining semantic coherence, as the probability of each utterance neighbor candidate is calculated independently, without considering the relationship between the candidate utterances and the contextual input [14]. Additionally, existing systems face a hurdle when dealing with unseen or out-of-vocabulary words from the knowledge graph. Pretrained models lack position information for referenced knowledge, making it challenging to identify specific meanings if certain words do not appear or are unlikely to appear in the model's tokenizer. This limitation can lead to misusing memorized knowledge in generating responses [2].

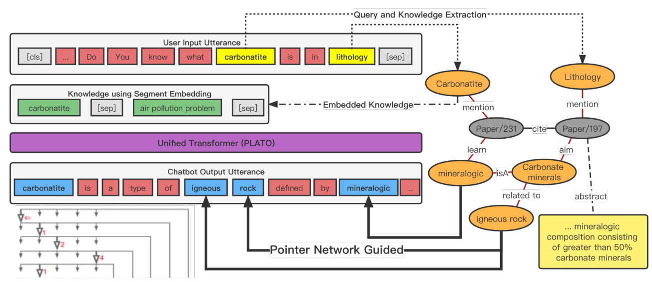

Figure 1: Overview of PK-Chat with GAKG [4].

Faced with the challenges of knowledge-driven generative dialogue models, we incorporate pointer networks [18, 21] to transfer information from the original input text to the output text while preserving detailed information. Consider text summarization, where words from the original corpus might not have been encountered during training. Models lacking a pointer network struggle to restore the original details, often resulting in generated summaries containing inaccuracies. In contrast, a pointer network enables the restoration of details by retaining original text information when encountering previously unseen words [18, 19].

In this study, we develop a knowledge-driven generative dialogue model guided by a pointer network. We train the model using GAKG [4] to facilitate natural and fluent knowledge-informed dialogues with users, explicitly focusing on geology-related knowledge. The system overview is depicted in Figure 1. The contributions of this paper can be summarized as follows:

1 This research paper introduces **PK-Chat**, a novel dialogue generation model that combines a pre-trained language model with a pointer generation network through a flexible self-attentive mechanism. By leveraging this advanced approach, PK-Chat demonstrates superior performance compared to established baselines across various benchmarks.

[1] https://github.com/iiot-tbb/Dialogue_DDE

2 PK-Chat innovatively adopts the pointer network with unified pre-trained language models to guide domain-specific conversation generation, a key contribution towards advancing the state-of-the-art in this domain.
3 Alongside PK-Chat, we construct GA-Dialogue, the first academic dialogue dataset with words sourced from the GAKG. The availability of this dataset represents a significant advancement, as it can be used to train other dialogue generation models, further contributing to the development of this promising field.

2 METHODOLOGY

In this section, we introduce **PK-Chat**, a novel dialogue generation model that leverages a unified pre-trained language model and pointer generation network through a self-attentive mechanism. PK-Chat is designed to generate fluent and natural text that draws on specific domain knowledge based on academic knowledge graphs. To achieve this, PK-Chat consists of a dialogue generation model, knowledge graph retrieval, and keyword extraction subsystems that work collaboratively to produce intelligent and context-sensitive responses to user inquiries.

First, when a user inputs an utterance, the model determines whether the input is relevant to the current referenced knowledge range. Suppose the current user utterance conflicts with the referenced knowledge, then the knowledge extraction algorithm will be activated to extract the text's keyword or entity information. The graph query statement is applied to query the specific connected triples information with the keyword of the entity in the knowledge graph, and all directly connected edges and tail entities with the node will be recalled as the knowledge input part of the model, which is combined with the utterance of the user to generate a response.

2.1 Dialogue Generation

In order to generate reasonable dialogue responses, the generative model should fully understand the above dialogue history information and background knowledge so that the responses are accurate and consistent enough.

Since the unification of bidirectional, unidirectional, and sequence-to-sequence objective functions enables us to straightforwardly fine-tune the pre-trained language UniLM[5] for both NLU and NLG tasks and dialogue tasks can benefit from it, we adopted a parameter sharing self-attention mechanism transformers like UniLM-PLATO based conversational language model as the backbone of the PK-Chat, which train from social media corpus. We fine-tune it with the conversational corpus from the data in GAKG illustrated in section 3. Besides, the loss functions for our task include the negative log-likelihood function used in the dialogue response generation and the pointer generation part, the bag-of-words model loss function used when predicting the words that should be in the responses, and the cross-entropy loss function used for topic switching.

Response Generation. For a given contextual information c and a selected hidden variable z, the reply is given as $P(r \mid c, z)$ based on this. Where $z \in [1, K]$, each specific z value corresponds to a potential semantic behavior, and the identification of the corresponding hidden variable can be accomplished by $\operatorname{argmax}(P(z \mid r, c))$ for the given contextual information and response content.

In PK-Chat, the response generation consists mainly of discrete hidden variables, content, and knowledge information. Moreover, take the maximum likelihood estimation function as the loss function like Equation 1.

$$\begin{aligned} \mathcal{L}_{NLL} &= -\mathbb{E}_{z \sim p(\mathbf{z} \mid c,k,r)} \log p(r \mid c, k, z) \\ &= -\mathbb{E}_{z \sim p(\mathbf{z} \mid c,k,r)} \sum_{t=1}^{T} \log p\left(r_t \mid c, k, z, r_{<t}\right), \end{aligned} \tag{1}$$

where z is the discrete hidden variable obtained from (c, k, r) and based on the probability $p(\mathbf{z} \mid c, k, r)$ are sampled. The hidden variable identification task obtains the distribution of the posterior probabilities of the hidden variables. c is the conversation information above, and k is the external knowledge information. And $p(\mathbf{z} \mid c, k, r)$ is a softmax activate function as Equation 2,

$$p(\mathbf{z} \mid c, k, r) = \operatorname{softmax}\left(W_1 h_{[M]} + b_1\right) \in \mathbb{R}^K, \tag{2}$$

where $z \in \mathbb{R}^K$, $h_{[M]} \in \mathbb{R}^K$ are the status tokens for the last layer of special status marker location. $W_1 \in \mathbb{R}^{K \times D}$ and $b_1 \in \mathbb{R}^K$ denotes the trainable parameters.

For each word w in the response generation, the prediction is made by the word corresponding word list, context, and knowledge-embedded information. PK-Chat adopts the pointer network, making the references to external knowledge more accurate. The probability is calculated as Equation 3,

$$P(w) = \lambda_{gen} P_{\text{vocab}}(w) + \left(1 - \lambda_{gen}\right) \sum_{i: w_i = w} a_i^t, \tag{3}$$

where $\lambda_{gen} = \operatorname{sigmoid}\left(W_2 h_D + b_2\right)$, W_2 and b_2 are trainable parameters, h_D is a hidden state of the intermediate generation result, and a is denoted as the prediction of the knowledge embedding location among the context and the pointer.

In addition to the negative log-likelihood estimation of the direct task goal of generating dialogue responses, the loss function of the bag-of-words model is added to the model training process to achieve the learning of the hidden variable z by predicting the words in the bag of words, specifically by predicting the words that should be in the responses through the hidden state of the last layer of z. Such a multi-task model can also accelerate the convergence speed of the model training. And the loss function is Equation 4.

$$\begin{aligned} \mathcal{L}_{BOW} &= -\mathbb{E}_{z \sim p(\mathbf{z} \mid c,k,r)} \sum_{t=1}^{T} \log p\left(r_t \mid c, k, z\right) \\ &= -\mathbb{E}_{z \sim p(\mathbf{z} \mid c,k,r)} \sum_{t=1}^{T} \log \frac{e^{f_{r_t}}}{\sum_{v \in V} e^{f_v}}, \end{aligned} \tag{4}$$

where V represents the size of the word list, f is the softmax function $f = \operatorname{softmax}\left(W_3 h_z + b_3\right) \in \mathbb{R}^{|V|}$ that predicts the words in the target generation, and f_{r_t} represents the probability value of the words generated at each moment. This prediction does not correlate to the order of each word but to the intention of making the hidden state variables capture more global information through this learning approach.

Topic Switch. In practice, we should select the correct external knowledge under the appropriate topic and judge whether we need to switch knowledge by comparing the current user's words and the context. Therefore, the judgment of topic switching is necessary, and we can select different knowledge at the appropriate moment. In this model, topic switching is a binary classification task to classify whether the current knowledge matches the current user utterance. If it does, the current topic knowledge is maintained and keeps chatting on the current topic. When the current knowledge does not match the question the user asks, the keyword extraction module is triggered, and the corresponding entity and edge information of the extracted keyword are queried in the knowledge graph. We choose the cross-entropy loss function as Equation 5.

$$\mathcal{L}_{TS} = -\log p\left(l_{\text{true}} = 1 \mid k, c, r^{+}\right) - \log p\left(l_{\text{true}} = 0 \mid k, c, r^{-}\right). \quad (5)$$

Give the knowledge during the dialogue with $l_{\text{true}} = 1$, randomly sample the knowledge in the other topic, and label it as $l_{\text{true}} = 0$. Overall, the loss function of the whole model is:

$$\mathcal{L} = \mathcal{L}_{NLL} + \mathcal{L}_{BOW} + \mathcal{L}_{TS}, \quad (6)$$

where $\mathcal{L}_{NLL}$ acts directly on the generation purpose, $\mathcal{L}_{BOW}$ acts on the hidden state learning and assists in the generation task. The $\mathcal{L}_{TJ}$ is used for topic classification, so the whole model uses a multi-task learning method.

2.2 Keyword Extraction

When the user's utterances mention entities that are in the knowledge graph, the critical information will be extracted via rule-based keyword extraction method, TF-IDF [17], TextRank [13] and BiLSTM+CRF [7] NER methods to extract the current entity during the communication with the user.

- We use a rule-based method by constructing regular expressions like *"(what|which|where)(is|are)(the)[a-z]{0,5}?"* to match the questioning phrase, which can quickly locate the corresponding keyword.
- We use TF-IDF and TextRank to obtain the most important words by multiplying the word frequency of a word and its inverse document frequency to indicate the importance of a word.
- We also use the BiLSTM-CRF model that defines the knowledge information extraction of user conversations as a sequence annotation task for keyword extraction.

Regarding the keyword extraction training data, TF-IDF and TextRank are unsupervised methods that do not require constructing labeled data for training. Hence, the dataset construction for information extraction mainly enables the BILSTM+CRF model to perform well on this keyword extraction task. According to the characteristics of the dialogue data in this paper, the entity information in the dialogue refers to the entities in the knowledge graph, so the annotation task in this part does not need much manual annotation. We only need to search and locate the entities in the dialogue and do the automatic annotation.

2.3 Retrieve over Knowledge Graph

In order to ensure the efficiency of the knowledge retrieval, we choose a reasonable storage method for the external knowledge graph. In this paper, we choose GAKG, an academic knowledge graph in geoscience, to deploy an academic dialogue system.

The GAKG is a collection of papers' illustrations, text, and bibliometric data. It is currently the largest and most comprehensive geoscience academic knowledge graph, consisting of more than 120 million triples with 11 kinds of concepts connected by 19 relations, stored in RDF format. We download the full copy of GAKG and store it in the graph database (Neo4J). After that, we built a GA-Dialogue dataset to train an academic chatbot. First, we randomly sampled all the information about the connected edges and tail entities of a single head entity on the knowledge graph of GAKG, constructed a specific dialogue scenario based on the sampled information, and started a specific dialogue around the information of the entity, i.e., we quoted the information of the entity in the dialogue to reply. In order to improve the quality of the dialogue dataset, we invited 20 geographers who understand the details of GAKG to participate in the construction and let them retain the label format of the entity. Five hundred fifty dialogue scenarios and 3,615 dialogues were constructed in GA-Dialogue. The average number of utterances of users per scenario is **6.7**.

However, the number of dialogue datasets is not large enough, so we increase the data by cleaning and constructing the public dataset. We used the Baidu DuConv [23] and Baidu DuRecDial [12] dialogue datasets as external datasets to introduce. For the DuConv dataset, there are 29,858 conversations in the scenes, with an average of **9** rounds of conversation per scene. In order to unify the data in this dialogue dataset with the dialogue data in our GAKG, we construct two types of knowledge in the conversation dataset: *conversation goals and* knowledge. We integrate the conversation goals and knowledge aggregated in the conversation dataset into the knowledge as the unified external knowledge. For the DuRecDial dataset, there are 10,200 conversations in the scenes, with an average of **15** rounds of conversation per scene. There are three types of valuable knowledge in the conversation dataset: conversation goals, knowledge, and user profile. We integrate the conversation goals, knowledge, and user profile aggregated in the conversation dataset as the knowledge. In this way, the data format is aligned with the conversation format of GAKG.

We sampled a few data, and finally, GA-Dialogue has 1,000 dialogue scenarios and 8219 dialogue rounds, with an average of **8.21** dialogue rounds per scenario.

3 EXPERIMENT

In this section, we evaluate the automatic evaluation results and the human evaluation results of the model of the dialogue system. This section details the models' benchmarks and evaluation metrics in experimental setup and evaluation results.

3.1 Experimental Setup

In this subsection, we briefly introduce the benchmarks, baselines, and metrics we selected to experiment with our model.

Benchmarks. We choose Persona-Chat [25] and DailyDialog [10] as the general benchmark, and we build an Academic Knowledge-Graph-based dialogue benchmark GA-Dialogue.

- Persona-Chat is a dataset of knowledge-based conversations on persona profiles (background knowledge).

		Automatic Evaluation			Human Evaluation				
Dataset	Model	BLEU-1/2	Distinct-1/2	Knowledge R/P/F1	Readability	Relevance	Consistency	Informativeness	Naturalness
GA-Dialogue (part1)	PLATO (Unidirect)	0.054/0.042	0.099/0.270	0.002/0.011/0.003	0.60	0.50	0.502	0.40	0.37
	PLATO	0.415/0.354	**0.165**/0.361	0.099/0.218/0.124	2.67	2.10	2.23	2.37	**2.20**
	PK-Chat (Ours)	**0.636/0.532**	0.139/**0.366**	**0.100/0.228/0.128**	**2.73**	**2.26**	**2.40**	**2.43**	1.90
GA-Dialogue (part2)	PLATO (Unidirect)	0.106/0.086	0.050/0.137	0.002/0.048/0.003	0.00	0.03	0.03	0.03	0.03
	PLATO	0.342/0.268	**0.105/0.322**	0.022/0.246/0.040	**2.93**	**2.53**	2.50	2.63	2.33
	PK-Chat (Ours)	**0.496/0.383**	0.065/0.237	**0.044/0.273/0.074**	2.80	2.43	**2.57**	**2.80**	**2.60**
Persona-Chat	PLATO (Unidirect)	-	0.003/0.010	0.018/0.084/0.028	0.30	0.24	0.20	0.27	0.13
	PLATO	0.231/0.178	0.014/0.053	**0.028/0.138/0.044**	**2.83**	**2.17**	**2.30**	1.97	2.33
	PK-Chat (Ours)	**0.257/0.199**	**0.015/0.062**	0.026/0.131/0.042	2.57	1.96	2.17	**2.03**	**2.34**
DailyDialog	PLATO (Unidirect)	0153/0.117	0.042/0.153	-	2.71	1.97	1.67	1.33	1.27
	PLATO	0.388/0.304	**0.055/0.303**	-	2.57	2.33	2.10	2.07	1.90
	PK-Chat (Ours)	**0.416/0.329**	0.049/0.282	-	**2.77**	**2.71**	**2.53**	**2.63**	**2.41**

Table 1: Comparison with baselines.

- DailyDialog is a chitchat dataset containing high-quality human conversations about everyday life.
- GA-Dialogue (Part 1 & 2), we divide the GA-Dialogue test dataset into two parts: the first part (part1) of the set contains dialogue data on the knowledge graph of GAKG, and the second part (part2) of the dialogue evaluation comes from the evaluation of the dialogue data introduced by external dataset. The datasets are available at Github Repo.[2]

Baselines. We choose the PLATO [1] and PLATO (Unidirect) as the baselines since PLATO achieves the sota results for the known model size of the same scale and PLATO (Unidirectional) is chosen as the baseline model to analyze the effect of the unidirectional attention mechanism on the final generation of the model, it is consistent with the GPT [16] series in the model self-attention structure.

Metrics. Unlike task-oriented dialogue systems, open-domain dialogue systems are complicated to evaluate the performance of dialogue systems through a specific metric due to the flexibility of the dialogue. In general, the open-domain dialogue systems are measured through objective and subjective evaluations, and the automatic and human evaluation methods used in this paper compare each model.

We choose BLEU [3] (bilingual evaluation understudy), Distinct [9] and Knowledge [1] as the **Automatic Evaluation Metric**, and greater the metrics are the better the models perform.

- BLEU is used to evaluate the generation task determined by calculating the overlap between the generated responses and the n-gram of the tags. In this paper, we set n as 1 and 2.
- Distinct is set up to measure diversity rubric for evaluating generated sentences by counting the ratio of unique n-gram of the words. In this paper, we set n as 1 and 2.
- Knowledge determines whether the cited knowledge is correct or incorrect.

For the human evaluation method, the human evaluation includes five indicators as described in [11], and we use them as the **Human Evaluation Metrics** in this paper: Readability, Relevance, Consistency, Informativeness, and Naturalness. In each benchmark, 500 generated dialogues and their contextual information are randomly selected as evaluation data, and 20 geoscientists are invited to analyze the dialogue performance evaluation and score them from [0,1,2,3] points in the above five aspects.

[2] https://github.com/davendw49/PK-Chat

3.2 Experimental Result

In GA-Dialogue (part1), the PK-Chat model outperforms the PLATO model in BLEU, Distinct, and Knowledge metrics, and in GA-Dialogue (part2), it outperforms the PLATO model and the baseline model in BLEU metrics and Knowledge metrics. The PK-Chat outperforms the PLATO model in all five dimensions of the human evaluation metrics, with four of the highest metrics in the first part of the evaluation set and three of the highest in the second part. Thus, both the automatic and human evaluation metrics have improved.

Similarly, in the Persona-Chat dataset, the PK-Chat model outperforms the baseline models on the automatic measures BLEU and Distinct but slightly underperforms the PLATO model on the Knowledge measure. The PK-Chat model outperforms the PLATO model in terms of the human evaluation metrics, and our proposed method does not have any gain on the final dialogue generation in this part of the dialogue dataset since the dataset references the knowledge of the user task portrait of the knowledge part is rarely directly referenced when answering the questions.

As for DailyDialog, the proposed method in this paper outperforms the baseline model on this dataset for automatic and human evaluation metrics. For the information whose context is a historical conversation, the model in this paper can enhance the metrics. By observation on the dataset, compared with the Persona-Chat, the conversation content usually revolves around the same topic, and the coherence between the conversation and context of the chitchat is more robust, so the model can replicate the learning of the words in the previous question through the pointer network so that the model will have a good performance effect.

4 CONCLUSION

This paper proposes PK-Chat, a knowledge graph-enhanced model via a unified pre-trained language model and pointer generation network to realize academic dialogues, aiming to develop a fluent, natural, and knowledge-informative dialogical interaction with scholars. By combining a unified pre-trained language model and a pointer network, the model could accurately refer to the knowledge mentioned in the KGs. Moreover, we put forward a GA-Dialogue as a benchmark to evaluate dialogue agents.

ACKNOWLEDGMENT

We would like to acknowledge supports from National key research and development plan No. 2022YFB3904204, National Natural Science Foundation of China under Grant (No. 42050105, 62202299, 62020106005, 62061146002, 61960206002), Natural Science Foundation of Shanghai No. 22ZR1429100, Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University. The first author would like to thank Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University. Moreover, this work is also a contribution to the Deep-time Digital Earth (DDE) Big Science Program.

REFERENCES

[1] Siqi Bao, H. He, Fan Wang, and Hua Wu. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *ACL*.

[2] Siqi Bao, H. He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In *FINDINGS*.

[3] Boxing Chen and Colin Cherry. 2014. A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. In *WMT@ACL*.

[4] Cheng Deng, Yuting Jia, Hui Xu, Chong Zhang, Jingyao Tang, Luoyi Fu, Weinan Zhang, Haisong Zhang, Xinbing Wang, and Chenghu Zhou. 2021. GAKG: A Multimodal Geoscience Academic Knowledge Graph. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (2021).

[5] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. arXiv:1905.03197 [cs.CL]

[6] MS Fabian, Kasneci Gjergji, WEIKUM Gerhard, et al. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*. 697–706.

[7] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv* abs/1508.01991 (2015).

[8] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web* 6, 2 (2015), 167–195.

[9] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL*.

[10] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP*.

[11] Hongru Liang and Huaqing Li. 2021. Towards Standard Criteria for human evaluation of Chatbots: A Survey. *ArXiv* abs/2105.11197 (2021).

[12] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *ACL*.

[13] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *EMNLP*.

[14] Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. 2019. Are Training Samples Correlated? Learning to Generate Dialogue Responses with Multiple References. In *ACL*.

[15] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. In *ArXiv*.

[16] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[17] Juan Enrique Ramos. 2003. Using TF-IDF to Determine Word Relevance in Document Queries.

[18] A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *ArXiv* abs/1704.04368 (2017).

[19] Fei Sun, Peng Jiang, Hanxiao Sun, Changhua Pei, Wenwu Ou, and Xiaobo Wang. 2018. Multi-Source Pointer Network for Product Title Summarization. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (2018).

[20] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. *ArXiv* abs/1904.09223 (2019).

[21] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks. In *NIPS*.

[22] Yanmeng Wang, Ye Wang, Xingyu Lou, Wenge Rong, Zhenghong Hao, and Shaojun Wang. 2021. Improving Dialogue Response Generation Via Knowledge Graph Filter. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), 7423–7427.

[23] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive Human-Machine Conversation with Explicit Conversation Goal. In *ACL*.

[24] Shiquan Yang, Rui Zhang, and Sarah Monazam Erfani. 2020. GraphDialog: Integrating Graph Knowledge into End-to-End Task-Oriented Dialogue Systems. In *EMNLP*.

[25] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur D. Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *ACL*.