Read Then Respond: Multi-granularity Grounding Prediction for Knowledge-Grounded Dialogue Generation

Yiyang Du[1(✉)], Shiwei Zhang[2], Xianjie Wu[1], Zhao Yan[3], Yunbo Cao[3], and Zhoujun Li[1]

[1] State Key Lab of Software Development Environment, Beihang University, Beijing, China
duyiyang@buaa.edu.cn
[2] Baidu Inc., Beijing, China
[3] Tencent Cloud Xiaowei, Beijing, China

Abstract. Retrieval-augmented generative models have shown promising results in knowledge-grounding dialogue systems. However, identifying and utilizing exact knowledge from multiple passages based on dialogue context remains challenging due to the semantic dependency of the dialogue context. Existing research has observed that increasing the number of retrieved passages promotes the recall of relevant knowledge, but the performance of response generation improvement becomes marginal or even worse when the number reaches a certain threshold. In this paper, we present a multi-grained knowledge grounding identification method, in which the coarse-grained selects the most relevant knowledge from each retrieval passage separately, and the fine-grained refines the coarse-grained and identifies final knowledge as grounding in generation stage. To further guide the response generation with predicted grounding, we introduce a grounding-augmented copy mechanism in the decoding stage of dialogue generation. Empirical results on MultiDoc2Dial and WoW benchmarks show that our method outperforms state-of-the-art methods.

Keywords: Knowledge-grounded dialogue · Retrieval-augmented · Grounding prediction

1 Introduction

Dialogue generation task faces the problem of producing non-informative or hallucinatory response [10,18]. Inspired by the retrieval-then-generation framework in open-domain QA [11,15,28], recent efforts have been made to address these concerns by knowledge-based dialogue generation. Those approaches typically

Y. Du and S. Zhang—Equal contribution.

X. Yang et al. (Eds.): ADMA 2023, LNAI 14177, pp. 291–306, 2023.
https://doi.org/10.1007/978-3-031-46664-9_20

involve knowledge searching, finding relevant knowledge according to the dialogue context, and producing final contextual responses [24,27].

Unexpectedly, as the number of retrieved documents increases, the performance of existing models either saturates or even degrades. As reported in [25,27], adding more knowledge documents to a vanilla response generation model leads to a more severe problem of hallucinations, i.e. plausible statements with factual errors. This might be because incorrectly retrieved passages with high lexical overlap with the input dialogue context can mislead the response generator, rather than providing reasonable knowledge. So, how to identify relevant knowledge from the numerous retrieved results to guide the response generation becomes a critical problem.

To identify relevant knowledge and improve response performance, paragraph-level methods are proposed to filter passages which contain knowledge related to the dialogue. EviGui-G [2] exclude noisy documents from retrieval results by predicting whether a retrieved document provides relevant evidence to response as an auxiliary task. Re^2G [9] purposes a retriever-ranker-generator framework to filter the retrieved knowledge fed to the generator and applies knowledge distillation to train the ranker and retriever jointly. DIALKI [29] extracts knowledge by first selecting the most relevant passage to the dialogue context and then selecting the final knowledge string within the selected passage to guide response generation. By selecting a exclusive span from multiple passages as grounding, this token-level method further locks the scope of relevant knowledge and achieves good results, especially in long documents. However, as a result of this method of selecting only one grounding, there is the risk of error propagation, which will contaminate the response once irrelevant knowledge is chosen.

In this paper, we propose a novel **M**ulti-**g**ranularity **G**rounding **G**uided **G**eneration (MG^4) model which introduces two types of token-level knowledge, namely coarse-grained and fine-grained groundings, and fuse them with weighted attention to encourage the generator to consider the importance of knowledge in different dialogue contexts. Our method has the ability to extract critical grounding information from a vast array of knowledge documents in a coarse-to-fine manner, thereby assisting in the generation of final responses. Furthermore, our experiments have shown that the coarse-to-fine approach outperforms any of its individual components. The framework imitates the process of human search for answers using a browser. Initially, it reads each relevant document retrieved and identifies the most relevant knowledge in each document as coarse-grained groundings for the query. Next, it assesses the importance of each piece of knowledge and combines the understanding from each document with a fine-grained grounding to generate a response.

Concretely, we introduce two distinct granularities of grounding: coarse-grained grounding and fine-grained grounding. The former aims to extract different spans of evidence from every retrieved passage through a question-and-answer system. To further identify the most relevant evidence from retrieved passages, we introduce a fine-grained grounding predicting task during the encoding

phase of generation, which can locate exclusive grounding as knowledge from all retrieved passages. Additionally, to enhance the guidance of responses by grounding, we devise a grounding-augmented copy mechanism during the decoding phase of generation to encourage the generator to utilize the predicted grounding when producing responses explicitly. Our experimental results demonstrate that different granularities of grounding can effectively direct the generator to improve response performance. Our best model achieves the state of the art on both MultiDoc2Dial [8] and WoW [20] at the time of writing. Our contributions are summed up as follows:

1. We propose a grounding-guided dialogue generation model based on two different granularities knowledge(coarse-grained and fine-grained).
2. We further incorporate grounding to guide generation by introducing a grounding-augmented copy mechanism, which give additional attention to two granularities grounding and the retrieved original paragraph text.
3. We achieve a new state-of-the-art on MultiDoc2Dial and WoW in automated metrics. Our method generates more accurate dialogue responses and alleviates hallucination problems in human evaluation and verification.

2 Related Work

Retrieval-augmented generation. The retrieval-augmented generator is a two-stage pipeline framework: (i) first to retrieve relevant passages from the knowledge source (the retriever) [3,12,13,26,30]; and (ii) second to generate an answer based on retrieved passages with the original query (the generator) [14,22]. RAG [15] retrieves relevant passages from external sources [13] and then generate the final response in a sequence-to-sequence style with marginalizing generation probabilities from different retrieved documents. FiD [11] retrieves a larger number of passages, encodes them independently, and then fuses the encoder results of multiple passages in the decoder phase. EMDR^2 [28] purpose an end-to-end training method and updates the retriever and reader parameters using an expectation-maximization algorithm. Recent work improves the retrieval component [19] or introduces passage re-ranking modules [7] for further improvements.

Knowledge-grounded Dialogues. Knowledge-grounded dialogue systems aim to generate knowledgeable and engaging responses based on context, and external knowledge [4,5,21,31–33]. EviGui-G [2] introduces a joint task whether a candidate passage provides relevant evidence to enhance the ability to identify gold passages. K2R [1] proposes a knowledge to response modular model to generate a knowledge sequence, then attends to its own generated knowledge sequence to produce a final response. To address knowledge identification in conversational systems with long grounding documents, DIALKI [29] extends multi-passage reader models in open question answering to obtain dense encodings of different spans in multiple passages in the grounding document, and it contextualizes them with the dialogue history. Re^{2}G [9] introduces a reranking

mechanism between the retriever and the reader, which permits merging retrieval results from sources with incomparable scores (Fig. 1).

3 Method

3.1 Overview

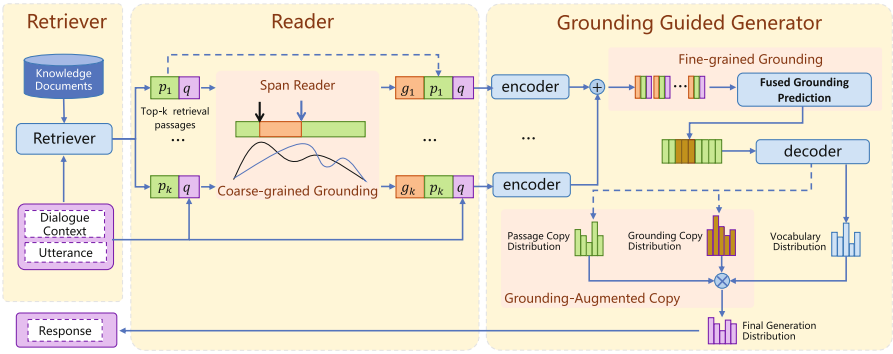

Fig. 1. Overview architecture of MG^4 framework

Problem Description. In knowledge-grounded response generation task, given a set of knowledge documents D, dialogue context U consisting of dialogue history $\{u_1, \ldots, u_{T-1}\}$ and user's utterance of current turn u_T , the goal is to generate response u_{T+1}. The probability of the generated responses can be written as:

$$p(y_t \mid P, U) = \prod_{t=1}^{n} p(y_t \mid P, U, y_1, \ldots, y_{t-1}) \tag{1}$$

where y_t is the t-th token in the agent response u_{T+1}, P is the split results of documents D. In order to distinguish the above D and P, we use "document" and "passage" respectively to denote the text of before and after segmentation. As the dialogue is knowledge-guided, the response is entailed by the grounding evidence in gold document among the provided multiple documents. In the gold passage related to the question, the grounding G_g is a span evidence to guide response generation. In this paper, we predict and exploit the grounding evidence in a multi-stage style to enhance the final response generation.

Method Overview. We propose a grounding-guided framework which extends the retrieval-augmented generation paradigm by adding a reader module to predict grounding - the token level evidence from retrieved passages to guide response generation. **Firstly**, the retriever retrieves the top-K knowledge passages (segments of the document) related to the last turn utterance and the

dialogue history. **Secondly**, the reader module (Sect. 3.2) is used to predict coarse-grained grounding evidence from every retrieved passage independently. It is worth noticing that coarse-grained grounding may be inaccurate considering that there is an error prediction of the reader, or the retrieved passage may not include the grounding evidence. **Thirdly**, the response generator finds the most relevant evidence from multiple retrieved passages. We propose using the grounding-guided encoder (Sect. 3.3) and copy-augmented decoder (Sect. 3.4) in the generator to produce the final response. The grounding-guided encoder uses the encoder representation of the generator to predict fine-grained grounding, and the copy-augmented decoder encourages the generator to borrow words from the predicted grounding explicitly.

3.2 Coarse-Grained Grounding Prediction in Reader

Firstly, by taking current utterance u_T with dialogue history $\{u_1, \ldots, u_{T-1}\}$ and a retrieved passage p_i as input, the grounding reader aims to infer important grounding evidence span from each retrieved passage p_i. We train our reader to use all the three tuples of dialogue context, gold passage, and grounding evidence span of gold passage in the training set. The grounding evidence span can be obtained in most cases since the response is written by human based on its provenance.

We use span-based reading comprehension model to predict coarse-grained grounding. The start and end probability are calculated by a linear projection from the last hidden states of reader's encoder:

$$\hat{\boldsymbol{p}}^{\,\text{start}} = \sigma(\varphi(H)) \quad \hat{\boldsymbol{p}}^{\,\text{end}} = \sigma(\varphi(H)) \tag{2}$$

where $\hat{\boldsymbol{p}}^{\,\text{start}}$ and $\hat{\boldsymbol{p}}^{\,\text{end}}$ is start and end probability distribution, H is the representation of reader's encoder, σ is softmax function and $\varphi(\circ)$ is MLP. The cost function is defined as :

$$J(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} \log\left(\hat{\boldsymbol{p}}_{y_t^s}^{\,\text{start}}\right) + \log\left(\hat{\boldsymbol{p}}_{y_t^e}^{\,\text{end}}\right) \tag{3}$$

where T is the number of training samples, y_t^s and y_t^e are the true start and end position of the t-th sample.

Then we use the well-trained reader to infer grounding evidence G for every retrieved passage in the training and evaluation set. The usage of coarse-grained grounding evidence G will be introduced in Sect. 3.3.

3.3 Fine-Grained Grounding Prediction in Generator Encoder

The generator is an encoder-decoder structure where the encoder part encodes every retrieved passage independently with the dialogue context and the coarse-grained grounding predicted in Sect. 3.2. The representation of j-th passage $h_{enc}^j \in R^{d \times l_j}$ can be calculated by encoder:

$$h_{enc}^j = \text{Encoder}\,(C; p_j; g_j) \tag{4}$$

where C is the dialogue context, p_j is the j-th passage and g_j is the predicted coarse-grained grounding in j-th passage from reader. The input form of j-th passage feed to the generator encoder can be described in detail as follows:

$$[\bar{S}_u, u_1, ...u_T; \bar{S}_p, p_j^0, p_j^1, ...\bar{S}_g, g_j^s, ...g_j^e, \bar{E}_g, ...p_j^{l_j}] \tag{5}$$

where u_T is the dialogue utterance of T-th turn, $p_j = \{p_j^0, \ldots, p_j^{l_j}\}$ is the context tokens of j-th retrieved passage with length l_j. $g_j = \{g_j^s, \ldots, g_j^e\}$ is coarse-grained grounding predicted by reader in the passage with the start and end position. $\bar{S}_u$, $\bar{S}_p$, $\bar{S}_g$, $\bar{E}_g$ are special tokens to indicate the start position of dialogue and passage context, the start and end position of coarse-grained grounding.

Fine-Grained Grounding Prediction. The encoder part of the generator incorporates fine-grained grounding prediction to identify the most relevant grounding evidence from all the retrieved passages to generate a response. Fine-grained grounding prediction can also fuse and denoise the coarse-grained groundings as it can be jointly trained with the response generation part. The error in coarse-grained groundings can arise from two sources: (1) errors in the prediction from the reader module, and (2) the possibility that the retrieved passage may not inherently contain any grounding evidence.

In the training phase, we consider that some retrieved passages may not contain the exact gold grounding evidence but rather similar useful information. Therefore, we use a token-level matching method to identify tokens present in the gold grounding and use them as fine-grained grounding labels. In the validation phase, the predicted grounding evidence is leveraged in the generator decoder part described in Sect. 3.4. The fine-grained grounding prediction is composed of a linear layer and a sigmoid function, which acts on the representation from the generator encoder. Since tokens included in the gold grounding accounts for a small proportion of the tokens in all retrieved passages, we sample negative tokens and apply focal loss [17] to train the grounding evidence prediction. The loss function can be defined as follows:

$$\boldsymbol{p}_{\mathrm{g}}(i) = \sigma(W_g h_i + b_g) \tag{6}$$

$$J(\boldsymbol{\theta}) = \sum_{y_i=1}^{M} \alpha \left(1 - \boldsymbol{p}_{\mathrm{g}}(i)\right)^{\gamma} \log \boldsymbol{p}_{\mathrm{g}}(i) + \sum_{y_i=0}^{N} (1-\alpha) \boldsymbol{p}_{\mathrm{g}}(i)^{\gamma} \log \left(1 - \boldsymbol{p}_{\mathrm{g}}(i)\right) \tag{7}$$

where h_i is the i-th position's representation from generator encoder, W_g and b_g are trainable parameters, σ is sigmoid function, $J(\boldsymbol{\theta})$ is the loss objective contributed by M positive grounding tokens and N negative tokens. α and γ is the hyperparameters in focal loss.

3.4 Copy Grounding Evidence in Generator Decoder

The decoder part of generator jointly decodes all encoded features of retrieved knowledge to generate response. We fuse the encoder inputs in a Fusion-in-Decoder style [11] to empower the decoder to attend all input passages and get

cross-attention result within a linear time complexity. As described in Sect. 3.3, the representation of j-th passage $h^j_{enc} \in R^{d \times l_j}$ can be calculated as follows :

$$h^j_{enc} = \text{Encoder}\,(C; p_j; g_j) \tag{8}$$

Then concatenate the h^j_{enc} to produce h_{enc} for decoder:

$$h_{enc} = h^1_{enc} \circ h^2_{enc} \circ h^3_{enc} \ldots h^K_{enc} \tag{9}$$

Our decoder is based on transformer style, so the cross-attention result can be calculated in the transformer layer itself:

$$e_{t,i} = \frac{(W_s s_t)^T W_h h_i}{\sqrt{d_k}} \tag{10}$$

$$\alpha_{t,i} = \text{softmax}\,(e_{t,i}) \tag{11}$$

where the h_i is the i-th position's representation of h_{enc}, s_t is the t-th step representation of h_{dec} calculated by self-attention and layer-normalization. W_s and W_h are learnable weights. d_k is the hidden size of k-th head, where we take out the last layer of transformer and the average of heads as the cross-attention output including cross-attention weights $e_{t,i}$ and cross-attention probs $\alpha_{t,i}$.

Grounding Augmented Copy Mechanism. We propose a grounding-augmented copy mechanism to encourage generator to explicitly borrow words from the predicted grounding. Let $L = \sum_{i=0}^{k} l_i$ denote the total encoder length after concatenation, g be the fine-grained grounding introduced by Sect. 3.3 to identify whether a token is present in gold grounding. The attention score from the response to predicted grounding can obtained by re-normalizing the cross attention weights in grounding token positions.

$$m_{t,i} = \begin{cases} 1, & g(i) = 1 \\ -\infty, & g(i) = 0 \end{cases} \tag{12}$$

$$n_{t,i} = \begin{cases} 1, & g(i) = 0 \\ -\infty, & g(i) = 1 \end{cases} \tag{13}$$

$$\beta_{t,i} = \text{softmax}\,(e_{t,i} \cdot m_{t,i}) \tag{14}$$

$$\gamma_{t,i} = \text{softmax}\,(e_{t,i} \cdot n_{t,i}) \tag{15}$$

The cross-attention probability from decoder time step t to token i in the fine-grained grounding evidence is denoted as $\beta_{t,i}$, while $\gamma_{t,i}$ represents the same probability for tokens in the other part of the passage except for the grounding evidence. This cross-attention probability can be used as a copied probability to contribute to the final probability distribution.

$$\boldsymbol{p}_{grounding}(w) = \sum_{i:x_i=w} \beta_{t,i}, \quad \boldsymbol{p}_{passage}(w) = \sum_{i:x_i=w} \gamma_{t,i} \tag{16}$$

where $P_{grounding}(w)$ is the vocabulary probability distribution by copying grounding evidence and $P_{passage}(w)$ is the vocabulary probability distribution by copying the other part of encoder input including passages and dialogue context. We reserve the distribution from passages and dialogue context because not all response words come from grounding and they may come from dialogue or other parts in passages. Then we add the copy vocabulary probability distribution to the generator vocabulary probability distribution with a learnable 3-way gate.

$$p_1, p_2, p_3 = \text{softmax}\left(W^3_{\text{gate}} \cdot h^{\text{dec}}_t + b^3_{\text{gate}}\right) \tag{17}$$

$$\boldsymbol{p}_{generate}(w) = \text{lm}_{head}\left(h^{dec}_t\right) \tag{18}$$

$$\boldsymbol{p}(w) = p_1 \cdot \boldsymbol{p}_{generate}(w) + p_2 \cdot \boldsymbol{p}_{grounding}(w) + p_3 \cdot \boldsymbol{p}_{passage}(w) \tag{19}$$

where $W^3_{\text{gate}} \in \mathbb{R}^{d\times 3}$, $b^3_{\text{gate}} \in \mathbb{R}^{d\times 3}$ are learnable parameters, lm_{head} is the output layer in transformer to calculate target vocab distribution. p_1, p_2, p_3 are 3-way gate probability. $\boldsymbol{p}(w)$ is the final target vocab distribution considering the contribution of generation, the predicted grounding and retrieved passages. According to $\boldsymbol{p}(w)$ to decode a word w step by step, the final response is generated.

4 Experiment

4.1 Datasets

MultiDoc2Dial. [8] is a new goal-oriented dialogue dataset based on multiple documents, containing 29,748 queries in 4800 dialogues with an average of 14 turns based on 488 documents from different domains. Each dialogue turn annotates the dialog data with the roles, dialogue behavior, human speech, and the grounding span with document information.

WoW. [6] is a large conversational dataset based on knowledge retrieved from Wikipedia. It covers a wide range of topics (a total of 1365), comprising 22311 dialogues and 201999 rounds. We verify the performance of our model on the WoW KILT version [20]. The KILT version requires model to find and fuse knowledge from all of Wikipedia pages rather than the provided knowledge candidates for each turn in original dataset, which is more suitable for our setting.

4.2 Baselines

RAG. [15] retrieves relevant passages from external sources and then generate the final response in a sequence-to-sequence style with marginalizing generation probabilities from different retrieved documents. **FiD** [11] Fusion-in-Decoder encodes all retrieved passages independently and then fuses the encoder result of multiple passages in the decoder phase. **EMDR**^{2} [28] provides an end-to-end approach to optimize retriever and generator parameters using model feedback

itself as "pseudo-labels" for latent variables. **DIALKI** [29] identifies the most relevant passage and grounding span in the passage from multiple documents and then only use the single passage and span to generate response. **EviGui-G** [2] incorporate evidentiality of passages and introduces a leave-one-out method to create pseudo evidentiality labels for model training.

Re^{2}G. [9] applies a retriever-ranker-generator framework to filter the retrieved knowledge fed to the generator and applies knowledge distillation to jointly train the ranker and retriever.

Table 1. Main Results. Results of automatic metrics on test set of MultiDoc2Dial and WoW. † denotes the model is based on T5-base while ‡ denotes T5-large and § denotes BART-large;

	MultiDoc2Dial		WoW	
	F1	R-L	F1	R-L
RAG §	34.25	31.85	13.11	11.57
FiD†	41.74	40.37	16.52	15.16
DIALKI §	38.95	37.64	17.04	15.65
EviGui-G†	43.14	41.33	17.30	15.93
EMDR^2 ‡	43.76	41.86	-	-
Re^{2}G §	44.26	42.40	18.90	16.76
MG^4 base†	45.30	43.38	18.69	16.80
MG^4 ‡	**45.72**	**43.94**	**19.28**	**17.26**

4.3 Experiment Setting

For the evaluation of our knowledge-based dialogue system, we evaluate the generated responses against the reference responses with automatic metrics, including token-level F1 score(F1) [23], Rouge-L(R-L) [16].

We report the results of RAG, FiD, and EviGui-G from [2], as well as Re2G from [9]. We reproduced the evaluation results using the same hyper-parameters, averaging over five runs with different seeds, and conducting a t-test with a p-value less than 0.05. The result of our model on the WoW dataset is from the KILT [20] version, which provides an online submission board[1].

To train the MG^4 model, we use the Adam optimizer with a learning rate of 5e-5. The number of top-k passages is set to 50. The input length of dialogue context and a single passage is set to 512, while the grounding span max length is set to 128, and the maximum response length is set to 50. α and γ in focal loss are set to 0.25 and 2.

[1] https://eval.ai/web/challenges/challenge-page/689/leaderboard/1909.

4.4 Quantitative Results

According to the Table 1, The end-to-end EMDR^2 model has a slight advantage over FiD by 1.12 F1 and 1.19 Rouge-L in MultiDoc2Dial, indicating that the end-to-end model may somewhat mitigate the problem of accumulating pipeline framework errors. Our MG^4 method outperforms nearly all benchmark results on both MultiDoc2dial and WoW. MG^4 outperforms DIALKI model on both Multidoc2Dial and WoW, indicating that selecting only one grounding from the most relevant passage has the risk of error propagation and multi-granularity grounding with weighted attention grounding copy mechanism can effectively identify multiple related information and improve the quality of generation. In particular, MG^4 outperforms the end-to-end model EMDR^2 by 2.46 F1 and 2.08 Rouge-L on MultiDoc2Dial, illustrating that grounding knowledge in retrieved passages can bring more performance gains than just training retriever and generator in an end-to-end way. Re^{2}G gets good performance on both MultiDoc2Dial and WoW by adding a reranker module to filter the retrieved knowledge, while MG^4 can also highlight some token-level evidence in retrieved knowledge and outperform Re^{2}G by 1.46 F1, 1.54 Rouge-L and 0.38 F1 and 0.50 Rouge-L.

Table 2. Ablation results. $\boldsymbol{G^{coarse}}$ denotes introducing the coarse-grained to the generator predicted by reader; ***Copy*** denotes introducing copy mechanism in the generator to copy words from fine-grained grounding. $\boldsymbol{G^{fine}}$ denotes fine-grained grounding prediction. **MG^4-CG** doesn't remove any module but replaces copying mechanism from fine-grained grounding with coarse-grained grounding.

	MultiDoc2Dial		WoW	
	F1	R-L	F1	R-L
MG^4	45.72	43.94	19.28	17.26
w/o G^{coarse}	43.63	41.81	17.72	16.38
w/o *Copy*	45.27	43.51	18.74	17.19
w/o G^{fine}	44.22	42.28	18.02	16.65
MG^4-CG	44.97	43.11	18.37	16.82

Table 2 presents the results of the ablation experiments. There is a clear drop when removing coarse-grounding, i.e. w/o G^{coarse}, illustrating the effectiveness of reader module and the influence of reader to generator. Removing grounding augmented copy mechanism, i.e. w/o *Copy*, drops the performance on both datasets, proving that copy mechanism can enhance the guidance from grounding to response generation. In w/o *Copy* setting, furtherly removing the fine-grained grounding prediction task, w/o G^{fine}, will continue to bring performance drop, indicating that training via joint tasks to predict evidentiality labels can bring help to the generation task. Finally, we conduct experiments on copying from coarse-grained grounding, i.e. MG^4-CG. It's performance is lower than MG^4, which can be understood as the superiority of fine-grained grounding compared to coarse-grained grounding in guiding response generation.

4.5 Human Evaluation

Human annotators are asked to evaluate our model by quantifying the three aspects of generated responses, as described below: (i) **Fluency**, a measure of whether the response is consistent and less repetitive. (ii) **Relevance**, which measures the relevance of the response to the dialogue context. (iii) **Factuality** measures the correctness and faithfulness of all facts involved in the generated response.

Table 3. Absolute human valuation results for MG^4 versus $EMDR^2$ on MultiDoc2Dial. The table presents each metric average value for all annotators and samples out of 3 points. The Fleiss' kappa between annotators is 0.58.

Model	Fluency	Relevance	Factuality
$EMDR^2$	2.54	2.33	2.13
MG^4	**2.67**	**2.73**	**2.51**

Table 4. Comparative evaluation results between MG^4 and $EMDR^2$, where the percentage indicates the proportion of preference by all evaluators.

Aspect	Win	Lose	Tie
Fluency	**32%**	20%	48%
Relevance	**57%**	13%	30%
Factuality	**62%**	22%	16%

We choose $EMDR^2$, which is the most important reference in terms of automatic measurements, for comparative purposes. We sample the evaluation dialogue turns from the MultiDoc2Dial, which is factually supported by knowledgeable customer service documents. Table 3 shows the absolute evaluation results of human annotation. To reduce the evaluation inconsistency caused by different evaluators, we also conduct a comparative evaluation with results shown in Table 4. We found that MG^4 outperformed $EMDR^2$ in both evaluation dimensions, indicating that MG^4 can improve knowledge utilization through our coarse-to-fine grounding prediction method and grounding-augmented copy mechanism. It is noteworthy that our model has a significant improvement on the factuality metric, demonstrating its ability to alleviate the dialogue hallucination problem to some extent.

5 Further Analysis

5.1 Can Grounding Guide the Response Generation?

Table 5. Oracle experiments to explore the upper bound impact of gold grounding on MultiDoc2dial and WoW dev dataset.

Model	Dataset	F1	R-L
FiD	MultiDoc	42.14	40.67
FiD with Gold-G	MultiDoc	51.39	49.79
FiD	WoW	16.15	15.86
FiD with Gold-G	WoW	29.53	28.46

We conduct two experiments to illustrate the influence of grounding information to final response generation.

Firstly, we introduce gold grounding to the generator in the way as Sect. 3.2 and conduct oracle experiments to explore the influence of grounding. As shown in Table 5, the grounding-guided model significantly improved in performance by 9.25 F1, 9.12 Rouge-L on MultiDoc2Dial and 13.38 F1, 12.6 Rouge-L on WoW. According to the experimental results, the introduction of gold grounding can significantly improve the generation performance in the knowledge-grounded dialogue generation task.

Secondly, we leverage a span-based model called reader in Sect. 3.2 to predict grounding as a replacement of gold grounding. According to the ablation experiment results in Table 2, we can find that the performance gain from the reader module is most notable, proving that knowledge-grounded dialogue can benefit from the coarse-grained grounding from the extractive reader module.

5.2 Why We Need a Multi-granularity Grounding Prediction?

In our paper, multi-granularity grounding includes coarse-grained and fine-grained grounding. We add a fine-grained grounding prediction introduced in Sect. 3.3 in the generator encoder to find most relevant evidence from all retrieved passages. Figure 2 shows an actual case of the predicted coarse-grained grounding and fine-grained grounding in WoW, and the bottom right shows their contribution to the final response by taking out the cross-attention weight (average in the output sequence dimension) in the generator. In Table 2, we compare the performance between MG^4-CG and MG^4 in which MG^4-CG means copying from coarse-grained grounding and MG^4 means copying from fine-grained grounding. The introduction of fine-grained grounding can improve 0.75 F1 and 0.83 Rouge-L in MultiDoc2dial as well as 0.91 F1 and 0.44 Rouge-L in WoW compared to coarse-grained grounding which illustrates the validity of our coarse-to-fine method to predict token level evidence from multiple retrieved passages.

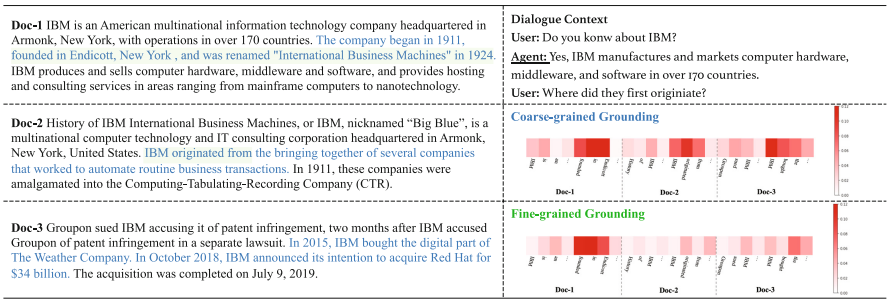

Fig. 2. Coarse-grained vs. Fine-grained. Retrieved passages are located on the left side. The blue portion is the coarse-grained grounding predicted from multiple passages while the light green portion is the fine-grained grounding predicted from grounding evidence denoiser. (Color figure online)

5.3 MG^4 Performs Better with More Passages

Figure 3 shows the Rouge-L score of our MG^4 model and the FiD model in different passage number settings. From the figure, we can see that the performance gains influenced by retrieved passage numbers is marginal as the number increases. It's worth noticing that our MG^4 model can get even higher improvement compared to FiD with larger retrieved passage numbers. The improvement is 1.92 Rouge-L in 10 passages setting and 3.27 Rouge-L in 50 passages setting. It can be interpreted as that larger number of retrieved passages means larger amount of relevant knowledge information as well as noise, which will bring more burden to the generator module. While our MG^4 can alleviate this problem by providing token-level multi-granularity grounding from retrieved passages to the generator.

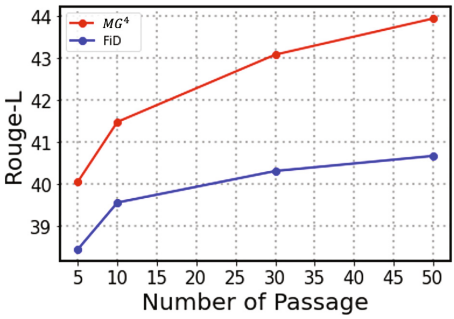

Fig. 3. Impact of the input passage number to response performance on MultiDoc2Dial.

6 Conclusion

In this work, our aim is to address the grounding identification issue in generating dialogues based on multiple documents. To achieve this goal, we propose a multi-granularity grounding prediction method in conjunction with a grounding-augmented copy mechanism that makes use of predicted key information from multiple documents. Our experimental results demonstrate that grounding information has a significant impact on guiding dialogue generation and that our proposed architecture, MG^4, can effectively utilize this information and mitigate the issue of hallucination in knowledge-based dialogue.

References

1. Adolphs, L., Shuster, K., Urbanek, J., Szlam, A., Weston, J.: Reason first, then respond: Modular generation for knowledge-infused dialogue. arXiv preprint arXiv:2111.05204 (2021)
2. Asai, A., Gardner, M., Hajishirzi, H.: Evidentiality-guided generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2112.08688 (2021)
3. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
4. Chen, X., et al.: Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3426–3437 (2020)
5. Davison, J., Feldman, J., Rush, A.M.: Commonsense knowledge mining from pretrained models. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1173–1178 (2019)
6. Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of Wikipedia: Knowledge-powered conversational agents. arXiv preprint arXiv:1811.01241 (2018)
7. Fajcik, M., Docekal, M., Ondrej, K., Smrz, P.: R2–d2: a modular baseline for open-domain question answering. arXiv preprint arXiv:2109.03502 (2021)
8. Feng, S., Patel, S.S., Wan, H., Joshi, S.: MultiDoc2Dial: modeling dialogues grounded in multiple documents. In: EMNLP (2021)
9. Glass, M., Rossiello, G., Chowdhury, M.F.M., Naik, A., Cai, P., Gliozzo, A.: Re2G: retrieve, rerank, generate. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2701–2715. Association for Computational Linguistics, Seattle, United States, July 2022. https://aclanthology.org/2022.naacl-main.194
10. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019)
11. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282 (2020)
12. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**, 1–11 (1972)
13. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)

14. Lewis, M., et al.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
15. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
16. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
18. Ma, Y., Nguyen, K.L., Xing, F.Z., Cambria, E.: A survey on empathetic dialogue systems. Inf. Fusion **64**, 50–70 (2020)
19. Paranjape, A., Khattab, O., Potts, C., Zaharia, M., Manning, C.D.: Hindsight: posterior-guided training of retrievers for improved open-ended generation. arXiv preprint arXiv:2110.07752 (2021)
20. Petroni, F., et al.: Kilt: a benchmark for knowledge intensive language tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2523–2544 (2021)
21. Prabhumoye, S., Hashimoto, K., Zhou, Y., Black, A.W., Salakhutdinov, R.: Focused attention improves document-grounded generation. arXiv preprint arXiv:2104.12714 (2021)
22. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)
23. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
24. Rashkin, H., Reitter, D., Tomar, G.S., Das, D.: Increasing faithfulness in knowledge-grounded dialogue with controllable features. arXiv preprint arXiv:2107.06963 (2021)
25. Reimers, N., Gurevych, I.: The curse of dense low-dimensional information retrieval for large index sizes. arXiv preprint arXiv:2012.14210 (2020)
26. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al.: Okapi at TREC-3. NIST Spec. Publ. SP **109**, 109 (1995)
27. Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J.: Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567 (2021)
28. Singh, D., Reddy, S., Hamilton, W., Dyer, C., Yogatama, D.: End-to-end training of multi-document reader and retriever for open-domain question answering. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
29. Wu, Z., Lu, B.R., Hajishirzi, H., Ostendorf, M.: DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 1852–1863. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, November 2021. https://doi.org/10.18653/v1/2021.emnlp-main.140, https://aclanthology.org/2021.emnlp-main.140
30. Xiong, L., et al.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv preprint arXiv:2007.00808 (2020)
31. Zhan, H., Zhang, H., Chen, H., Ding, Z., Bao, Y., Lan, Y.: Augmenting knowledge-grounded conversations with sequential knowledge transition. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5621–5630 (2021)

32. Zhang, S., Du, Y., Liu, G., Yan, Z., Cao, Y.: G4: Grounding-guided goal-oriented dialogues generation with multiple documents. In: Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pp. 108–114. Association for Computational Linguistics, Dublin, Ireland, May 2022. https://doi.org/10.18653/v1/2022.dialdoc-1.11, https://aclanthology.org/2022.dialdoc-1.11
33. Zhu, W., Mo, K., Zhang, Y., Zhu, Z., Peng, X., Yang, Q.: Flexible end-to-end dialogue system for knowledge grounded conversation. arXiv preprint arXiv:1709.04264 (2017)