Towards Faithful Dialogs via Focus Learning

Anonymous ACL submission

Abstract

Maintaining faithfulness between responses and knowledge is an important research topic for building reliable knowledge-grounded dialogue systems. Existing models heavily rely on elaborate data engineering or increasing the model's parameters ignoring to track the tokens that significantly influence losses, which is decisive for the optimization direction of the model in each iteration. To address this issue, we propose Focus Learning (FocusL), a novel learning approach that adjusts the contribution of each token to the optimization direction by directly scaling the corresponding objective loss. Specifically, we first introduce a positioning method by utilizing relevance distributions between knowledge and each response token to locate knowledge-aware tokens. Then, we further design a relevance-to-weight transformation to provide dynamic token-level weights for adjusting the cross-entropy loss. Finally, we use the weighted loss to encourage the model to pay special attention to the knowledge utilization. Experimental results demonstrate that our method achieves the new state-of-the-art results and generates more reliable responses while maintaining training stability.

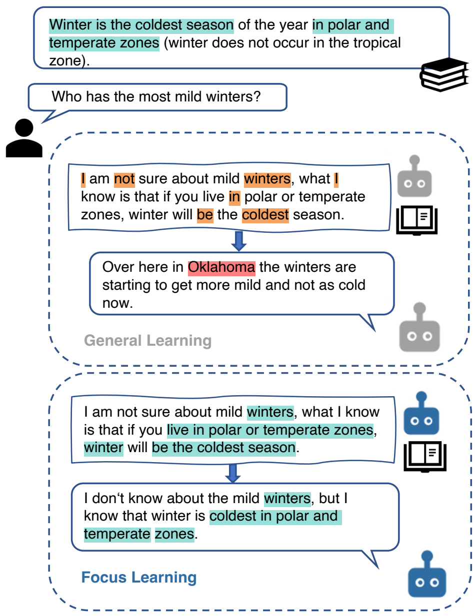


Figure 1: The different learning focus (i.e., the tokens that corresponding loss significantly influence total objective losses) between general learning and focus learning. Original learning focus without guidance in general learning are fragmented with no rules to follow. Our methods make the model focus on the knowledge-aware tokens (i.e., tokens that have high semantic similarity to knowledge) to alleviate the hallucinations.

1 Introduction

Although open-domain conversation systems can generate smooth and fluent responses with the help of large-scale pre-trained models (Raffel et al., 2020; Lewis et al., 2020), vacuous responses (Li et al., 2016) continue to be prevalent. To enrich the content of responses, an effective way is to introduce external knowledge (Dinan et al., 2019; Zhou et al., 2018). The knowledge-grounded model, however, frequently generates responses that appear knowledgeable but are not actually derived from the given knowledge. This means that the correctness of the knowledge used in responses cannot be guaranteed. As shown in Figure 1, the "Oklahoma" in the response is not present in the given knowledge and irrelevant knowledge is unverifiable. This phenomenon is known as *hallucination* (Dziri et al., 2022a). Due to the inability to verify knowledge, hallucinations will mislead users and reduce the model's credibility.

Numerous methods have been developed to tackle the hallucinations problem by knowledge graph (Kang et al., 2022; Dziri et al., 2021), contrast learning (Sun et al., 2022) or control code (Rashkin et al., 2021). These models enhance the model's attention to knowledge by increasing parameters or elaborate data engineering. An important assumption for them is that the model has the ability to give more attention to knowledge dur-

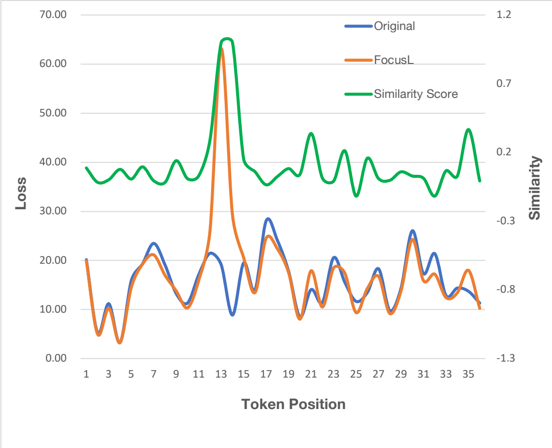


Figure 2: Distributions of the loss and the similarity between knowledge and response tokens. We select a response as an example and visualize the loss, semantic similarity to knowledge, and the adjusted loss (FocusL) at the beginning of training. In the original loss, the model is less sensitive to optimization of knowledge-aware tokens. In contrast, the loss of knowledge-aware tokens in FocusL are larger than the others, and the knowledge-irrelevant tokens' loss are scaled down.

ing training, yet this is not always held true. We consider this to be a common problem with general training methods which neglect to track the tokens that significantly influence the objective loss (i.e., learning focus). As the example shown in Figure 1, in general learning scenario, the learning focus is often out of control, and the model focus on the simple words (e.g., *be, the*), which lead to neglect of the tokens that have high relevance to knowledge referred as knowledge-aware tokens (e.g., *polar or temperate zones*). Intuitively, knowledge-aware tokens are even more critical for improving consistency, and focusing the model's attention on them can make the optimization goal fitter for the task. Therefore, it is necessary to revise the original learning focus. However, there are two main challenges: (1) How to locate the desired learning focus? Due to the fact that the learning focus is on different words in each sentence and the token-level manual annotation of responses is extremely time-consuming and labor-intensive, the existing datasets do not have a fine-grained annotation of key semantic words in the responses. (2) Given the desired learning focus, how to correct the original learning focus? Existing training methods with cross-entropy loss lack direct guidance on learning focus.

To address above issues, we propose a novel learning approach, **Focus L**earning (FocusL). Instead of impacting knowledge utilization implicitly, we directly scale the corresponding objective loss to adjust the contribution of each token to the optimization direction. Specifically, for the first challenge, we first define the desired learning focus in knowledge-grounded dialogue task as knowledge-aware tokens. Then we devise a positioning method to get the relevance score distribution between knowledge and each response token. For the second challenge, we explore a relevance-to-weight transformation method to provide dynamic token-level weights for the cross-entropy loss. Finally, we use the corrected learning focus to guide model training. As we can see in Figure 2, the losses of knowledge-aware tokens do not gain a high proportion of the original loss distribution. In contrast, our approach can expand the gap between knowledge-aware tokens and the others, which increases the impact of the change of knowledge-aware tokens' loss on the final loss, thus affecting the optimization direction and guiding the model to pay more attention to knowledge utilization.

Our main contributions are summarized as below:

- We rethink existing models and the learning method, and propose a novel learning approach to address the hallucination problem by directly adjusting the learning focus.

- We propose a positioning method and relevance-to-weight transformation method to adaptively scale the loss of each token in response.

- Experimental results demonstrate that our approach significantly outperforms the current state-of-the-art baselines, and effectively reduce hallucinations while maintaining high quality of responses.

2 Related Work

Knowledge-grounded dialogue generation. Knowledge-grounded dialogue systems aim to alleviate vacuous responses by injecting external knowledge into the dialogue model. Recently, various forms of external knowledge have been used in dialogue systems, such as tables (Moghe et al., 2018), graphs (Bollacker et al., 2008; Moon et al., 2019; Zhou et al., 2020), documents (Ghazvininejad et al., 2017; Zhou et al., 2018; Zhao et al., 2019). In spite of research on the forms of knowledge, most existing systems focus on knowledge selection (Lian et al., 2019; Kim et al., 2020; Zheng et al., 2020; Meng et al., 2020; Li et al., 2022) and response generation with given knowledge (Xu et al., 2020; Ma et al., 2020; Cai et al., 2020; Zhao et al., 2020). In this work, we mainly focus on avoiding models using unverifiable knowledge in response generation with given knowledge.

Hallucinations In text generation Generating responses that are unfaithful to the provided knowledge, known as hallucination, is a tricky problem in knowledge-grounded dialogue systems. Recently, the hallucination problem has attracted increasing attention because the generated hallucination text looks smooth and fluent but usually contains false knowledge, which significantly threatens the model's credibility. Some studies reduce hallucinations by introducing knowledge graph (Kang et al., 2022; Dziri et al., 2021), controlled generation (Rashkin et al., 2021), and contrast learning (Sun et al., 2022). In a recent study, Dziri et al. (2022b) analyzed the source of hallucination in detail and found that the commonly knowledge-grounded conversation datasets (Dinan et al., 2019; Zhou et al., 2018) inherently contains hallucinations, and models trained on such dataset further amplify the hallucinations, which demonstrate that the pattern of hallucinatory responses is more likely to be learned by the model. To address this problem, Dziri et al. (2022a) further proposed FaithDial, a new dataset that removes the hallucinations in the Wizard of WikiPedia (Dinan et al., 2019). Different from these studies about models and datasets, we found that the training method with unexpected learning focus also plays a vital role in the hallucination problem and then presented a method to revise the original focus.

3 Methods

3.1 Our Approach

The overview of FocusL is presented in Figure 3. Given the conversation context $C = (c_1, ..., c_n)$ consisting of a sequence of n dialogue turns and the corresponding knowledge $K = (k_1, ..., k_m)$ for the current turn, where m is the number of tokens in K, the goal of our task is to generate responses $Y = (y_1, ..., y_T)$ where T is the number of tokens in Y. We first form the input I with joint knowledge K and conversation context C as follows:

$$I = [K; C] \tag{1}$$

where the utterances of C are delimited by the speaker identifier (either <`user`> or <`bot`>). Then we use T5 (Raffel et al., 2020) as the base model, which is a pre-trained encoder-decoder model that uses a transformer architecture (Vaswani et al., 2017). Taking I as input, the base model outputs a logit distribution $h = (h_1, ..., h_T)$, where h_t is the corresponding logit distribution of the t-th token in Y. The positioning module locate knowledge-aware tokens in the response Y and calculate the corresponding adjust weight w_a. The focus shifting module adjusts the original logit distribution h to obtain the final logit distribution h_w. Finally, we train the model to produce the next conversation utterance $y_1...y_T$ by minimizing the cross-entropy loss.

In the following, we will detail introduce the three steps of the FocusL training process: (1) locate knowledge-aware tokens which used as new learning focus (§3.2); (2) calculate adjust weihgts based on the relevance of knowledge with each token in the response (§3.3); (3) switch original learning focus to the knowledge-aware tokens (§3.4).

3.2 Learning Focus Positioning

To adjust the learning focus of the model, we first define knowledge-aware tokens as the new learning focus which is more in line with the knowledge-grounded dialogue task. And then we use the distance between token and knowledge in semantic space to measure its relevance:

$$relevance(y_t^r, k) = \frac{y_t^r \cdot \mathcal{K}}{\|y_t^r\| \cdot \|\mathcal{K}\|} \tag{2}$$

To get the semantic representation of the token y_t and the knowledge k_i, we use the embedding layer $Emb(\cdot)$ of the base model to obtain a dense representation:

$$y_t^r = Emb(y_t) \tag{3}$$

$$\mathcal{K} = \frac{1}{m}\sum_{i=1}^{m} Emb(k_i) \tag{4}$$

Note that we do not use the model's encoder to obtain the representation vector of knowledge and responses, we think that the output of the embedding layer is sufficient to provide the needed semantic information, and also has less impact on the

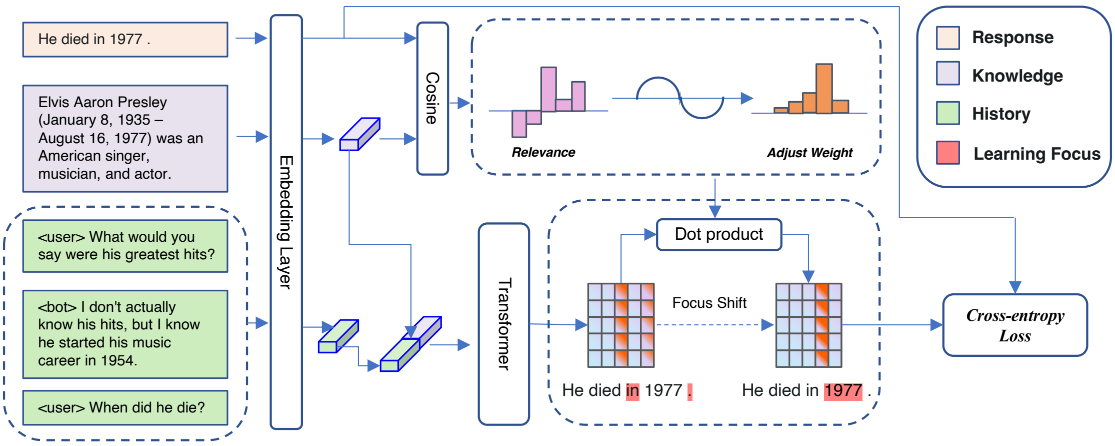


Figure 3: Training process of FocusL. We first calculate original model output based on the given knowledge and context. Then we calculate the relevance score between each token in response and knowledge, and further convert it to the adjust weight distribution. Finally, we use the adjust weight to scale the original loss and get the corrected loss.

training speed. Instead of outputting knowledge-aware tokens directly, the positioning method uses relevance matrix to provide more information for §3.3.

3.3 Adjust Weihgt

To adaptively assign a weight for each token to adjust the corresponding logit value, we can simply define adjust weight w_a as follows:

$$w_a = \begin{cases} 2, & \text{if } relevance(y_t^r, \mathcal{K}) \geq \theta \\ 1, & otherwise \end{cases} \quad (5)$$

where θ is a threshold value. We rigidly define knowledge-aware tokens by setting a specific θ. The tokens with relevance greater than the threshold are regarded as knowledge-aware tokens and obtain a high adjust weight to increase corresponding loss. We keep the original logit value unchanged for tokens with relevance less than the threshold.

However, the boundaries of knowledge-aware tokens are difficult to define, and the threshold value easily influences the learning effect of the model. To solve this problem, we further propose two different methods for converting relevance distribution into adjust weights.

Liner Weihgt To make full use of the information in the relevance matrix, we propose to assign a different adjust weight to each token. We obtain a non-negative distribution by following formula:

$$w_a = 1 + relevance(y_t^r, \mathcal{K}) \quad (6)$$

This method adaptively scales up the loss of knowledge-aware tokens while scaling down the loss of rest tokens. Although this can adjust the weights of all tokens, a linear weight distribution is not a good simulation due to the complexity of focus changes in real-world human learning. Meanwhile, the weights between knowledge-aware and irrelevant tokens are not significantly different, which does not have a large enough impact on the loss.

Non-linear Weihgt To ensure the stability of training, we aim to increase the loss of knowledge-aware tokens as much as possible while keeping that the adjusted final loss is not too different from the original loss. Therefore the distribution of weights should be smoother at low relevance interval and sharper at high relevance interval. We map the original linear distribution to a logarithmic distribution with following formula:

$$w_a = -\ln(1 - relevance(y_t^r, \mathcal{K}) + \lambda) + 1 \quad (7)$$

where $\lambda \in (0, e-2)$ is a small constant that we call it smoothing factor. A large smoothing factor represents a smoother distribution of the obtained weights.

3.4 Focused Cross-Entropy

After obtaining the adjust weight w_a, we scale the original logit and then use the new logit to calculate the probability of each token. At the time step t, given original model outputs h_t, the probability of the token y_t is calculated as follows:

$$p_w(y_t|y_{<t}, \mathcal{I}) = softmax(w_a \cdot h_t) \quad (8)$$

We define the final loss for optimization as the Focused Cross-Entropy(FCE) loss:

$$\mathcal{L}_{FCE} = -\frac{1}{T}\sum_{t=1}^{T} \log p_w(y_t|y_{<t}, \mathcal{I}) \qquad (9)$$

where T is the length of the response. FCE changes the original loss distribution, which leads the model to shift original learning focus to desired tokens. To reduce this loss function, the gradient descent approach is used to update all parameters.

4 Experiments

To evaluate the effectiveness of our method, we conduct experiments following the settings in (Dziri et al., 2022a). We use pre-trained T5 (Raffel et al., 2020)[1] from the HuggingFace library (Wolf et al., 2020) as our base language model and trained 10 epochs via accumulating gradients for 4 steps. We utilized a learning rate of 6.25E-5, and AdamW (Loshchilov and Hutter, 2019) for optimization. We set the warmup ratio to 4% followed by a linear decay. The max length of the input and output is 256 and 128 respectively. We set the batch size to 8. For adjust weihgt, we set the smoothing factor λ to 0.01 and the threshold value θ to 0.5. As for decoding, we use nucleus sampling with p=0.6. We trained our model on a single NVIDIA Tesla V100 GPU with 32GB memory. Each epoch takes about 130 min for WoW and 35min for FaithDial.

4.1 Datasets

We conduct experiments on two knowledge-grounded dialogue datasets: (1) Wizard of WikiPedia (**WoW**) published in (Dinan et al., 2019); (2) FaithDial published in (Dziri et al., 2022a)

WoW is a widely used dataset for knowledge-grounded dialogue based on WikiPedia. WoW is collected by two crowdsourcing workers, one of which is a knowledgeable wizard and the other is an inquisitive apprentice. The wizard can access the knowledge of WikiPedia, while the apprentice cannot. The dataset includes 22,311 dialogues with 201,999 turns, and the test set has two subsets: Test Seen and Test Unseen. Test Seen comprises 533 topics that overlap with the training set and contain new dialogues. Test Unseen contains 58 topics that have never been encountered in training or validation.

[1] https://huggingface.co/t5-base

FaithDial Since the current knowledge conversation dataset (Dinan et al., 2019) contains a large number of hallucination responses (Dziri et al., 2022b), Dziri et al. (2022a) proposes FaithDial, which corrects the responses in WoW to be more faithful to knowledge. The percentage of corrections to the original Wizard's responses exceeded 80%. The dataset contains a total of 5,649 conversations with 50,761 turns.

4.2 Baselines

We compare our model with the following baselines:

GPT2 (Radford et al., 2019) is an autoregressive model based on the Transformer decoder architecture (Vaswani et al., 2017).

DIALOGPT (Zhang et al., 2020) is pre-trained on a large scale dialogue datasets based on GPT2 to be more applicable to conversation generation.

DOHA (Prabhumoye et al., 2021) equips the BART (Lewis et al., 2020) model with a knowledge-aware attention module, enabling specific attention to the information in the knowledge.

CTRL (Rashkin et al., 2021) utilizes control codes to guide the model to generate responses that are more faithful to knowledge. Following (Dziri et al., 2022a), we use T5 as the base model of CTRL.

4.3 Evaluation Metrics

We aim to verify the effectiveness of our method in two aspects: **fluency** and **faithfulness**. We use both automatic metrics and human evaluations to compare baseline models.

Automatic Metrics We use **BLEU** (Papineni et al., 2002), **ROUGE** (Lin, 2004) to evaluate the fluency of the generated responses, which reflect the similarity of the generated responses to the reference responses and both are widely used in text generation evaluation (Dziri et al., 2022a; Zhou et al., 2022). To evaluate the faithfulness of the generated responses to knowledge, we use **BERTScore** (Zhang et al., 2019), **F1** and **Q^2** (Honovich et al., 2021). BERTScore can measure the semantic similarity of responses to knowledge with sentence embeddings from BERT (Devlin et al., 2019), while F1 measures the lexical overlap between responses and knowledge, and Q^2 uses an automated question-and-answer technique to evaluate the consistency of responses and knowledge.

Human Evaluation To mitigate the unreliability of automatic evaluation. We further conduct a human evaluation to verify the effectiveness of our method. We randomly select 100 dialogues from the test set of FaithDial and ask three human evaluators to evaluate. We asked the human evaluators to rate the fluency (**Fluency**), informativeness (**Inform.**) and faithfulness (**Faithful.**) of the generated responses on a 5-point scale, where 1, 3, and 5 indicate unacceptable, moderate, and perfect performance, respectively. Among the metrics, **Fluency** evaluate the response generation quality, **Inform.** evaluate the whether the response is safe or vacuous, and **Faithful.** focus on whether the knowledge used in response is come from the given knowledge, which is stricter than **Inform.**. We then calculate the average score of the three human evaluators as the final score.

5 Results

The results on FaithDial and WoW are shown in Table 1, 2 and 3. As can be seen, FocusL outperforms all baselines in both faithfulness and fluency.

5.1 Automatic Evaluation

FCE vs CE To test the effectiveness of FocusL equipped with FCE, We compare our method with baselines on FaithDial dataset, and report the results in Table 1. We use the test results of baselines from (Dziri et al., 2022a) and keep the same metric calculate method to evaluate FocusL. We can see that our method outperforms the state-of-the-art baselines on all automatic metrics. In particular, FocusL achieves a significant improvement in BERTScore, F1, BLEU and ROUGE, and a large improvement in Q^2 F1 and Q^2 NLI. We also found that the models based on transformer decoder architecture (GPT2, DIALOGPT) perform worse than the encoder-decoder architecture (T5, CTRL, DOHA). Noticably, despite the fact that CTRL performs well in terms of faithfulness, it doesn't improve fluency much. In contrast, FocusL achieves a significant improvement in both faithfulness and fluency. This indicates that FocusL reasonably utilizes knowledge during conversation.

Moreover, to demonstrate the learning focus of our approach, we analyze the trend of loss during training as shown in Figure 4. FocusL achieves higher performance with nearly the same trend as

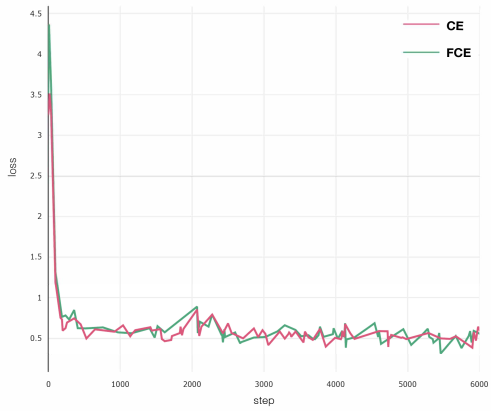


Figure 4: The loss during training. FCE can maintain the same training stability as CE.

the original CE loss variation and also shows that our FCE loss does not destabilize training. Compared to the original CE loss, FCE has a higher loss at the beginning of training, which the more significant adjustment of our approach to the learning focus at the beginning of training can explain.

Robustness to Out-of-Domain Knowledge To evaluate the ability to apply knowledge of out-of-domain, We further test our method on WoW, and report the results in Table 2. We select the baselines which perform well on FaithDial and use T5 as the backbone for comparison. Results show that FocusL outperforms all baselines on faithfulness metrics, significantly improved model's reliability with slightly impact on fluency. It is worth noting that our model improves more significantly in the out-of-domain setting, which indicates that our method is more robust to out-of-domain knowledge.

Robustness to Data Size In order to verify the learning efficiency of our approach with adjusted learning focus, we also conducted experiments in a low-resource setting. We randomly select 1/2, 1/4, 1/8, 1/16, and 1/32 of the training data and report the results in Figure 5. We can see that our method has higher faithfulness even with 1/32 training data. Results also show that FocusL has more significant improvement compared with baselines in the low-resource setting, which demonstrates that our approach can learn how to use knowledge more efficiently. Meanwhile, the faithfulness of both T5 and CTRL does not change significantly. However, their fluency decrease severely, which might be explained by that models tend to copy knowledge

Models	Faithfulness				Fluency	
	BERTScore	F1	Q^2 F1	Q^2 NLI	BLEU	ROUGE
GPT2	0.36	50.41	58.4	69.8	9.50	33.43
DIALOGPT	0.36	52.25	56.5	66.2	9.63	33.13
DOHA	0.39	58.32	69.1	78.3	9.89	31.78
T5	0.41	59.22	70.4	79.5	10.31	33.89
CTRL	0.46	62.21	72.4	81.5	10.41	33.97
FocusL	**0.50****	**65.07***	**73.25**	**82.58**	**11.58****	**35.41****

Table 1: Automatic results on FaithDial to evaluate the Faithfulness and Fluency of the generated responses. The the best performance are **bolded**. One "*" denotes statistical significant with $p < 0.05$, and "**" denotes significant improvement with $p < 0.01$.

Test Set Split	Models	Faithfulness				Fluency	
		BERTScore	F1	Q^2 F1	Q^2 NLI	BLEU	ROUGE
seen topic	T5	0.48	61.88	69.08	75.02	12.44	32.79
	CTRL	0.49	62.99	70.56	76.35	12.61	**33.20**
	FocusL	**0.52**	**65.25**	**71.41**	**77.32**	**12.63**	32.95
unseen topic	T5	0.47	60.68	67.13	73.09	**12.63**	32.81
	CTRL	0.46	59.81	66.70	72.59	12.30	32.73
	FocusL	**0.51**	**63.99**	**69.09**	**74.97**	12.48	**32.84**

Table 2: Automatic results on WoW to evaluate the Faithfulness and Fluency of the generated responses. The the best performance are **bolded**.

Models	Faithful.	Fluency	Inform.
T5	2.80	**3.62**	3.23
CTRL	2.98	3.53	3.14
FocusL	**3.11***	3.59	**3.44***

Table 3: Human evaluation on WoW. **Bolded** numbers indicate the best performance. Numbers marked with * indicate that the improvement is statistically significant (p-value < 0.05).

Model	BERTScore	F1	BLEU
FocusL	**0.51**	**66.11**	11.65
-CW	0.38	52.63	9.10
-LW	0.42	57.86	11.62
w/o FCE	0.40	57.17	**11.89**

Table 4: The ablation study of various adjust weight distribution. **Bolded** numbers indicate the best performance.

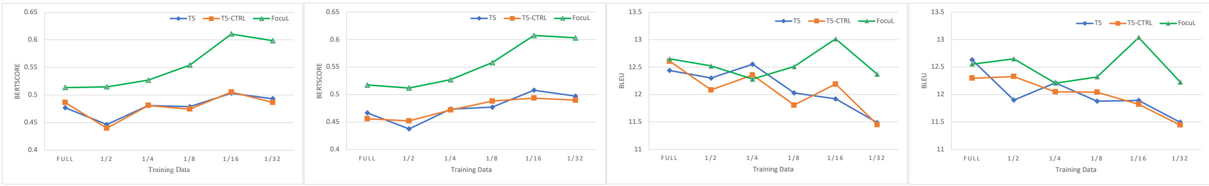
and ignore the fluency of the response. In comparison, FocusL can achieve a better trade-off between fluency and faithfulness with limited data.

5.2 Human Evaluation

In addition to automatic evaluation, we present human evaluation results in Table 3. We choose T5 and CTRL as baselines for comparison. Results show that FocusL receives higher scores on both **Faithful.** and **Inform.**, and fluency is slightly lower than T5. Overall, our approach can make model more reliable with almost as much fluency as the baselines.

5.3 Ablation Study

Finally, we attempt to study the performances of variation of FCE described in §3.3. Results for different adjust weight distribution are shown in Table 4, and Table 5 is for different λ in non-linear weight. In Table 4, we compared the weight distribution with the threshold (CW), linear weight (LW), non-linear weight (FocusL), and without FCE (w/o FCE). Among them, CW performs the worst, which may be influenced by the threshold. In contrast, the effect of LW is more stable than CW and does not suffer from hyperparameter effects. Even though BLEU is slightly lower than CE, FocusL has sig-



(a) BERTScore (b) BERTScore (c) BLEU (d) BLEU

Figure 5: Automatic results on WoW with limited training data. (a) and (b) show the results of BERTScore on seen and unseen test set, respectively. (c) and (d) show the results of BLEU on seen and unseen test set, respectively.

Model	BERTScore	F1	BLEU
FocusL	**0.51**	**66.11**	11.65
$\lambda = 0.05$	0.45	61.51	11.78
$\lambda = 0.1$	0.50	64.55	11.53
$\lambda = 0.2$	0.44	60.88	12.08
$\lambda = 0.4$	0.43	60.19	12.13
$\lambda = 0.7$	0.43	60.81	**12.53**

Table 5: The ablation study of various λ for non-linear adjust weight distribution. **Bolded** numbers indicate the best performance.

nificantly improved BERTScore and F1.

To further study the effect of λ in non-linear weight, we set λ to 0.05, 0.1, 0.2, 0.4, 0.7, and present the results in Table 5. Note that FocusL used $\lambda = 0.01$ in the experiments. As the λ increases, faithfulness metrics of the model gradually decrease, and fluency metrics gradually increase. This indicates that smaller λ with sharper weight distribution makes the model more sensitive to knowledge-aware tokens' losses, which increases the accuracy of knowledge utilization. In contrast, larger λ with smoother weight distribution makes the model focus on the quality of the response.

5.4 Case Study

To better illustrate the advantage of our approach, we present an example case in Table 6. We randomly chose one dialog from the test set of FaithDial, and compare the responses generated by T5, CTRL, and FocusL. It can be observed that the response generated by T5 used a wrong year "*1998*" while the given knowledge is about "*1997*", and its causality also cannot be inferred from the given knowledge. CTRL misunderstands the given knowledge and ignores the impact of "*Bring It On*" on the global presentation of cheerleading. In contrast, FocusL can generate a response more related to the given knowledge and closest to the gold response, which contains all the knowledge entities in the gold response.

Given Knowledge:
The global presentation of cheerleading was led by the 1997 broadcast of ESPN's International cheerleading competition, and the worldwide release of the 2000 film " Bring It On ".

Context:
<user> She has done a lot of dance and tumbling already. She will try it out and see what works best for her.
<bot> Got it, are you from the United States? Cheerleading is an activity that originated there, it is also predominantly in America.
<user> Yes we are, she wants to be a cheerleader since she was a little kid, I am sure she will be fine. I could see her going on to do it in college as well.

Gold Response: Nice, have you watched the film Bring It on ? it is from 2000 .

T5: I see, did you know that the 1998 televised ESPN's International cheerleading competition led to the global presentation of cheerleading? That's interesting.

CTRL: Yes, the world presented cheerleading in 1997 .

FocusL: I see, did you know that the movie Bring It On was released in 2000 ?

Table 6: An example case from FaithDial. We highlight the knowledge-aware tokens in the gold response with green and the hallucination with red.

6 Conclusion

In this paper, we propose a novel learning approach with more direct guidance on the training process to improve the faithfulness of knowledge-grounded dialogue systems, referred to as FocusL. By leveraging semantic relevance between the response and knowledge, FocusL correct the model's learning focus, leading to more consistent and fluent response generation. We empirically show that our approach has the best performance with a stable training process and is robust to data size and out-of-domain knowledge. FocusL is simple yet effective and can achieve state-of-the-art results in two knowledge-grounded datasets.

Limitations

As we have shown, there is much room to improve the learning approach, which incur lower costs than increasing model's parameters or elaborate data engineering. This paper is an exercise in guiding learning focus, and we argue that FocusL is not perfect for the positioning method and the relevance-to-weight transformation method. For example, our positioning method will contain noise, and some words that are not important in given knowledge may be used as our learning focus. We will continue to explore better methods to guide the model's learning focus. Meanwhile, our method only experiments on the basic cross-entropy loss, and still needs to be explored for other learning approaches such as contrast learning.

Ethics Statement

FocusL aims to to convey correct knowledge to users rather than misleading hallucinations. We hope to see a reliable and trustworthy dialogue system impact from better guiding the model's learning focus. However, even if the dialogue system does not produce hallucinations, there is still a risk of potential misuse. For example, the dialogue systems may be used to spread misinformation or to mislead users. If possible, we would prefer that the model itself has the ability to identify undesirable knowledge and block it.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Yuanyuan Cai, Min Zuo, Qingchuan Zhang, Haitao Xiong, and Ke Li. 2020. A bichannel transformer with context encoding for document-driven conversation generation in social media. *Complex.*, 2020:3710104:1–3710104:13.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *arXiv preprint, arXiv:2204.10757*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, William B. Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *ArXiv*, abs/1702.01932.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2022. Knowledge-consistent dialogue generation with knowledge graphs. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2823, Seattle, United States. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Longxuan Ma, Wei-Nan Zhang, Runxin Sun, and Ting Liu. 2020. A compare aggregate transformer for understanding document-grounded dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1358–1367, Online. Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and M. de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Nikita Moghe, Siddharth Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Conference on Empirical Methods in Natural Language Processing*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, M. de Rijke, and Zhaochun Ren. 2022. Contrastive learning reduces hallucination in conversations. *ArXiv*, abs/2212.10400.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1845, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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*, pages 270–278, Online. Association for Computational Linguistics.

Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. 2019. A document-grounded matching network for response selection in retrieval-based chatbots. In *International Joint Conference on Artificial Intelligence*.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations*.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.