BoK: Introducing Bag-of-Keywords Loss for Interpretable Dialogue Response Generation

Suvodip Dey and **Maunendra Sankar Desarkar**
Indian Institute of Technology Hyderabad, India
suvodip15@gmail.com, maunendra@cse.iith.ac.in

Abstract

The standard language modeling (LM) loss by itself has been shown to be inadequate for effective dialogue modeling. As a result, various training approaches, such as auxiliary loss functions and leveraging human feedback, are being adopted to enrich open-domain dialogue systems. One such auxiliary loss function is Bag-of-Words (BoW) loss, defined as the cross-entropy loss for predicting all the words/tokens of the next utterance. In this work, we propose a novel auxiliary loss named Bag-of-Keywords (BoK) loss to capture the central thought of the response through keyword prediction and leverage it to enhance the generation of meaningful and interpretable responses in open-domain dialogue systems. BoK loss upgrades the BoW loss by predicting only the keywords or critical words/tokens of the next utterance, intending to estimate the core idea rather than the entire response. We incorporate BoK loss in both encoder-decoder (T5) and decoder-only (DialoGPT) architecture and train the models to minimize the weighted sum of BoK and LM (BoK-LM) loss. We perform our experiments on two popular open-domain dialogue datasets, DailyDialog and Persona-Chat. We show that the inclusion of BoK loss improves the dialogue generation of backbone models while also enabling post-hoc interpretability. We also study the effectiveness of BoK-LM loss as a reference-free metric and observe comparable performance to the state-of-the-art metrics on various dialogue evaluation datasets.

1 Introduction

Open-domain dialogue generation is a dynamic area of research, aiming to generate contextually relevant and meaningful responses given a dialogue context. As deep learning models continue to thrive in the field of natural language processing (NLP), a widely adopted strategy to solve any natural language generation (NLG) task involves pre-training and fine-tuning large language models (LLMs).

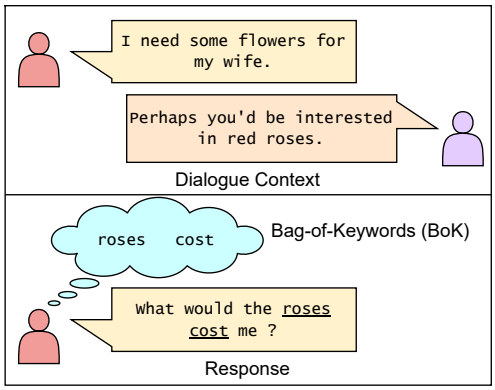


Figure 1: A motivating example for Bag-of-Keywords loss in open-domain dialogue system.

The LLMs are predominantly trained with language modeling (LM) loss, which essentially corresponds to cross-entropy loss for predicting the next word or token. While LM loss remains effective in training NLG models for diverse tasks, including dialogue generation (Sordoni et al., 2015; Wolf et al., 2019; Zhang et al., 2020; Roller et al., 2021), it may not be the optimal choice for training models specifically tailored for dialogue generation. It is well-established that perplexity, a measure associated with LM loss, primarily gauges fluency and weakly correlates with human dialogue evaluation (Dinan et al., 2019; Mehri and Eskenazi, 2020b; Phy et al., 2020). Consequently, relying solely on LM loss may not guarantee generations with desirable conversational qualities. Therefore, exploring alternative loss functions and training methods is crucial to advance the development of generative open-domain dialogue models.

In order to mitigate the exclusive dependence on LM loss in the training of open-domain dialogue models, various approaches have been explored in the existing literature. These techniques can be broadly categorized into two classes - a) auxiliary loss and b) human feedback. The first approach combines one or more auxiliary losses

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 566–578
September 18–20, 2024.

with LM loss to train the dialogue models. Various types of auxiliary losses have been explored in the context of open-domain dialogue learning. For instance, Bag-of-Words (BoW) loss computes the cross-entropy loss to predict words/tokens of the next utterance from the given dialogue context (Zhao et al., 2017; Li et al., 2021; Dey et al., 2023). Some methodologies involve predicting the sentence-level encoding of the next utterance and determining the loss through L1/L2 norms and KL divergence (Serban et al., 2017; Li et al., 2021; Chen et al., 2022; Dey et al., 2023). Few approaches incorporate a next-utterance classification loss (Wolf et al., 2019), wherein the auxiliary loss is computed for a classification or ranking task to predict the true utterance from a set of candidate responses. On the other hand, the second approach is based on refining the pre-trained dialogue model through human feedback. These methods mostly follow the training principle of Reinforcement learning from human feedback (RLHF), where the model is fine-tuned to maximize the reward associated with the generated response using Reinforcement learning. RLHF has gained significant interest recently, particularly with the popularity of models like Chat-GPT (Long and et al., 2022). However, acquiring quality human feedback data is challenging and expensive (Casper and et al., 2023). Furthermore, relying on automated dialogue evaluation metrics as a substitute for human feedback can pose challenges, as they may not strongly correlate with human judgments (Liu et al., 2016; Yeh et al., 2021).

In this work, our objective is to propose a novel auxiliary loss for open-domain dialogue systems. Specifically, we address the limitation of BoW loss by introducing Bag-of-Keywords (BoK) loss, which is defined as the cross-entropy loss to predict the keywords of the next utterance. While training, we extract the keywords of the ground-truth response using YAKE! (Campos et al., 2018, 2020), an unsupervised feature-based keyword extractor. The keywords can be seen as a proxy for the core idea of the response. In a conversation, a reply can be generated in multiple ways. As a result, BoW loss can induce training data bias since it considers all the words/tokens of the ground-truth response for prediction. In contrast, BoK loss focuses on the core idea (as shown in Fig. 1) that alleviates the problem of generalization. The main contributions of this work are summarized as follows[1]:

- We propose BoK loss, a novel auxiliary loss for open-domain dialogue systems. BoK loss can be easily incorporated into any generative model and trained using a weighted sum of BoK and LM (BoK-LM) loss.
- We show that BoK loss enhances the dialogue generation of backbone models on DailyDialog and Persona-Chat datasets. We note an improvement in the specificity of the generated responses with the inclusion of BoK loss.
- We perform a qualitative analysis of the generated responses and discuss how BoK loss enables post-hoc interpretability.
- We study the effectiveness of BoK-LM loss as a reference-free metric. We observe that it exhibits moderate correlations with human judgments on different evaluation datasets.

2 Background and Related Works

Open-domain dialogue generation is a challenging NLG task. Let $D_{<t} = \{u_1, u_2, ...u_{t-1}\}$ be a multi-turn conversation where u_j represents the utterance at turn j. Let C_t be the condition (like persona, document, etc.) other than dialogue history for generating u_t. The task of open-domain dialogue generation is to generate u_t given $D_{<t}$ and C_t. Like any NLG task, it is modeled using language models and generally trained using the next word/token prediction task. The corresponding language modeling (LM) loss is defined as,

$$\mathcal{L}_{\text{LM}} = -\sum_{n=1}^{T} \log p(u_{t_n} | u_{t_{<n}}, D_{<t}, C_t; \theta) \quad (1)$$

where u_{t_n} denotes the n^{th} word/token of utterance u_t and θ indicates the parameters of the language model. Training transformer (Vaswani et al., 2017) based large language models (LLMs) with LM loss on large dialogue corpora has shown remarkable performance in open-domain dialogue generation (Zhang et al., 2020; Roller et al., 2021). However, it has been shown that perplexity ($e^{\mathcal{L}_{\text{LM}}}$), a metric that is a function of LM loss, can measure fluency but shows a weak correlation with other conversational aspects (Dinan et al., 2019; Mehri and Eskenazi, 2020b; Phy et al., 2020). The root

[1]Code is available at github.com/SuvodipDey/BoK

cause of this behavior stems from the inherent one-to-many nature of dialogue, where a given context can elicit multiple possible responses (Liu et al., 2016). Consequently, simply increasing the size of training data may not always yield improvement, as it is impractical to collect all potential response variations. To tackle this challenge, researchers employ various techniques, broadly categorized into two classes: i) incorporating one or more auxiliary losses alongside LM loss, and ii) leveraging human feedback to finetune pre-trained dialogue models. Given our focus on proposing a new auxiliary loss, we keep our related works limited to different auxiliary losses utilized for open-domain dialogue generation, described as follows.

- The first kind of auxiliary loss estimates the error in predicting the sentence-level encoding of the next utterance given the dialogue context. Authors of VHRED (Serban et al., 2017) and DialogVED (Chen et al., 2022) use Kullback-Leibler (KL) divergence to compute the distance between the approximate and true posterior distribution of the next utterance. Models like DialoFlow (Li et al., 2021) and DialoGen (Dey et al., 2023) use the L1/L2 norm for the same purpose. Predicting the encoding of the next utterance is challenging and may lead to issues like posterior collapse while using KL divergence (Chen et al., 2022).

- The second approach is based on the next utterance classification loss. In this method, the task is to classify the ground-truth response from a given set of candidate utterances (Wolf et al., 2019). It is worth noting that this method requires negative samples, which are usually not included in the datasets. Hence, different kinds of negative sampling techniques are adopted to obtain them. However, finding high-quality negative samples is difficult for dialogues (Lan et al., 2020).

- The third approach focuses on predicting the words/tokens of the next utterance. This loss is popularly known as Bag-of-Words (BoW) loss (Zhao et al., 2017). Models like DialoFlow (Li et al., 2021) and DialoGen (Dey et al., 2023) utilize BoW loss to support LM loss. DialogVED (Chen et al., 2022) uses BoW loss to tackle the posterior collapse that is caused due to minimizing KL divergence. As discussed earlier, a dialogue context can have many relevant responses. Hence, the task of predicting all the words/tokens can induce training data bias. In this work, we aim to address this limitation of BoW loss.

3 Methodology

In this section, we describe Bag-of-Keywords loss followed by its application in open-domain dialogue systems.

3.1 Bag-of-Keywords (BoK) loss

As discussed, BoW loss is computed as the cross-entropy loss to predict all the tokens of the next utterance. Say the model has to generate utterance u_t given dialogue context $D_{<t}$. Let $\phi_t \in \mathbb{R}^d$ be the representation of the context for generating u_t. Then the BoW loss ($\mathcal{L}_{\text{BoW}}$) is defined as,

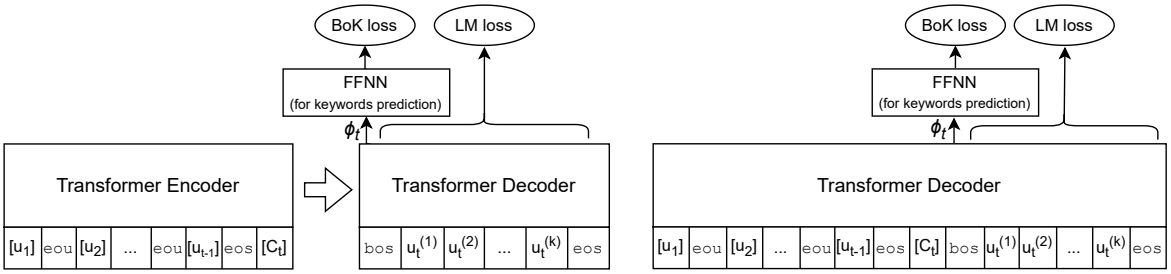
$$\mathcal{L}_{\text{BoW}} = -\sum_{w \in u_t} \log p(w|\phi_t) \tag{2}$$

where $p(w|\phi_t)$ is the probability of predicting the word/token $w \in u_t$ given ϕ_t. Predicting all the words of a dialogue response may cause training data bias because there can be multiple ways to generate a response. Additionally, dialogue responses often contain stopwords that are necessary for sentence construction and fluency. Therefore, predicting these stopwords in BoW loss is unnecessary since LM loss already takes care of it.

One simple approach to address this limitation of BoW loss is to predict only the keywords of the response. By keywords, we mean the critical words that capture the core concept of the response. This approach can help reduce the training data bias and increase its generalizability for open-domain dialogue generation. To achieve this, we propose Bag-of-Keywords (BoK) loss, which is computed as the cross-entropy loss to predict the keywords of the next utterance. We define BoK loss ($\mathcal{L}_{\text{BoK}}$) as,

$$\mathcal{L}_{\text{BoK}} = -\sum_{w \in K_t} \log p(w|\phi_t) \tag{3}$$

where K_t is the set of keywords (or tokens associated with the keywords) in u_t. Note that the annotations regarding the keywords are not available in the existing dialogue datasets. In this work, we find the keywords using YAKE! (Campos et al., 2018, 2020), an unsupervised feature-based keyword extraction algorithm that leverages statistical features extracted directly from the text, thereby supporting



(a) Encoder-decoder framework

(b) Decoder-only framework

Figure 2: Incorporating BoK loss in open-domain dialogue models. $[u_j]$ and $[C_j]$ represents the list of tokens after tokenizing utterance u_j and condition C_j, respectively. $u_t^{(i)}$ denotes the i^{th} token of utterance u_t, whereas {eos, bos, eou} are special tokens. $\phi_t \in \mathbb{R}^d$ is the hidden state of the final layer of bos token, representing the context.

texts of multiple domains and languages. However, one can adopt any strategy for keyword extraction. We chose YAKE! because it is unsupervised and has already been utilized to extract keywords from dialogue responses (Dey and Desarkar, 2023). For example, in Fig. 1, YAKE! extracted the keywords "*roses*" and "*cost*" from the response "*What would the roses cost me?*".

3.2 Application of BoK loss

BoK loss can be easily applied to any open-domain dialogue model. Currently, all state-of-the-art dialogue generation models are based on Transformer (Vaswani et al., 2017). These models can be broadly classified into two architectures - i) encoder-decoder and ii) decoder-only. Incorporating BoK loss into both these architectures is described as follows.

- **Encoder-Decoder Architecture**: Fig. 2a shows the method of applying BoK loss in encoder-decoder architecture. The encoder takes the concatenation of the past utterances ($D_{<t}$) along with the condition C_t as input. Note that C_t may be present or absent based on the task or dataset. In the decoder, we add an extra component for computing the BoK loss. Let $\phi_t \in \mathbb{R}^d$ be the hidden state representation of the final layer corresponding to the bos token, representing the context. Then, the BoK loss is computed as follows:

$$\alpha_t = \text{softmax}(\text{FFNN}(\phi_t)) \in \mathbb{R}^{|V|} \quad (4)$$

$$\mathcal{L}_{\text{BoK}} = -\sum_{w \in K_t} \log p(w|\phi_t) = -\sum_{w \in K_t} \log \alpha_{t_w} \quad (5)$$

where FFNN denotes a single layer feed-forward neural network, and $|V|$ is the vocabulary size of the decoder tokens.

Dataset	Type	#Dialog	#Turns	T_{max}	T_{min}	T_{avg}
DailyDialog	Train	11118	87170	35	2	7.84
	Dev	1000	8069	31	2	8.07
	Test	1000	7740	26	2	7.74
Persona-Chat	Train	8939	131438	50	12	14.70
	Dev	1000	15602	26	14	15.60
	Test	968	15024	34	14	15.52

Table 1: Basic statistics of DailyDialog and Persona-Chat dataset. T_{max}, T_{min}, and T_{avg} indicate maximum, minimum, and average dialogue turns.

- **Decoder-only Architecture**: Fig. 2b shows the process of incorporating BoK loss in decoder-only architecture. The BoK loss computation follows the same equations (Eqn. 4 and 5) as encoder-decoder architecture.

The training objective for both architectures is to minimize the weighted sum of BoK and LM loss. We term this loss as BoK-LM loss ($\mathcal{L}_{\text{BoK-LM}}$).

$$\mathcal{L}_{\text{BoK-LM}} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{BoK}} \quad (6)$$

where $\lambda \in \mathbb{R}$ is a hyper-parameter to set the weight of the BoK loss. Note that both the loss components depend on the context vector ϕ_t. Hence, the BoK-LM loss helps to learn ϕ_t such that it can capture the core idea of the response and align the generation towards a meaningful response.

4 Experimental Set up

4.1 Datasets

We conduct our experiments on two datasets: DailyDialog (Li et al., 2017) and Persona-Chat (Zhang et al., 2018a). DailyDialog is a popular chit-chat dataset in which the task is to generate responses conditioned only on the dialogue history. On the other hand, Persona-Chat is a knowledge-grounded dataset where a response needs to be generated

Model	Referenced Metrics										Reference-Free Metric			
	Bleu-1	**Bleu-2**	**Bleu-3**	**Bleu-4**	**Nist-2**	**Nist-4**	**Meteor**	**Div-1**	**Div-2**	**Entropy**	**U**	**S**	**L_S**	**USL_S-H**
DialoFlow	48.75	26.73	16.35	10.70	3.76	3.97	16.44	0.039	0.216	9.98	0.96	0.88	0.21	0.6777
DialoGen	49.13	27.25	16.88	11.07	3.76	3.98	16.40	0.043	0.223	9.88	0.83	0.90	0.32	0.6685
DialogVED	50.50	28.95	18.38	12.29	3.94	4.18	16.90	0.037	0.204	9.82	0.86	0.88	0.30	0.6642
T5	51.56	29.22	18.29	12.05	3.99	4.23	16.27	0.044	0.219	9.62	0.97	0.89	0.18	0.6718
$T5_{BoW}$	**51.75**	29.70	18.89	12.75	4.05	4.32	16.64	0.045	0.230	9.79	0.97	0.89	0.19	0.6791
$T5_{BoK}$	51.74	**29.74**	19.19	13.24	**4.09**	**4.37**	16.62	0.045	0.233	9.84	0.97	0.90	0.20	0.6793
DialoGPT	49.30	27.63	17.37	11.68	3.78	4.01	16.67	0.037	0.193	9.66	0.97	0.89	0.19	0.6731
$DialoGPT_{BoW}$	49.60	27.85	17.60	11.82	3.80	4.04	16.83	0.037	0.190	9.60	0.97	0.89	0.20	0.6759
$DialoGPT_{BoK}$	49.16	29.10	**20.00**	**14.92**	4.01	4.35	**17.72**	**0.048**	**0.257**	**10.19**	0.97	0.89	0.31	**0.7064**

Table 2: Comparison of dialogue generation performance on DailyDialog test data with automated metrics. The highest and second-highest scores are written in bold and underlined respectively.

based on both dialogue history and a persona profile that defines the speaker. Table 1 displays the basic statistics of the two datasets.

4.2 Implementation Details

We choose T5 (Raffel et al., 2020) and DialoGPT (Zhang et al., 2020) as our encoder-decoder and decoder-only architecture, respectively. We perform our experiments with T5-large[1] (≈770M parameters) and DialoGPT-large [2] (≈774M parameters) for both DailyDialog and Persona-Chat datasets. All the implementations are done using PyTorch and Huggingface (Wolf et al., 2020) libraries in Python 3.10, and executed on a Nvidia A100 with 40GB memory. We use AdamW optimizer with a learning rate of 5e-5, batch size of 16, maximum training epochs of 20, and early stopping to train the models. We use beam search with a beam width of 5, maximum sequence length of 40, minimum sequence length of 11, and length penalty of 0.1 to generate responses for all the models. The rest of the details are provided in Appendix A.1.

4.3 Baselines

We refer to T5 and DialoGPT trained with BoK-LM loss as $T5_{BoK}$ and $DialoGPT_{BoK}$, respectively. We compare them with vanilla T5 and DialoGPT models, trained only with LM loss. To measure the improvement over BoW loss, we also train T5 and DialoGPT with a weighted sum of BoW and LM loss (like Eqn. 6), denoted as $T5_{BoW}$ and $DialoGPT_{BoW}$ respectively. We also have some dataset-specific baselines. For DailyDialog, we use DialoFlow (Li et al., 2021), DialogVED (Chen et al., 2022), and DialoGen (Dey et al., 2023). All these three baselines use BoW loss and sentence-level next utterance prediction loss. For Persona-Chat, we use TransferTransfo (Wolf et al., 2019) and DialogVED. TransferTransfo utilizes the next utterance classification as the auxiliary loss.

[1] huggingface.co/google-t5/t5-large

[2] huggingface.co/microsoft/DialoGPT-large

Model	**U**	**S**	**L_S**	**USL_S-H**	**Dial-M**
TransferTransfo	0.75	0.63	0.44	0.5502	1.7730
DialogVED	0.74	0.84	0.38	**0.6348**	1.7499
T5	0.71	0.73	0.39	0.5756	0.9288
$T5_{BoW}$	0.72	0.75	0.40	0.5867	0.8781
$T5_{BoK}$	0.72	0.76	0.41	0.5947	**0.8556**
DialoGPT	0.76	0.72	0.36	0.5788	1.0312
$DialoGPT_{BoW}$	0.77	0.71	0.40	0.5868	1.0013
$DialoGPT_{BoK}$	0.77	0.72	0.42	0.5923	1.0004

Table 3: Comparison of dialogue generation performance on Persona-Chat test data.

5 Results and Analysis

5.1 DailyDialog Dataset

Table 2 compares the performance of various models on DailyDialog test data. We use BLEU (Papineni et al., 2002), NIST (Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), Diversity (Li et al., 2016), and Entropy (Zhang et al., 2018b) for referenced evaluation, and USL-H (Phy et al., 2020) for reference-free evaluation. As word-overlapping based metrics are not reliable with only one reference, we conduct the referenced evaluation using multi-reference DailyDialog (Gupta et al., 2019) that contains four additional references along with the original response. For BoK loss, we set the maximum number of keyword tokens $|K_t| = 8$ (refer Eqn. 3). For BoK-LM loss in Eqn. 6, we set λ to 0.1 and 0.3 for T5 and DialoGPT architecture, respectively. The effect of varying λ and $|K_t|$ is studied in the ablation study. The key observations from Table 2 are discussed below.

Referenced Evaluation: Firstly, we observe that the inclusion of BoW loss enhances the performance of both vanilla T5 and DialoGPT across all metrics. BoW loss is optimized to predict all the words/tokens of the next utterance, thereby improving the unigram match i.e. Bleu-1 score. Our findings corroborate this observation, demonstrating that $T5_{BoW}$ and $DialoGPT_{BoW}$ attain higher Bleu-1 scores compared to their other counterparts. Sec-

Comparisons	Dataset	Coherence			Engagingness			Informativeness			Interactiveness			Overall		
		W	L	T	W	L	T	W	L	T	W	L	T	W	L	T
$T5_{BoK}$ vs. $T5_{BoW}$	DailyDialog	24	18	58	30	26	44	20	14	66	26	18	56	32	26	42
	Persona-Chat	26	18	56	24	20	56	24	18	58	20	18	62	28	24	48
$DialoGPT_{BoK}$ vs. $DialoGPT_{BoW}$	DailyDialog	42	34	24	30	30	40	44	26	30	34	30	36	46	34	20
	Persona-Chat	28	18	54	14	20	66	24	18	58	14	16	70	28	22	50

Table 4: Human evaluation for comparing the impact of BoK and BoW loss on the performance of the backbone models. "W", "L", and "T" denote the percentage of win, loss, and tie, respectively.

ondly, we note that both $T5_{BoK}$ and $DialoGPT_{BoK}$ perform better than their BoW counterpart in most of the cases. Furthermore, they also outperform the three baselines (DialoFlow, DialoGen, and DialogVED) that rely on BoW loss. This indicates that BoK loss effectively improves the generalizability of BoW loss, making it more efficient.

Reference-free Evaluation: We use USL-H as our reference-free metric, which is a combination of three sub-metrics - Understandability (U), Sensibility (S), and Likability (L). We specifically make use of the USL_S-H variant, where the likability of a response is captured through its specificity. USL_S-H estimates understandability, sensibility, and specificity using valid prediction, next utterance prediction, and MLM task, respectively (Phy et al., 2020). Similar to the results of the referenced evaluation, $T5_{BoK}$ and $DialoGPT_{BoK}$ achieve better USL_S-H scores than their other counterparts. Moreover, we note that for $T5_{BoK}$ and $DialoGPT_{BoK}$, USL_S-H improves because of the likability or specificity aspect. We also observe this behavior in Table 3, which indicates that incorporating BoK loss enhances the specificity of the generated responses.

5.2 Persona-Chat Dataset

The results of the Persona-Chat test data are presented in Table 3. Unlike DailyDialog, Persona-Chat does not have any multi-referenced test data. Therefore, we use only reference-free metrics to ensure a fair evaluation. In addition to USL_S-H, we also evaluate using Dial-M (Dey and Desarkar, 2023), a masking-based reference-free metric that is effective in evaluating knowledge-grounded dialogues. It is worth mentioning that in Dial-M, a lower score is indicative of better performance as it is based on cross-entropy loss. In Table 3, we again observe that $T5_{BoK}$ and $DialoGPT_{BoK}$ attain better USL_S-H and Dial-M scores than their other counterparts. Furthermore, we observe that DialogVED outperforms all the models on USL_S-H. This is because it does not use persona profiles explicitly and relies on specially trained latent variables (on next utterance prediction) for persona-grounded response generation. Furthermore, USL_S-H only considers dialogue history as context and ignores persona. As a result, DialogVED performs better in understandability and sensibility, which are estimated using valid and next utterance prediction tasks, respectively. However, it falls short in specificity and Dial-M as it does not use persona.

We observe that for both DailyDialog and Persona-Chat, BoK performs better than BoW in most of the cases. For DailyDialog, $DialoGPT_{BoK}$ outperforms $DialoGPT_{BoW}$ significantly, which correlates with the automated result shown in Table 2. For Persona-Chat, as the generation is conditioned mainly on the persona profiles, the responses are very similar for both models, resulting in a lot of ties. We also observe that BoK loss results in better informativeness, which correlates with the improved specificity (in USL_S-H) shown in Table 2 and Table 3.

5.3 Human Evaluation

Table 4 shows the human evaluation to compare the impact of BoK and BoW loss on the backbone models. We randomly picked 50 test instances from both DailyDialog and Persona-Chat datasets. Four human evaluators (graduate students proficient in English) were presented with the generated responses from two models (A and B) who reported their judgment (A wins, B wins, or a tie) on various aspects. We asked the evaluators to evaluate five aspects, described as follows.

- *Coherence*: Captures which model produces more contextually coherent responses.
- *Engagingness*: Identifies which model generates more engaging or interesting responses.
- *Informativeness*: Determines which response contains more knowledge or specific information.
- *Interactiveness*: Captures which model produces more interactive responses that encourage the user to continue the conversation.

Model	λ	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Nist-2	Nist-4	Meteor	Div-1	Div-2	Entropy	U	S	L_S	USL_S-H
$T5_{BoK}$	0.05	51.61	29.43	18.69	12.63	4.03	4.29	16.47	0.044	0.224	9.74	0.97	0.89	0.19	0.6748
	0.10	**51.74**	**29.74**	**19.19**	**13.24**	**4.09**	**4.37**	16.62	0.045	0.233	9.84	0.97	0.90	0.20	0.6793
	0.20	51.53	29.58	19.06	13.07	4.06	4.34	**16.71**	0.046	0.231	9.85	0.97	0.90	0.21	0.6802
	0.30	51.08	28.91	18.44	12.55	4.00	4.26	16.58	0.046	**0.234**	**9.88**	0.97	0.90	0.21	**0.6820**
	0.40	50.45	28.21	17.59	11.64	3.93	4.16	16.04	0.046	0.233	9.82	0.97	0.89	0.21	0.6787
	0.50	50.59	28.16	17.54	11.55	3.92	4.15	16.02	0.046	0.233	9.82	0.97	0.89	0.21	0.6779
	0.60	50.33	27.93	17.30	11.28	3.89	4.12	15.88	**0.047**	0.234	9.81	0.97	0.89	0.21	0.6764
$DialoGPT_{BoK}$	0.05	49.59	27.79	17.51	11.72	3.79	4.02	16.84	0.037	0.191	9.61	0.97	0.89	0.20	0.6765
	0.10	**49.62**	27.91	17.68	11.90	3.81	4.05	16.84	0.038	0.195	9.65	0.97	0.89	0.21	0.6788
	0.20	49.36	27.59	17.39	11.64	3.77	4.01	16.75	0.037	0.192	9.64	0.97	0.89	0.20	0.6770
	0.30	49.16	**29.10**	**20.00**	**14.92**	**4.01**	**4.35**	**17.72**	0.048	**0.257**	**10.19**	0.97	0.89	0.31	**0.7064**
	0.40	49.18	28.84	19.50	14.31	3.98	4.29	17.51	0.048	0.254	10.17	0.97	0.89	0.30	0.7048
	0.50	48.83	28.40	19.07	13.92	3.92	4.23	17.11	**0.048**	0.253	10.16	0.97	0.89	0.30	0.7048
	0.60	48.72	28.21	18.82	13.60	3.89	4.19	17.14	0.048	0.252	10.15	0.97	0.89	0.29	0.7032

Table 5: Effect of varying λ on DailyDialog test performance with $|K_t| = 8$.

Model	$\|K_t\|$	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Nist-2	Nist-4	Meteor	Div-1	Div-2	Entropy	U	S	L_S	USL_S-H
$T5_{BoK}$	4	**51.87**	29.69	19.08	13.06	4.07	4.35	16.58	0.046	0.232	9.83	0.97	0.89	0.20	0.6772
	8	51.74	**29.74**	**19.19**	**13.24**	**4.09**	**4.37**	16.62	0.045	0.233	9.84	0.97	0.90	0.20	**0.6793**
	16	51.59	29.57	19.00	13.06	4.06	4.33	16.60	**0.046**	0.233	9.83	0.97	0.89	0.20	0.6780
	24	51.66	29.58	18.96	12.96	4.06	4.33	**16.63**	0.046	**0.234**	**9.85**	0.97	0.89	0.20	0.6781
$DialoGPT_{BoK}$	4	49.08	29.02	19.88	14.81	4.00	4.33	17.69	0.048	0.255	10.18	0.97	0.89	0.31	0.7051
	8	49.16	**29.10**	**20.00**	**14.92**	4.01	**4.35**	**17.72**	0.048	0.257	10.19	0.97	0.89	0.31	**0.7064**
	16	49.18	29.05	19.98	14.92	**4.01**	4.35	17.62	0.048	**0.258**	**10.19**	0.97	0.89	0.31	0.7054
	24	**49.34**	29.02	19.83	14.74	4.00	4.34	17.67	0.048	0.255	10.17	0.97	0.89	0.30	0.7040

Table 6: Effect of varying maximum number of keyword tokens ($|K_t|$) on DailyDialog test performance.

- *Overall*: This is the overall judgment or impression of the evaluator on the given responses.

The inter-annotator agreement (Fleiss' kappa) for the overall judgment was 0.81. The Fleiss' kappa for Coherence, Engagingness, Informativeness, and Interactiveness were 0.75, 0.64, 0.63, and 0.60, respectively.

5.4 Ablation Study

This section analyzes the impact of varying λ and $|K_t|$ in the BoK-LM loss. We conduct this ablation study on DailyDialog test data to perform both referenced and reference-free evaluations.

Table 5 shows the results of changing λ with $|K_t| = 8$ fixed. A higher value of λ denotes higher weightage to BoK loss in Eqn. 6. For Bleu, Nist, Meteor, Entropy, and USL_S-H, we observe that increasing λ improves the performance up to a certain threshold and then starts declining. In general, $T5_{BoK}$ and $DialoGPT_{BoK}$ perform well with λ values of 0.1 and 0.3, respectively. Div-1 metric measures diversity by counting distinct unigrams. This is why it shows better performance with higher λ values, where the context vector ϕ_t is learned to predict the keywords with more precision.

Table 6 shows the effect of varying the maximum number of keyword tokens ($|K_t|$) in Eqn. 3, keeping λ fixed at 0.1 and 0.3 for $T5_{BoK}$ and $DialoGPT_{BoK}$, respectively. Increasing $|K_t|$ makes BoK loss behave more like BoW loss. As a result, we observe $DialoGPT_{BoK}$ with $|K_t| = 24$ achieves the best Bleu-1 score, which aligns with our expectation. However, this is not the case for $T5_{BoK}$, which we believe happens due to the lower value of λ (0.1). Overall, $T5_{BoK}$ and $DialoGPT_{BoK}$ shows better performance with $|K_t| = 8$.

6 Discussions

6.1 Qualitative Analysis and Interpretability

In this section, we perform a qualitative analysis of the models trained with BoK-LM loss. Table 7 and Table 8 show comprehensive examples of dialogue generation using $DialoGPT_{BoK}$ on a DailyDialog and Person-Chat test instances, respectively. For each dialogue turn, we show the model-generated response. Additionally, we also show the top-8 tokens predicted by the feed-forward neural network (for computing BoK loss) given context vector ϕ_t. In Table 8, the "Persona" column denotes the persona profile used for response generation. Some tokens have a special character "Ġ", which can be interpreted as a space.

In both tables, we observe an overlapping of tokens between the generated response and the predicted tokens. For example, in Turn 4 of Table 7, the critical words in the generated response (sounds, good, deal) are in the top-8 predictions. Moreover, for the cases with less overlap, the generated response still aligns with the predicted tokens thematically. For instance, in Turn 3 of Table 8, the concept of the response matches with the predicted

Turn	Utterance (ground-truth)	Top-8 predicted tokens	Model Generated Response
1	i need to buy some flowers for my wife .	**kind**, Ġfresh, Ġred, Ġroses, ros, es, **Ġkind**, l	we have lots. what kind of flowers do you want?
2	perhaps you 'd be interested in red roses .	**ros**, **Ġred**, **es**, Ġsounds, good, <nok>, expensive, ide	no, i don't like red roses at all.
3	what would the roses cost me ?	**dozen**, Ġprice, **Ġdollars**, ars, oll, es, ten, Ġred	$ 20 per dozen. that sounds like a good deal to me.
4	they 're only $ 20 a dozen .	**Ġsounds**, dozen, bad, expensive, **Ġgood**, price, **deal**, reasonable	that sounds like a good deal. i'll take a dozen.

Table 7: A comprehensive example of dialogue generation with DialoGPT$_{\text{BoK}}$ on a DailyDialog test instance.

Persona of Speaker 1 (P1): i am engaged to be married. i am a college graduate. my favorite sport is soccer. i used to play sports in high school.
Persona of Speaker 2 (P2): i work in a warehouse driving a forklift. i lift weights every weekend. i am big and tall. i like watching horror movies. i like making barbecue.

Turn	Utterance (ground-truth)	Persona	Top-8 predicted tokens	Model Generated Response
1	janice here . what are you up to tonight ?	P2	**watching**, **Ġwatching**, Ġwork, <nok>, work, Ġworking, Ġlove, **Ġwatch**	hi! i am watching a horror movie. how are you?
2	they call me big shirley . i love scary movies . you ?	P1	**Ġlove**, Ġsports, **Ġwatch**, Ġplay, **Ġmovies**, **ovies**, Ġschool, **Ġwatching**	scary movies are great. i love watching soccer.
3	well , i am more into sports than i am horror flicks .	P2	Ġwork, Ġwatch, **Ġlove**, work, s, Ġworking, Ġsports, Ġgym	i love to lift weights. do you lift weights?
4	i see . i workout lifting weights . you ?	P1	Ġsports, **school**, **Ġplay**, Ġlove, Ġcollege, **Ġsoccer**, **soc**, **cer**	i used to play soccer in high school. you?

Table 8: A comprehensive example of dialogue generation with DialoGPT$_{\text{BoK}}$ on a Persona-Chat test instance.

token "gym". This refers to the effectiveness of BoK-LM loss in learning the context vector ϕ_t that guides the model to generate meaningful responses. Furthermore, ϕ_t can be interpreted by looking at the predicted tokens. This is how BoK loss enables post-hoc interpretability in the backbone model.

6.2 BoK-LM loss as Reference-Free Metric

In this section, we study the utility of BoK-LM loss as a reference-free metric for open-domain dialogue evaluation. We conduct our evaluation on various benchmark datasets like USR (Mehri and Eskenazi, 2020b), GRADE (Huang et al., 2020), PredictiveEngage (Ghazarian et al., 2020), and FED (Mehri and Eskenazi, 2020a) that contain human judgments for context-response pairs. We use DialoGPT$_{\text{BoW}}$ and DialoGPT$_{\text{BoK}}$ to compute the BoW-LM and BoK-LM loss, respectively. BoW-LM and BoK-LM losses are based on cross-entropy loss, where a lower score indicates better quality. As a result, they show a negative correlation with the human scores of the benchmark datasets.

In Table 9, we can observe that BoK-LM achieves comparable performance to the state-of-the-art metrics on the chit-chat datasets (GRADE-Dailydialog, PredictiveEngage, and FED). However, it shows weaker correlations for knowledge-grounded datasets (USR-Persona and Grade-Convai2) but still performs better than the referenced metrics such as BERTScore, BLEURT, and BERT-RUBER. Moreover, BoK-LM performs better than BoW-LM except for GRADE-DailyDialog dataset. Metrics typically exhibit better performance when applied to the dataset on which they were trained (Yeh et al., 2021). Since DialoGPT$_{\text{BoW}}$ is trained on DailyDialog and has more training data bias than DialoGPT$_{\text{BoK}}$, BoW-LM shows superior performance on GRADE-DailyDialog. However, it performs poorly on FED, a relatively difficult dataset. Nevertheless, BoK-LM achieves a decent performance on FED compared to the other metrics. This again verifies that BoK loss is more generalizable than BoW loss.

7 Conclusion

This paper proposes Bag-of-Keywords (BoK) loss, a novel auxiliary loss for training open-domain dialogue systems. The main idea of BoK loss is to improve the generalizability of Bag-of-Words (Bow) loss by predicting only the keywords or the core idea of the next response. We show that BoK loss enhances the generative performance of the vanilla T5 and DialoGPT models on the DailyDialog and

Metric	USR-Persona		GRADE-Convai2		GRADE-Dailydialog		PredictiveEngage		FED	
	P	S	P	S	P	S	P	S	P	S
BLEU-4	0.135	0.090*	0.003*	0.128	0.075*	0.184	-	-	-	-
METEOR	0.253	0.271	0.145	0.181	0.096*	0.010*	-	-	-	-
BERTScore	0.152	0.122*	0.225	0.224	0.129	0.100*	-	-	-	-
BLEURT	0.065*	0.054*	0.125	0.120	0.176	0.133	-	-	-	-
BERT-RUBER	0.266	0.248	0.309	0.314	0.134	0.128	-	-	-	-
MAUDE	0.345	0.298	0.351	-0.304	-0.036*	-0.073*	0.104	0.060*	0.018*	-0.094*
DEB	0.291	0.373	0.426	**0.504**	**0.337**	**0.363**	0.516	0.580	**0.230**	**0.187**
GRADE	0.358	0.352	**0.566**	**0.571**	0.278	0.253	**0.600**	0.622	0.134	0.118
HolisticEval	0.087*	0.113*	-0.030*	-0.010*	0.025*	0.020*	0.368	0.365	0.122	0.125
USR	**0.440**	**0.418**	**0.501**	**0.500**	0.057*	0.057*	**0.582**	**0.640**	0.114	0.117
USL-H	**0.495**	**0.523**	**0.443**	0.457	0.108*	0.093*	**0.688**	**0.699**	**0.201**	**0.189**
Dial-M	**-0.464**	**-0.486**	-0.310	-0.312	-0.111	-0.120	-0.570	-0.592	-0.127	-0.097
BoW-LM	-0.156	-0.124	-0.286	-0.252	**-0.419**	**-0.443**	-0.534	-0.572	-0.048*	-0.082*
BoK-LM	-0.261	-0.255	-0.318	-0.301	**-0.367**	**-0.383**	-0.581	**-0.632**	**-0.135**	**-0.151**

Table 9: Comparison of dialogue evaluation metrics with top-3 scores highlighted in bold. P and S indicate Pearson and Spearman's coefficients, respectively. All values are statistically significant to $p < 0.05$, unless marked by *.

Persona-Chat datasets when trained with BoK-LM loss. We also notice an improvement in the specificity of the generated response with the inclusion of BoK loss. We discuss the notion of interpretability that comes with the incorporation of BoK loss with comprehensive examples. Finally, we show that BoK-LM loss shows a moderate performance as a reference-free dialogue evaluation metric. In future work, we want to explore better keyword extraction methods and study the applicability of BoK loss in other NLG tasks.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. YAKE! Collection-Independent Automatic Keyword Extractor. In *Advances in Information Retrieval*, pages 806–810, Cham. Springer International Publishing.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Stephen Casper and et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. Survey Certification.

Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, Biao Cheng, and Nan Duan. 2022. DialogVED: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4852–4864, Dublin, Ireland. Association for Computational Linguistics.

Suvodip Dey and Maunendra Sankar Desarkar. 2023. Dial-M: A masking-based framework for dialogue evaluation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 77–84, Prague, Czechia. Association for Computational Linguistics.

Suvodip Dey, Maunendra Sankar Desarkar, Asif Ekbal, and Srijith P. K. 2023. DialoGen: Generalized long-range context representation for dialogue systems. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 372–386, Hong Kong, China. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail S. Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *ArXiv*, abs/1902.00098.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement:

An efficient metric for automatic evaluation of open-domain dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7789–7796.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Zhihua Jiang, Guanghui Ye, Dongning Rao, Di Wang, and Xin Miao. 2022. IM^2: an interpretable and multi-category integrated metric framework for automatic dialogue evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11091–11103, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1).

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.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. 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 128–138, Online. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ouyang Long and et al. 2022. Training language models to follow instructions with human feedback.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational*

Linguistics, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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. 21(1).

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddharth Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 553–562, New York, NY, USA. Association for Computing Machinery.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI Conference on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. DynaEval: Unifying turn and dialogue level evaluation. 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 5676–5689, Online. Association for Computational Linguistics.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021b. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2502–2516.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

A.1 Additional Implementation Details

For training data preparation related to BoK loss, we first extract the keywords from the next utterance using YAKE! (Campos et al., 2018, 2020). It outputs the keywords as a list with a decreasing order of relevance. We concatenate this list of keywords into a string and then tokenize it using the T5/GPT tokenizer. We consider the top-k tokens based on the maximum token limit ($|K_t|$). There are instances where the YAKE! could not find any keywords. In those cases, we add a special token (<nok>) in the label. In other words, the model is trained to predict <nok> for generic responses with no keywords.

As discussed, we studied the effectiveness of our proposed BoK loss by applying it to T5 and DialoGPT. We performed our experiment with DailyDialog and Persona-Chat datasets. For each dataset, we train a separate T5 and DialoGPT model. The two datasets and models only support the English language. The best model was selected for each training based on the validation loss. The training time of all the models is around 12-20 hours. Since we do not have any sampling during training and use a fixed seed (10), the models are reproducible. Furthermore, we generate the responses using beam search with a fixed configuration (described in Section 4.2). Because of that, we report the results of the model with a single run since they are deterministic. We use four data-specific baselines - DilaoFlow (≈ 900M parameters) [3], DialogVED (≈ 400M parameters) [4], DialoGEN (≈ 900M parameters) [5], and TransferTransfo (≈ 200M parameters) [6]. Codes of all the baselines are publicly available and have free license.

[3] github.com/ictnlp/DialoFlow
[4] github.com/lemuria-wchen/DialogVED
[5] github.com/SuvodipDey/DialoGen
[6] github.com/huggingface/transfer-learning-conv-ai

The referenced evaluation of the generated dialogues was conducted following the evaluation of DSTC7 Task 2 [7]. We used two different models to compute the BoK-LM loss in Table 9. For the knowledge-grounded datasets (USR-Persona, GRADE-Convai2), we used the $\text{DialoGPT}_{\text{BoK}}$ model trained on the Persona-Chat dataset. For the chit-chat datasets (GRADE-DailyDialog, Predictive Engage, and FED), we utilized the $\text{DialoGPT}_{\text{BoK}}$ model trained on the DailyDialog dataset. The same process is followed to compute the BoW-LM loss as well.

A.2 Related Works on Open-domain Dialogue Evaluation

Since we study the usefulness of our proposed loss as a reference-free metric, we add a short literature survey on open-domain dialogue evaluation. There are primarily two kinds of dialogue evaluation metrics- i) referenced and ii) reference-free. In referenced metrics, the generated response is compared with one or more reference utterances to evaluate its goodness. The most popular referenced metrics are word-overlapping based metrics like BLEU (Papineni et al., 2002), NIST (Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), Diversity (Li et al., 2016), and Entropy (Zhang et al., 2018b). There are also learning-based referenced metrics like ADEM (Lowe et al., 2017), RUBER (Tao et al., 2017), BERT-RUBER (Ghazarian et al., 2019), PONE (Lan et al., 2020), BERTScore (Zhang* et al., 2020), BLEURT (Sellam et al., 2020), etc. Conversely, the reference-free metrics are designed to evaluate dialogues without any references. As collecting good-quality references is expensive and needs human effort, most of the recent research focuses on developing reference-free metrics. Most of the methods formulate the dialogue evaluation problem as a classification task and use the classification score as the metric (Sinha et al., 2020; Sai et al., 2020; Huang et al., 2020; Zhang et al., 2021a). Metrics such as USR (Mehri and Eskenazi, 2020b), USL-H (Phy et al., 2020), FED (Mehri and Eskenazi, 2020a), HolisticEval (Pang et al., 2020), D-score (Zhang et al., 2021b), and QualityAdapt (Mendonca et al., 2022) combine various sub-metrics to provide more holistic evaluation. Dial-M (Dey and Desarkar, 2023) adopts a masking-based approach that utilizes masked language modeling (MLM) loss as the evaluation

[7] github.com/mgalley/DSTC7-End-to-End-Conversation-Modeling/tree/master/evaluation/src

Turn	Utterance (ground-truth)	Top-8 predicted tokens (BoK)	Top-8 predicted tokens (BoW)
1	i need to buy some flowers for my wife .	(kind, 0.1113), (Ġfresh, 0.0913), (Ġred, 0.0629), (Ġroses, 0.0332), (ros, 0.0304), (es, 0.0277), (Ġkind, 0.0249), (l, 0.0199)	(Ġ?, 0.0856), (Ġroses, 0.0649), (Ġyou, 0.0382), (Ġkind, 0.0347), (Ġ., 0.0314), (Ġlike, 0.0295), (how, 0.0234), (Ġare, 0.0224)
2	perhaps you 'd be interested in red roses .	(ros, 0.2161), (Ġred, 0.2063), (es, 0.0894), (Ġsounds, 0.0227), (good, 0.0147), (<nok>, 0.0118), (expensive, 0.0083), (ide, 0.0079)	(Ġ?, 0.0896), (Ġ., 0.0816), (Ġ,, 0.0686), (Ġlike, 0.0453), (Ġi, 0.0379), (Ġroses, 0.0285), (Ġthey, 0.0215), (how, 0.0161)
3	what would the roses cost me ?	(dozen, 0.7592), (Ġprice, 0.0139), (Ġdollars, 0.0111), (ars, 0.009), (oll, 0.006), (es, 0.0037), (ten, 0.0033), (Ġred, 0.0032)	(Ġ., 0.121), (Ġeach, 0.1), (Ġdollars, 0.0324), (Ġper, 0.0272), (Ġ$, 0.0264), (Ġdozen, 0.0259), (they, 0.0228), (the, 0.0223)
4	they 're only $ 20 a dozen .	(Ġsounds, 0.1743), (dozen, 0.1095), (bad, 0.0831), (expensive, 0.0669), (Ġgood, 0.0486), (price, 0.0395), (deal, 0.0219), (reasonable, 0.0185)	(Ġ., 0.0788), (Ġ?, 0.0409), (Ġ,, 0.0356), (that, 0.0326), (Ġi, 0.0251), (Ġa, 0.0233), (i, 0.0231), (how, 0.0226)

Table 10: Comparison of predicted tokens on a DailyDialog test instance.

score. Metrics like IM^2 (Jiang et al., 2022) leverage various evaluation metrics to enhance the evaluation of different dialogue aspects.

A.3 Comparison of top-k Predicted Tokens (BoK vs. BoW)

Table 10 shows the top-k tokens associated with the BoW and BoK loss (along with the probability scores) for the examples shown in Table 7. We use the DialoGPT$_{\text{BoK}}$ and DialoGPT$_{\text{BoW}}$ to find the top-k BoK and BoW, respectively. We can observe that the top-8 tokens associated with the BoW loss contain a lot of punctuation and stopwords as they are trained to predict all the words/tokens of the next utterance. In contrast, the top-k tokens associated with the BoK are more aligned with the conversation topic. For example, in Turn 4 of Table 10, all the tokens predicted by the BoK method are relevant and can potentially steer the conversation in a meaningful direction. However, for the BoW method, the predicted words are mostly punctuations and stopwords.