Entriever: Energy-based Retriever for Knowledge-Grounded Dialog Systems

Yucheng Cai[1], **Ke Li**[1], **Yi Huang**[2,*], **Junlan Feng**[2], **Zhijian Ou**[1,*]
[1]SPMI Lab, EE Department, Tsinghua University, [2]China Mobile Research Institute
{cyc22,ke-li24}@mails.tsinghua.edu.cn, ozj@tsinghua.edu.cn
{huangyi,fengjunlan}@chinamobile.com

Abstract

A retriever, which retrieves relevant knowledge pieces from a knowledge base given a context, is an important component in many natural language processing (NLP) tasks. Retrievers have been introduced in knowledge-grounded dialog systems to improve knowledge acquisition. In knowledge-grounded dialog systems, when conditioning on a given context, there may be multiple relevant and correlated knowledge pieces. However, knowledge pieces are usually assumed to be conditionally independent in current retriever models. To address this issue, we propose Entriever, an energy-based retriever. Entriever directly models the candidate retrieval results as a whole instead of modeling the knowledge pieces separately, with the relevance score defined by an energy function. We explore various architectures of energy functions and different training methods for Entriever, and show that Entriever substantially outperforms the strong cross-encoder baseline in knowledge retrieval tasks. Furthermore, we show that in semi-supervised training of knowledge-grounded dialog systems, Entriever enables effective scoring of retrieved knowledge pieces and significantly improves end-to-end performance of dialog systems.

1 Introduction

Recently, with the development of large language models (LLMs), dialog systems have attracted increasing research interests. Although LLMs have shown an astonishing ability in open-domain question answering, they still often lack accuracy and make mistakes about certain facts in specific domains, such as customer services. Recent studies (Shuster et al., 2022; Izacard et al., 2022b; Cai et al., 2023) have shown that the integration of knowledge retrieval into dialog systems can substantially enhance the precision of knowledge and mitigate the occurrence of hallucinations.Therefore, knowledge retrieval is crucial to improve dialog systems, especially for those that require knowledge grounding.

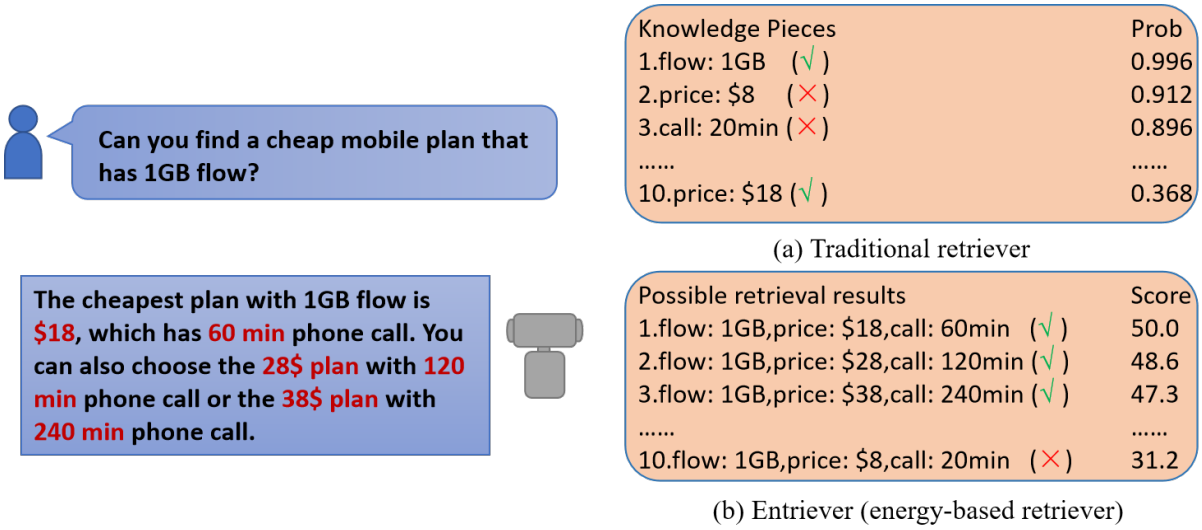
Currently, two types of methods are prevalent for knowledge retrieval, statistical-based methods (like BM25) and dense retrieval methods (like DPR (Karpukhin et al., 2020a)). Both methods aim to find the most relevant piece of knowledge from a given knowledge base (KB). However, when dealing with situations where multiple knowledge pieces from the KB might be relevant given certain context, which are common in real-life applications, both methods fail to account for the interrelationship among knowledge pieces and instead only model them separately. When a retrieval task requires the most relevant n pieces of knowledge as a collective whole, these methods typically obtain the top n results by relying on individual similarity scores. This ignores the relationship between these pieces, which could cause the retrieved pieces to contain repetitive information or miss important information.

To address the problems mentioned above, we propose using energy-based language models (ELMs) to model multiple knowledge pieces as a whole token sequence, rather than modeling the relevant knowledge pieces separately. A candidate retrieval result consists of multiple knowledge pieces. ELMs assign an energy score to each candidate result and use the score to distinguish positive candidates from negative ones, which is suitable for retrieval tasks. In previous research (Wang et al., 2015, 2017; Bakhtin et al., 2021), ELMs have been successfully used to calculate sentence scores in automatic speech recognition (ASR) and natural language generation (NLG). Different training strategies, such as noise contrastive estimate (NCE) and maximum likelihood estimate (MLE), have been explored (Wang et al., 2017; Wang and Ou, 2018a; Liu et al., 2023). However, to the best of our knowledge, our energy-based retriever, referred to as Entriever, is the first to use ELMs in retrieval tasks. We explore various MLE training approaches and find that the use of residual ELMs can greatly improve performance, shedding light on future work.

Moreover, note that the energy score of a candidate result is defined as the negative log probability up to an additive constant. The unnormalized nature of the energy score enables the proposed Entriever to model the probability of a candidate result without the need to access the entire KB. This feature is useful in building semi-supervised knowledge-grounded dialog systems. Semi-supervised knowledge-grounded dialog systems have seen significant progress recently (Deng et al.,

*Corresponding authors. This work is partly supported by the National Science and Technology Major Project (2023ZD0121401) and Guangxi Science and Technology Project (2022AC16002). The code for this paper is available at https://github.com/thu-spmi/entriever

Findings of the Association for Computational Linguistics: ACL 2025, pages 21462–21474



(a) Traditional retriever

(b) Entriever (energy-based retriever)

Figure 1: An illustration of the difference between (a) traditional retriever and (b) Entriever in the retrieval task for knowledge-grounded dialog systems. The traditional retriever judges the probability of each slots independently while Entriever assigns a score for each candidate retrieval result as a whole. The traditional retriever ignores the interrelationship between knowledge pieces – in this case the plan with 1GB flow should be at least $18, and makes mistakes on the value of the price and call. In contrast, Entriever can better model the interrelationships between knowledge pieces and demonstrate better overall performance in the task.

2023; Cai et al., 2023). These systems can make use of both labeled and unlabeled dialog data, reducing the reliance on costly manually labeled data and improving the efficiency of model training. However, it has been pointed out that the development of such systems can be difficult when the KB is not available in unlabeled data (Cai et al., 2023). This difficulty can be overcome by introducing Entriever that can model the probabilities of candidate results without the need to access the entire KB. Experiments show that using the Entriever significantly improves the performance of semi-supervised knowledge-grounded dialog systems.

Experiments are conducted on several knowledge-grounded dialog datasets, including MobileCS (Ou et al., 2022), Camrest (Wen et al., 2017), In-Car (Eric et al., 2017), and Woz2.1 (Eric et al., 2020). We evaluate the performance of Entriever itself through the retrieval task and the performance gain that Entriever brings for semi-supervised dialog systems through the response generation task.

In summary, the main contributions of this work are as follows.

- We propose to use an energy-based retriever (Entriever) to model each candidate retrieval result as a whole, which consists of multiple knowledge pieces, instead of modeling the knowledge pieces separately in knowledge-grounded dialog systems.
- We explore different architectures of energy functions and different training methods to train the Entriever and demonstrate the superiority of the proposed Entriever over previous methods.
- The proposed Entriever can model the retrieval probability without the need to access the entire KB, which improves the performance of semi-supervised knowledge-grounded dialog system.

2 Related Work

2.1 Knowledge Retrieval for Dialog Systems

Recent research such as RAG (Lewis et al., 2020) and REALM (Guu et al., 2020) have introduced knowledge retrieval models into conditional generation, which greatly improves the quality of generated responses in knowledge-intensive tasks such as open-domain question answering and knowledge-grounded dialog systems. Retriever, which ranks relevant knowledge pieces from the KB given a context, is important for knowledge retrieval. Previous works use statistic-based retrievers (e.g. BM25 (Robertson et al., 2004)) or neural network based retrievers (e.g. DPR (Karpukhin et al., 2020b)). There are several recent studies that improve over the original DPR retrievers. Re2G (Glass et al., 2022) proposed to use a cross-encoder reranker to improve the performance over the dual-encoder retriever. Contriever (Izacard et al., 2022a) performed unsupervised pretraining using contrastive learning to improve the performance of the retriever. RetroMAE (Xiao et al., 2022) was pretrained with mask auto-encoding objective function to better capture the information of the whole sentence. Recently, LLM-based retrievers have gained much attention and achieve improvements over the traditional retrievers in some aspects (Shen et al., 2024; Khramtsova et al., 2024). However, all these works aim to retrieve the most relevant knowledge piece from the KB given certain searching context. Yet in real-life applications,

multiple knowledge pieces from the KB might be relevant and helpful for response generation. Our work proposes to use an energy-based retriever (Entriever) to better model the inter-relationship among knowledge pieces instead of modeling them independently.

2.2 Energy-based Language Models (ELMs)

Energy-based language models (ELMs) parameterize an unnormalized distribution for natural sentences via an energy function, which can be very flexibly defined. In previous studies, ELMs have shown promising performances in scoring for sentences in various applications such as computation of sentence likelihoods (Wang et al., 2015, 2017; Wang and Ou, 2017, 2018a,b; Gao et al., 2020), text generation (Deng et al., 2020), language model pretraining (Clark et al., 2020), calibrated natural language understanding (He et al., 2021) and calculating sentence scores in automatic speech recognition (ASR) (Liu et al., 2023). Our work leverages the modeling flexibility of ELMs to model whether the ensemble of multiple knowledge pieces is suitable given a context.

There are two main training methods for ELMs, the maximum likelihood estimate (MLE) and the noise contrastive estimate (NCE) (Gutmann and Hyvärinen, 2010). In this work, we mainly explore different architectures of energy functions and different sampling methods using MLE methods. In MLE, calculating gradients of the log likelihood usually resorts to Monte Carlo sampling methods. Two widely-used classes of sampling methods are importance sampling (IS) and Markov Chain Monte Carlo (MCMC) (Liu, 2001). MCMC covers a range of specific algorithms and Metropolis independent sampling (MIS), where the proposed Markov move is generated independent of the previous state, is explored in this work. Meanwhile, residual ELM (Deng et al., 2020), which models the ELM over a normalized model instead of modeling from scratch, is explored to study whether it can bring performance gain to the non-residual ELM.

2.3 Knowledge-Grounded Dialog Systems

Knowledge-Grounded Dialog Systems aim to generate informative and meaningful responses based on both conversation context and external knowledge sources (Dinan et al., 2018; Kim et al., 2020; Zhao et al., 2020; Li et al., 2022). Semi-supervised knowledge-grounded dialog systems have seen significant progress recently (Li et al., 2020; Paranjape et al., 2021; Deng et al., 2023; Cai et al., 2023, 2024). The use of semi-supervised training in knowledge-based dialog systems has been shown to greatly improve performance (Paranjape et al., 2021; Deng et al., 2023; Cai et al., 2023). In a semi-supervised knowledge-grounded dialog system, the knowledge required for response generation is not annotated in unlabeled data, and needs to be predicted by an inference model. An annoying difficulty in semi-supervised training is to accurately score the pseudo knowledge labels generated by the inference model on the unlabeled data. In previous efforts to build semi-supervised knowledge-grounded dialog systems (Cai et al., 2023), researchers have to approximate the retrieval probability of the latent knowledge, using only the positive samples predicted by the inference model and ignoring the possible negative samples, due to that the KB is unavailable over unlabeled data. In contrast, Entriever directly models the retrieval probability of the latent knowledge as a whole. Our work explored using Entriever to better score the generated knowledge on the unlabeled data and significantly improved the performances of semi-supervised knowledge-grounded dialog system.

3 Preliminary

3.1 Knowledge-Grounded Dialog Systems

Knowledge-grounded dialog systems retrieve relevant knowledge pieces given the dialog context and generate system response using the retrieved knowledge. Our settings of the knowledge-grounded dialog system is similar to (Cai et al., 2023). Assume that we have a dialog with T turns of user utterances and system responses, denoted by $u_1, r_1, \cdots, u_T, r_T$ respectively. At turn t, based on the dialog context, the system queries a task-related KB to obtain relevant knowledge and generates appropriate responses. The KB is made up of N knowledge pieces[1], denoted by $\{k^1, k^2, \cdots, k^N\}$ and the knowledge pieces that are relevant for the system to respond at turn t are denoted by ξ_t. In knowledge-grounded dialog systems, the joint likelihood of the relevant knowledge pieces ξ_t and the response r_t given the context c_t and user input u_t is optimized at each turn t in a dialog session. The likelihood is decomposed into a knowledge retrieval probability p_θ^{ret} and a response generation probability p_θ^{gen}, as follows:

$$p_\theta(\xi_t, r_t|c_t, u_t) = p_\theta^{\text{ret}}(\xi_t|c_t, u_t) \times p_\theta^{\text{gen}}(r_t|c_t, u_t, \xi_t) \tag{1}$$

The model parameters are collectively denoted by θ, which can actually split into two parts $\theta = (\theta^{\text{ret}}, \theta^{\text{gen}})$.

The retrieval model p_θ^{ret} is introduced to retrieve knowledge from the KB. Traditionally, knowledge pieces in the KB are modeled independently when multiple knowledge pieces are retrieved. Particularly, the knowledge piece ξ_t necessary for turn t is represented by $\xi_t \triangleq \xi_{t,1} \oplus \xi_{t,2} \oplus \cdots \oplus \xi_{t,N}$, where $\oplus$ denotes sequence concatenation. $\xi_{t,i} = k^i$ if the knowledge piece k^i is relevant to the response r_t; otherwise, $\xi_{t,i}$ is set to be empty. Therefore, the retrieval probability can be written as follows:

$$p_\theta^{\text{ret}}(\xi_t|c_t, u_t) = \prod_{i=1}^{N} p_\theta^{\text{ret}}(\xi_{t,i}|c_t, u_t) \tag{2}$$

which we refer to as traditional retriever. However, a drawback of traditional retrievers is that it ignores

[1]The form of knowledge pieces can be flexible, for examples, documents or items. In this paper, the knowledge pieces are mainly in the form of entities with attributes, or, say, slot-value pairs.

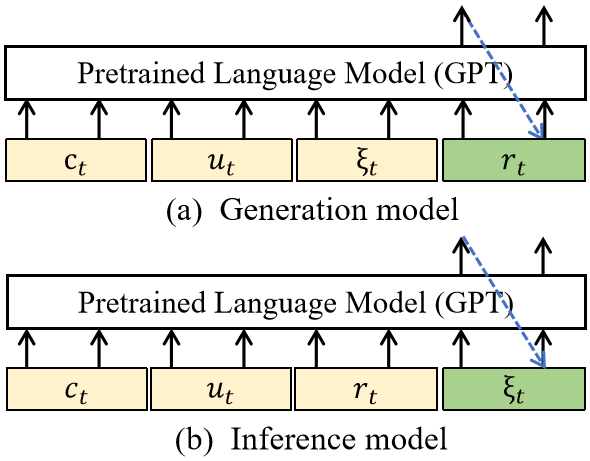


Figure 2: The models in semi-supervised training procedure: (a) the generation model, (b) the inference model. All of the variables c_t, u_t, ξ_t, r_t are represented by token sequences in our experiments.

the interrelationship between the knowledge pieces and only models the knowledge pieces independently, since different knowledge pieces may contain similar or correlated information.

Traditional retrievers using Eq. (2) can be implemented based on dual-encoders or cross-encoders. In vertical domains where the KBs are small, cross-encoder retrievers are often used for their improved retrieval performance compared to dual-encoder retrievers. In general domains where KBs are large and cross-encoders are computationally prohibitive, dual-encoder based retrievers are preferable for their reduced computation cost with fast k-nearest neighbor search library such as FAISS (Karpukhin et al., 2020a). In this work, the KBs in our experiments are small ones; thus a cross-encoder retriever based on BERT is used to realize $p_\theta^{\text{ret}}(\xi_{t,i}|c_t, u_t)$, using $c_t \oplus u_t \oplus k^i$ as input, as shown in Figure 3(a).

The generation probability $p_\theta^{\text{gen}}(r_t \mid c_t, u_t, \xi_t)$ is instantiated with a GPT2 model using an autoregressive loss function, which is shown in Figure 2(a):

$$p_\theta^{\text{gen}}(r_t \mid c_t, u_t, \xi_t) = \prod_{l=1}^{|r_t|} p_\theta^{\text{gen}}(r_t^{(l)} \mid c_t, u_t, \xi_t, r_t^{(1)}, \ldots, r_t^{(l-1)}) \quad (3)$$

where $|\cdot|$ denotes the length in tokens, and $r_t^{(l)}$ the l-th token of the response r_t.

In training, the ground truth ξ_t, which is annotated in the dataset, is used to maximize the log probabilities in Eq. (2) - (3). In testing, according to Eq. (1), we firstly retrieve relevant slot-value pairs ξ_t; then, we generate a_t and r_t, based on retrieved ξ_t. To be specific, to retrieve knowledge pieces using the cross-encoder retriever in Eq. (2), we threshold $p_\theta^{\text{ret}}(\xi_{t,i} = k^i|c_t, u_t), i = 1, \cdots, N$, similar to (Cai et al., 2023).

3.2 Semi-Supervision in Knowledge-Grounded Dialog Systems

Semi-Supervision aims to leverage both labeled and unlabeled data. Following (Cai et al., 2023), our semi-supervised dialog systems use latent variable model and the joint stochastic approximation (JSA) algorithm (Ou and Song, 2020) to optimize the latent variable model. As the knowledge pieces are annotated in labeled data and unavailable in unlabeled data, the relevant knowledge pieces $\xi_{1:T}$ are viewed as the latent variable for a dialog. Therefore, the generation model can be written as $p_\theta(\xi_{1:T}, r_{1:T}|u_{1:T})$ and the inference model can be written as $q_\phi(\xi_{1:T}|u_{1:T}, r_{1:T})$ to approximate the true posterior $p_\theta(\xi_{1:T}|u_{1:T}, r_{1:T})$. Both probabilities can be decomposed into the turn , as pointed out in (Cai et al., 2022) and shown in Figure 2:

$$p_\theta(\xi_{1:T}, r_{1:T}|u_{1:T}) = \prod_{t=1}^{T} p_\theta(\xi_t, r_t|c_t, u_t) \quad (4)$$

$$q_\phi(\xi_{1:T}|u_{1:T}, r_{1:T}) = \prod_{t=1}^{T} q_\phi(\xi_t|c_t, u_t, r_t) \quad (5)$$

In supervised training, the ground truth knowledge ξ_t is annotated in the dataset and thus can be directly used to maximize the probabilities in Eq. (4) and Eq. (5). In semi-supervised training, the ground truth knowledge ξ_t is not annotated for unlabeled data and should be inferred. Particularly, Metropolis independent sampling (MIS) is applied to draw samples from the true posterior $p_\theta(\xi_t|c_t, u_t, r_t)$ using the inference model $q_\phi(\xi_t|c_t, u_t, r_t)$ as a proposal. A recursive turn-level MIS sampler is used to sample $\xi_{1:T}$, as developed in (Cai et al., 2022). At each turn t, the MIS sampler works in a propose, accept or reject way, as follows:

1) Propose $\xi_t' \sim q_\phi(\xi_t|c_t, u_t, r_t)$.

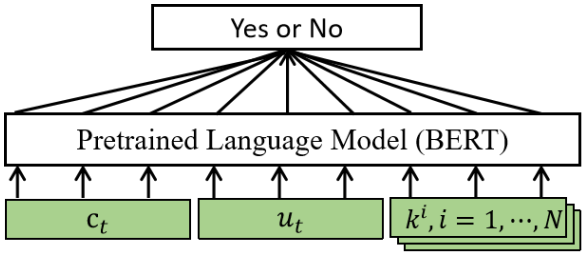
2) Simulate $\eta \sim \text{Uniform}[0, 1]$ and let

$$\xi_t = \begin{cases} \xi_t', & \text{if } \eta \le \min\left\{1, \frac{w(\xi_t')}{w(\tilde{\xi}_t)}\right\} \\ \tilde{\xi}_t, & \text{otherwise} \end{cases} \quad (6)$$

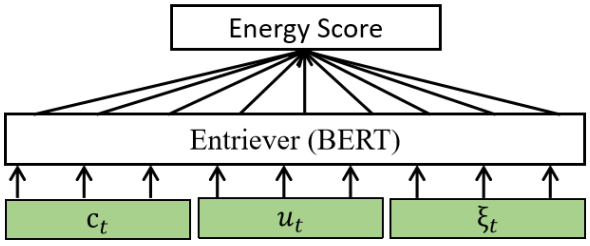
where $\tilde{\xi}_t$ denotes the cached latent knowledge, and the importance weight $w(\xi_t)$ between the target and the proposal distribution is defined as follows:

$$\begin{aligned} w(\xi_t) &= \frac{p_\theta(\xi_t|c_t, u_t, r_t)}{q_\phi(\xi_t|c_t, u_t, r_t)} \\ &= \frac{p_\theta(\xi_t, r_t|c_t, u_t)}{q_\phi(\xi_t|c_t, u_t, r_t)} \frac{1}{p_\theta(r_t|c_t, u_t)} \\ &\propto \frac{p_\theta^{\text{ret}}(\xi_t|c_t, u_t) \times p_\theta^{\text{gen}}(r_t|c_t, u_t, \xi_t)}{q_\phi(\xi_t|c_t, u_t, r_t)} \end{aligned} \quad (7)$$

The term $p_\theta(r_t|c_t, u_t)$ is canceled out, since it appears in both the numerator and denominator of $w(\xi_t')/w(\tilde{\xi}_t)$; and we only need to calculate the last line in Eq. (7) and use it as the importance weight. The details of the JSA algorithm for training semi-supervised knowledge-grounded dialog systems are given in Appendix A.



(a) Cross-encoder retrieval model



(b) Entriever

Figure 3: The architecture of retrieval models: (a) the cross-encoder retrieval model, which models the knowledge pieces independently, and (b) Entriever, which models the ensemble of relevant knowledge pieces. All of the variables $c_t, u_t, \xi_t, r_t, k^i$ are represented by token sequences.

Note that the retrieval probability $p_\theta^{\text{ret}}(\xi_t|c_t, u_t)$ is needed to calculate the importance weight $w(\xi_t)$ in Eq. (7). However, in previous works, using Eq. (2) to calculate the retrieval probability requires to access the entire KB, which, however, is often not available for unlabeled data in semi-supervised knowledge grounded systems. To address this issue, we propose to use Entriever to calculate the retrieval probability of the latent knowledge ξ_t as defined in Eq. (8).

4 Method

As introduced in Section 3, there are two motivations to develop energy-based retriever (Entriever) - better modeling interdependencies between knowledge pieces and better enabling semi-supervised knowledge-grounded dialog systems. The former aims to model the candidate retrieval result as a whole, and the latter to model the retrieval probability without the need to access the entire knowledge base. In Entriever, as shown in Figure 3(b), the retrieval probability is defined by:

$$p_\theta^{\text{ret}}(\xi_t|c_t, u_t) = \frac{\exp(-U_\theta(c_t, u_t, \xi_t))}{Z_\theta(c_t, u_t)} \propto \exp(-U_\theta(c_t, u_t, \xi_t)) \quad (8)$$

where $U_\theta(c_t, u_t, \xi_t)$ is the energy function. In this work, we initialize U_θ with BERT, similar to (Deng et al., 2020). $Z_\theta(c_t, u_t)$ denotes the normalizing constant.

4.1 Architecture of Entriever

The architecture of the energy function $U_\theta(c_t, u_t, \xi_t)$ in Eq. (8) can be very flexibly defined (Liu et al., 2023). In our work, a bi-directional text encoder (e.g., BERT) is used to encode the input c_t, u_t, ξ_t and we denote the encoder output (hidden vectors) by $\text{enc}_\theta(x)$. At position i, we have $\text{enc}_\theta(x)[i]$. Then, the energy is defined as:

$$U_\theta(c_t, u_t, \xi_t) = -\text{Linear}\left(\sum_{i=1}^{|x|} \text{enc}_\theta(x)[i]\right) \quad (9)$$

where $\text{Linear}(\cdot)$ denotes a trainable linear layer whose output is a scalar and $x \triangleq c_t \oplus u_t \oplus \xi_t$ is the concatenation of the input sequence c_t, u_t, ξ_t.

Orthogonal to the neural architecture used to define an energy function, we can define a residual form for an energy function, i.e., in the form of exponential tilting of a reference distribution (Wang et al., 2017; Deng et al., 2020). Specifically in our case, the retrieval probability can be defined as follows:

$$p_\theta^{\text{ret}}(\xi_t|c_t, u_t) \propto p^{\text{ref}}(\xi_t|c_t, u_t) \exp(-U_\theta(c_t, u_t, \xi_t)) \quad (10)$$

where a reference distribution $p^{\text{ref}}(\xi_t|c_t, u_t)$ is introduced. For simplicity, though with abuse of notation, we still use $U_\theta(c_t, u_t, \xi_t)$ to denote the residual energy function, as for both non-residual and residual forms, we still use Eq. (9) to realize U_θ and we will see in Section 4.2 that the formulas in model training share the same expressions. The role of residual energy is to fit the difference between the target distribution and the reference distribution.

In this work, the reference distribution $p^{\text{ref}}(\xi_t|c_t, u_t)$ is set to be the traditional retrieval distribution shown in Eq. (2), which is usually closer to the target distribution than from uniform[2]. Therefore, the residual Entriever only needs to learn the difference between the target distribution and the baseline distribution, which is easier to train. Remarkably, as $p^{\text{ref}}(\xi_t|c_t, u_t)$ is irrelevant to θ, the residual Entriever can be optimized the same as the non-residual Entriever, which is introduced in Section 4.2. In our experiments, we compare both forms of Entrievers and find that the residual Entriever reduces the training difficulty and brings substantial improvement to the overall performance of dialog systems.

4.2 Training of Entriever

First, it should be noted that the formulas presented in Section 4.2 apply to both non-residual and residual forms of Entrievers, defined in Eq. (8) and Eq. (10) respectively, unless otherwise specified.

MLE base model training of Entriever is to learn the energy function $U_\theta(x)$, by using the negative log likelihood as the loss function:

$$\mathcal{J}_\theta = -\log p_\theta^{\text{ret}}(\xi_t|c_t, u_t). \quad (11)$$

The gradient of the loss function $\frac{\partial J_\theta(x)}{\partial \theta}$ can be derived

[2]The non-residual form in Eq. (8) can be viewed as a constrained subclass of the residual form in Eq. (10), where the reference distribution is chosen to be uniform.

as follows (Ou et al., 2024):

$$\frac{\partial \mathcal{J}_\theta(\xi_t|c_t,u_t)}{\partial\theta} = -\frac{\partial U_\theta(c_t,u_t,\xi_t)}{\partial\theta} + \mathbb{E}_{\xi_t\sim p_\theta^{\text{ret}}}\left[\frac{\partial U_\theta(c_t,u_t,\xi_t)}{\partial\theta}\right] \tag{12}$$

The challenge in calculating the gradient in Eq. (12) is that calculating the second term as an expectation requires sampling from the unnormalized distribution $p_\theta^{\text{ret}}(\xi_t|c_t,u_t)$, which is generally intractable. Similar to (Parshakova et al., 2019; Liu et al., 2023), we compare two sampling approaches, Metropolis independence sampling (MIS) and important sampling (IS). Both approaches require a proposal distribution, which is set to be the traditional retrieval distribution in Eq. (2) in this work and is denoted by $q(\xi_t|c_t,u_t)$. Note that here we drop any parameters related to the proposal distribution, since it is always fixed during the training of Entriever.

For the residual Entrievers in our experiments, we use the reference distribution $p^{\text{ref}}(\xi_t|c_t,u_t)$ as the proposal distribution $q(\xi_t|c_t,u_t)$ (i.e., both set to be the traditional retrieval distribution), we have the importance weight in the following simple form:

$$\frac{p_\theta^{\text{ret}}(\xi_t|c_t,u_t)}{q(\xi_t|c_t,u_t)} \propto \exp(U_\theta(c_t,u_t,\xi_t)). \tag{13}$$

4.2.1 Importance Sampling (IS)

Instead of directly sampling from the intractable distribution $p_\theta^{\text{ret}}(\xi_t|c_t,u_t)$, importance sampling draw proposal samples from a tractable distribution $q(\xi_t|c_t,u_t)$ (Liu, 2001). The importance weight $\frac{p_\theta^{\text{ret}}(\xi_t|c_t,u_t)}{q(\xi_t|c_t,u_t)}$ is calculated and renormalization is taken to calculate the expectation. Specifically, to estimate the second term in Eq. (12), we can use the importance sampling method with the proposal distribution $q(\xi_t|c_t,u_t)$:

$$\mathbb{E}_{\xi_t\sim p_\theta^{\text{ret}}(\xi_t|c_t,u_t)}\left[\frac{\partial}{\partial\theta}U_\theta(c_t,u_t,\xi_t)\right] \approx \frac{\sum_{\xi_t}\frac{p_\theta^{\text{ret}}(\xi_t|c_t,u_t)}{q(\xi_t|c_t,u_t)}\frac{\partial U_\theta(c_t,u_t,\xi_t)}{\partial\theta}}{\sum_{\xi_t}\frac{p_\theta^{\text{ret}}(\xi_t|c_t,u_t)}{q(\xi_t|c_t,u_t)}}, \xi_t\sim q(\xi_t|c_t,u_t)$$

where the samples are from the proposal distribution $q(\xi_t \mid c_t,u_t)$, which is set to be the traditional retrieval distribution in Eq. (2). They can be trivially obtained by sampling from N independent binary distributions.

4.2.2 Metropolis Independence Sampling (MIS)

Similar to the IS approach, the Metropolis Indepenence Sampling (MIS) approach draws proposal samples from the tractable proposal distribution $q(\xi_t \mid c_t,u_t)$. Unlike the IS approach, which uses renormalization and weighted averaging techniques to estimate the expectation term, MIS uses Markov Chain Monte Carlo (MCMC) to obtain samples from the target distribution $p_\theta^{\text{ret}}(\xi_t \mid c_t,u_t)$. MIS is a special case of Metropolis-Hasting (Liu, 2001) and has been applied for ELM in (Wang and Ou, 2017; Liu et al., 2023).

In experiments, we run the Markov chain for T steps. $\xi_t^{(0)}$ is randomly initialized. At step $\tau = 1,\cdots,T$, generate a proposal sample ξ_t' from $q(\xi_t \mid c_t,u_t)$, and accept $\xi_t^{(\tau)} = \xi_t'$ with probability

$$\min\left\{1, \frac{p_\theta^{\text{ret}}(\xi_t' \mid c_t,u_t)/q(\xi_t'|c_t,u_t)}{p_\theta^{\text{ret}}(\xi_t^{(\tau-1)} \mid c_t,u_t)/q(\xi_t^{(\tau-1)}|c_t,u_t)}\right\},$$

otherwise set $\xi_t^{(\tau)} = \xi_t^{(\tau-1)}$. Then we can use the samples $\{\xi_t^{(1)},...,\xi_t^{(T)}\}$ to approximate the second term in Eq. (12) via Monte Carlo averaging:

$$\mathbb{E}_{\xi_t\sim p_\theta^{\text{ret}}(\xi_t|c_t,u_t)}\left[\frac{\partial}{\partial\theta}U_\theta(c_t,u_t,\xi_t)\right] \approx \frac{\sum_{\tau=1}^{T}\frac{\partial U_\theta(c_t,u_t,\xi_t^{(\tau)})}{\partial\theta}}{T}$$

4.3 Using Entriever to retrieve knowledge

4.3.1 Testing retrieval capability of Entriever

During testing, for Entriever in Eq. (8), the retrieval candidate with the highest score is taken as the final retrieval result. The retrieval task in the knowledge grounded dialog system aims at retrieving the complete set of useful knowledge pieces from the KB given the context. Therefore, a KB with N knowledge pieces will yield 2^N possible combination of retrieval candidates, which is impossible to enumerate. Therefore, some traditional retriever (as defined in Eq. (2)) is firstly used as a proposal[3]. Only the K retrieval candidates with the highest retrieval probabilities proposed from the traditional retriever are scored by Entriever. The 2^N possible retrieval results are defined according to the traditional retriever, which does not consider dependency among different knowledge pieces. Therefore, we can apply the Viterbi algorithm (Forney, 1973) to find the top-K retrieval candidates from the traditional retriever, instead of enumerating the probability of all 2^N possible combination of retrieval results to reduce computational complexity. We perform an ablation study to study the effect of K on the retrieval results in Table 5.

The pseudocode of the used Viterbi algorithm is shown in Algorithm 1. It runs like a beam search, with beam width K. The index of knowledge pieces $(1,2,\cdots,N)$ can be viewed as time steps, and at each step, there are two possible (step dependent) states to select from (select or not select). Each retrieval result is a knowledge piece subset, which can be viewed as a path through the $2 \times N$ lattice. The Viterbi algorithm can be applied to find the top-K paths $(\eta_1,\eta_2,\cdots,\eta_K)$ traversing the lattice with their probabilities $p_{\eta_1},p_{\eta_2},\cdots,p_{\eta_K}$.

4.3.2 Leveraging Entriever in Semi-Supervised Knowledge-Grounded Dialog Systems

In semi-supervised training of knowledge-grounded dialog systems, we need to calculate the importance weight $w(\xi_t)$ in Eq. (7) in order to properly filter the generated pseudo labels for unlabeled data. This involves calculating the retrieval probability $p_\theta^{\text{ret}}(\xi_t|c_t,u_t)$ for the pseudo

[3]In this work, we use the cross-encoder retriever for the proposal, but dual-encoders can be used as well.

Algorithm 1 The Viterbi algorithm to find the top-K retrieval candidates from the traditional retriever

Require: N knowledge pieces $k^1, k^2, \cdots, k^N$ with corresponding retrieval probabilities $p_1, p_2, \cdots, p_N$ calculated by the traditional retriever Eq. (2).
Initialize the retrieval knowledge piece subsets $\eta_1, \eta_2, \cdots, \eta_K$ to empty;
Initialize the retrieval probabilities $p_{\eta_1}, p_{\eta_2}, \cdots, p_{\eta_K}$ to be 1;
for i =1 to N **do**
  Consider to expand the K knowledge piece subsets by adding k^i or not. If k^i is selected and added to some $\eta_j, j = 1, \cdots, K$, the probability p_{η_j} is multiplied with p_i; otherwise, p_{η_j} is multiplied with $1 - p_i$;
  Calculate the possible $2K$ probabilities for the subset expansion: $p_{\eta_1} * p_i, p_{\eta_1} * (1-p_i), p_{\eta_2} * p_i, p_{\eta_2} * (1-p_i), \cdots, p_{\eta_K} * p_i, p_{\eta_K} * (1-p_i)$;
  Select the top-K results from the $2K$ results based on their probabilities, update $\eta_1, \eta_2, \cdots, \eta_K$;
end for
return The top-K retrieval results $\eta_1, \eta_2, \cdots, \eta_K$ with their retrieval probabilities.

labels ξ_t, generated from the inference model. However, in unlabeled data such as customer service logs, KBs are often unavailable. This poses a significant challenge to the traditional retriever, which calculates the retrieval probability based on the entire KB by Eq. (2). In contrast, the proposed Entriever can directly calculate the retrieval probability without the need to access the entire KB by Eq. (8). Note that in semi-supervised experiments, since the KB is unavailable for unlabeled data, we only use the non-residual form of Entriever, i.e. Eq. (8). In this setting, the unknown normalizing constant $Z_\theta(c_t, u_t)$ is canceled out, since it appears in both the numerator and denominator of $w(\xi'_t)/w(\tilde{\xi}_t)$ in Eq. (6); and we can calculate the importance weight as follows, for ξ_t generated from the inference model:

$$w(\xi_t) \propto \frac{\exp(-U_\theta(c_t, u_t, \xi_t)) \times p_\theta^{\text{gen}}(r_t|c_t, u_t, \xi_t)}{q_\phi(\xi_t|c_t, u_t, r_t)} \tag{14}$$

The two-stage training of semi-supervised is detailed in Appendix A. The first stage is supervised pre-training of the retrieval model p_θ^{ret}, the generation model p_θ^{gen}, and the inference model q_ϕ on labeled data. In the second stage, the retriever is frozen, and only the generation model p_θ^{gen} and the inference model q_ϕ are further trained on the mix of labeled and unlabeled dialogs.

5 Experiments

5.1 Experiment Settings and Baselines

Experiments are conducted on several dialog datasets: **(1) MobileCS dataset**, a real-life human-human dialog dataset, focuses on mobile customer service, released from the EMNLP 2022 SereTOD Challenge (Ou et al., 2022). MobileCS contains a total of around 100K dialogs. The labeled part was officially split into training/validation/test sets with 8,953/1014/955 dialogs, respectively. The remaining 87,933 dialogs are unlabeled. **(2) CamRest dataset** (Wen et al., 2017) focuses on dialogs in the restaurant domain, consisting of 676 dialogues. Each dialogue contains a KB. The average size of the KB is 22.5 triples. Following previous work, the dataset is split into training/validation/test sets with 406/135/135 dialogs. **(3) In-Car Assistant dataset** (Eric et al., 2017) comprises 3,031 dialogs spanning three domains: weather, navigation, and schedule. The average size of the KB for each dialogue is 62.3 triples. Following previous work, the dataset is split into training/validation/test sets with 2425/302/304 dialogs. **(4) Woz2.1 dataset** (Eric et al., 2020) contains three domains: hotel, attraction and restaurant. The average size of the KB for each dialogue is 54.4 triples. Following (Ding et al., 2024), the dataset is split into training/validation/test sets with 1,839/117/141 dialogs.

For evaluation, we follow the scripts in (Cai et al., 2023) and (Ding et al., 2024). We evaluate the knowledge retrieval ability of Entriever on all four dialog datasets. Three metrics, *Joint Accuracy* (whether the whole knowledge in a dialog turn is accurate or not), *Inform* (whether the retriever provides all the key information for completing a dialog session), and *F1* (the accuracy of the knowledge pieces retrieved), are reported. To evaluate the improvement that Entriever brings to the semi-supervised knowledge-grounded dialog systems, experiments are taken on the MobileCS dataset, as only the MobileCS dataset contains unlabeled data. Two metrics, *Success rate* and *BLEU*, are used to evaluate the quality of the generated responses. *Success rate* measures how often the system is able to provide all the entities and values requested by the user, which is crucial in performing a successful dialog. *BLEU* is used to measure the fluency of the generated responses by analyzing the amount of n-gram overlap between the real responses and the generations. The overall performance of the semi-supervised knowledge-grounded dialog system is measured by *Combined score*, which is *Success* + 2***BLEU*, as in the original SereTOD challenge evaluation scripts (Liu et al., 2022).

For the knowledge retrieval task, we select the most prevalent retriever, the dual-encoder retriever (Karpukhin et al., 2020b), and the most competitive retriever, the cross-encoder retriever (Glass et al., 2022; Cai et al., 2023) (mostly used in the reranking tasks), as our baselines. For the semi-supervised knowledge-grounded dialog systems on MobileCS, several baselines are reported in the experiments. We implement Entriever upon the current state-of-the-art (SOTA) method JSA-KRTOD (Cai et al., 2023).

In our experiments, BERT (Devlin et al., 2019) is used to initialize the retrievers (including the dual-encoder baseline, cross-encoder baseline, and the Entriever) and GPT-2 (Radford et al., 2019) is used to initialize the response generator in the semi-supervised knowledge-grounded dialog systems following previous

Table 1: Results on knowledge retrieval task for the MobileCS, Camrest, In-Car, and Woz2.1 datasets. Joint-acc, Inform, and F1-score are reported. Residual Entrievers are used and trained with different methods (MIS and IS).

Method	MobileCS			Camrest			In-Car			Woz2.1		
	Joint-acc	Inform	F1	Joint-acc	Inform	F1	Joint-acc	Inform	F1	Joint-acc	Inform	F1
Cross-encoder	73.15	35.95	0.589	81.38	63.84	0.816	74.70	42.16	0.870	75.00	32.86	0.508
Entriever (MIS)	76.67	39.81	0.620	**83.17**	68.05	0.824	**78.66**	49.64	**0.875**	**80.24**	43.78	0.524
Entriever (IS)	**77.21**	**42.45**	**0.628**	**83.17**	**68.28**	**0.825**	78.51	**50.53**	**0.875**	79.72	**45.02**	**0.530**

Table 2: Comparison over the MobileCS dataset for different semi-supervision methods (pseudo labeling (PL) and JSA) and whether Entriever is used or not during semi-supervised training. Ratio means the ratio between the number of unlabeled dialogs and the number of labeled dialogs in training. The p-value denotes the significant test result for Combined score. The first colomun of p-value means whether JSA + Entriever outperforms the PL methods, and the second colomun of p-value means whether the JSA + Entriever method significantly improves the JSA method.

Ratio	Method	Success	BLEU-4	Combined	p-value	
1:1	PL	87.5	8.853	105.21		
	JSA	88.0	8.713	105.43	0.025	0.013
	JSA + Entriever	90.6	9.816	**110.23**		
2:1	PL	87.8	9.196	106.19		
	JSA	88.7	9.490	107.68	0.006	0.018
	JSA + Entriever	92.1	9.725	**111.55**		
4:1	PL	88.5	9.341	107.18		
	JSA	90.9	9.398	109.70	0.049	0.088
	JSA + Entriever	92.8	9.554	**111.91**		
9:1	PL	89.4	9.532	108.46		
	JSA	91.8	9.677	111.15	0.083	0.192
	JSA + Entriever	93.0	9.627	**112.25**		

Table 3: Semi-supervised response generation results on the MobileCS dataset. Success, BLEU-4, and Combined score are reported.

Method	Success	BLEU-4	Combined
Baseline (Liu et al., 2022)	31.5	4.170	39.84
Passion (Lu et al., 2022)	43.2	6.790	56.78
TJU-LMC (Yang et al., 2022)	68.9	7.54	83.98
PRIS (Zeng et al., 2022)	78.9	14.51	107.92
JSA-KRTOD (Cai et al., 2023)	91.8	9.677	111.15
JSA-KRTOD+Entriever (ours)	93.0	9.627	**112.25**

settings (Cai et al., 2023; Ding et al., 2024). Hyperparameters are chosen based on the development set, and evaluated on the test set.

5.2 Main Results

The experiments mainly explore the following research questions: **RQ1**: Whether Entriever can improve the knowledge retrieval performance? **RQ2**: Whether introducing Entriever can improve the overall performance of the semi-supervised knowledge-grounded dialog system?

As shown in Table 1, Entriever greatly improves over the cross-encoder (current SOTA method) on all of the Joint-acc, Inform, and F1 metrics across the four datasets. Regardless of the training methods (MIS and IS), Entriever consistently outperforms the strong cross-encoder baseline. Based on these results, we can further discuss the reason for the improvement. The cross-encoder model (Figure 3(a)) models the knowledge pieces in the KB independently given the context. In contrast, Entriever (Figure 3(b)) models the collection of all relevant knowledge pieces given the dialog context. In knowledge-grounded dialog systems, the interconnectivity and interdependence among relevant knowledge fragments are of great importance. Therefore, through the explicit modeling of such interrelationships, the retrieval results produced by Entriever tend to be more accurate as a whole, therefore achieving significantly higher scores on the Joint-acc and Inform metrics. These findings answer **RQ1** and show that Entriever can substantially improve the knowledge retrieval performance.

Table 4: Knowledge retrieval capability on MobileCS for different model architectures and training methods. Joint-acc, Inform, and F1-score are reported.

Setting	Joint-acc	Inform	F1
Dual-encoder (Karpukhin et al., 2020b)	65.60	32.17	0.563
Cross-encoder (Cai et al., 2023)	73.15	35.95	0.589
Entriever (Non-residual, MIS)	76.94	31.89	0.593
Entriever (Non-residual, IS)	72.19	32.22	0.596
Entriever (Residual, MIS)	76.67	39.81	0.620
Entriever (Residual, IS)	**77.21**	**42.45**	**0.628**

Considering the difference the importance sampling (IS) method and Metropolis independence sampling (MIS) method, there is no significant difference between the results. The sample size in IS and the markov steps in MIS are both set to be 12. MIS and IS sampling methods perform equally well. This is similar to the results of using ELMs in rescoring for speech recognition (Liu et al., 2023).

To answer **RQ2**, we conduct experiments on the semi-supervised knowledge-grounded dialog systems with Entriever. To systematically study the effect of the Entriever, different semi-supervised methods and label ratio are explored. As shown in Table 2, the introduction of Entriever substantially improves the overall performance (Combined Score) of the system regardless of the label ratio. Remarkably, the introduction of Entriever can greatly improve the Success rate metric for the dialog systems, indicating that the system's ability to provide the important knowledge is improved. Moreover, the significant test results in Table 2 show that almost in all settings, introducing Entriever can significantly improve the performances (p-value<0.1) over the original JSA method (filter the generated knowledge label with a less accurate knowledge retriever) and the

Table 5: Ablation study on how the number of proposed candidate retrieval results (K) for Entriever to score influences the final test results over MobileCS.

Config	Joint-acc	Inform	Precision	Recall	F1
$K = 4$	76.02	39.33	0.7162	0.5376	0.6142
$K = 8$	76.73	40.70	0.7054	0.5580	0.6231
$K = 16$	**77.21**	42.45	0.6855	0.5789	**0.6277**
$K = 32$	76.79	**42.60**	0.6455	0.6076	0.6260

Table 6: The computational resource overhead when training and testing with different models on the MobileCS dataset. The table reports the time cost for one epoch of training on the training set and the complete inference on the test set, as well as the maximum GPU memory usage. The unit of time in the table is seconds, and the unit of GPU memory usage is megabytes (MB). For all the metrics, the smaller the numerical value, the better. We use 3090 GPUs, and the training batch size is 8 and the batch size for inference is 32.

Model	Training		Inference	
	Time	GPU Memory	Time	GPU Memory
Dual-encoder	905	16356	34.89	3034
Cross-encoder	1469	21740	65.58	3242
Entriever	3803	24538	170.91	3606

pseudo labeling (PL) method (do not filter the generated knowledge label at all). Furthermore, as shown in Table 3, our Entriever improves over the current SOTA semi-supervised method JSA-KRTOD (Cai et al., 2023) in MobileCS. These findings answer **RQ2** and show that introducing Entriever can improve the overall performance of semi-supervised knowledge-grounded dialog systems.

5.3 Analysis and Ablation

To further study the influence of using different architectures and training methods of the retrievers, an ablation study is conducted on the MobileCS dataset to evaluate the knowledge retrieval performance. As shown in Table 4, the cross-encoder architecture greatly outperforms the commonly-used dual-encoder architecture, making it a strong baseline. For Entriever, the results show that the residual form of Entriever greatly improves the stability and performance of the training. Presumably, this is because that the residual form is built upon a trained cross-encoder retriever, which reduces the training burden and improves the training efficiency.

We also conduct an ablation study to explore how the number of proposed candidate retrieval results (K) for Entriever to score will affect the knowledge retrieval results, and the experiment results are shown in Table 5. From Table 5, it can be seen that the test results generally increase with the increase of K, when K is relatively small, presumably because the oracle retrieval result is more likely to be covered. However, although continuously increasing K can increase the possibility of providing the correct knowledge, more noisy samples are introduced as well. Moreover, the computational budgets increase linearly with K. As shown in Table 5, increasing K from 16 to 32 does not improve the performance significantly. Therefore, in experiments related to Entriever, the number of proposed candidate retrieval results during testing is set to $k = 16$.

Moreover, regarding the concern that introducing Entriever into the dialogue system may lead to unacceptable computational overhead, we report the time cost and maximum GPU memory usage during the training and inference with different retrieval models. From the results in Table 6, it can be seen that although the computational overhead of the dual-encoder based model is the smallest, the increase in computational overhead of the other retrievers is acceptable. Especially in scenarios such as vertical domain dialogues where the database size is relatively small, using more computational resources to improve the retrieval performance is preferable.

6 Conclusion

In this work, an energy-based retriever (Entriever) is proposed to collectively model the relevant knowledge pieces from a knowledge base given a context. Entriever can better model the inter-relationship between knowledge pieces, and can substantially improve the knowledge retrieval performance in knowledge-grounded dialog systems. Moreover, we conduct an in-depth exploration of various architectures of energy functions and training methods for Entriever and find out that using the residual form can improve the quality of the retrieval results. Furthermore, in semi-supervised training of knowledge-grounded dialog systems, Entriever enables effective scoring of retrieved knowledge pieces, and leads to significant improvement in the end-to-end performance of dialog systems. The above results show that Entriever has great potential for developing advanced knowledge-grounded dialog systems. We open-source the code and data to facilitate reproducibility and encourage further exploration in this direction.

7 Limitations

In this work, training Entrievers with maximum likelihood estimate (MLE) methods is explored. However, in previous works, noise contrastive estimate (NCE) (Gutmann and Hyvärinen, 2010) methods have also been used to train energy-based language models. Therefore, training Entrievers with NCE methods can be studied in future works and compared with the MLE methods explored in this work.

In addition, recent studies have explored using large language models (LLMs) for knowledge retrieval and reranking tasks. However, in this work, Entriever is implemented with relatively small models (BERT). Therefore, conducting experiments on Entriever with larger backbone models and studying the scaling effects of Entriever can be further explored.

References

Anton Bakhtin, Yuntian Deng, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. 2021. Residual energy-based models for text. *The Journal of Machine Learning Research*, 22(1):1840–1880.

Yucheng Cai, Si Chen, Yuxuan Wu, Yi Huang, Junlan Feng, and Zhijian Ou. 2024. The 2nd futuredial challenge: Dialog systems with retrieval augmented generation (futuredial-rag). In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1091–1098. IEEE.

Yucheng Cai, Hong Liu, Zhijian Ou, Yi Huang, and Junlan Feng. 2022. Advancing semi-supervised task oriented dialog systems by JSA learning of discrete latent variable models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 456–467.

Yucheng Cai, Hong Liu, Zhijian Ou, Yi Huang, and Junlan Feng. 2023. Knowledge-retrieval task-oriented dialog systems with semi-supervision. In *INTERSPEECH*.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 285–294.

Jingcheng Deng, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. Regavae: A retrieval-augmented gaussian mixture variational auto-encoder for language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2500–2510.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Zeyuan Ding, Zhihao Yang, Ling Luo, Yuanyuan Sun, and Hongfei Lin. 2024. From retrieval to generation: A simple and unified generative model for end-to-end task-oriented dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17907–17914.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Kumar Goyal, Peter Ku, and Dilek Hakkani-Tür. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *LREC*.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Silin Gao, Zhijian Ou, Wei Yang, and Huifang Xu. 2020. Integrating discrete and neural features via mixed-feature trans-dimensional random field language models. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938.

Tianxing He, Bryan McCann, Caiming Xiong, and Ehsan Hosseini-Asl. 2021. Joint energy-based model training for better calibrated natural language understanding models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1754–1761.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022b. Few-shot learning with retrieval augmented language models. *arXiv e-prints*, pages arXiv–2208.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020b. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, and Guido Zuccon. 2024. Leveraging llms for unsupervised dense retriever ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1307–1317.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218.

Hong Liu, Zhaobiao Lv, Zhijian Ou, Wenbo Zhao, and Qing Xiao. 2023. Exploring energy-based language models with different architectures and training methods for speech recognition. *arXiv preprint arXiv:2305.12676*.

Hong Liu, Hao Peng, Zhijian Ou, Juanzi Li, Yi Huang, and Junlan Feng. 2022. Information extraction and human-robot dialogue towards real-life tasks: A baseline study with the mobilecs dataset. In *EMNLP 2022 SereTOD Workshop*.

Jun S Liu. 2001. *Monte Carlo strategies in scientific computing*. Springer.

Weichen Lu, Yifei Wang, Weizhen Zhao, Buxian Chen, Xiaojie Chen, Jinsong Pan, Wentao Liang, and Yongquan Lai. 2022. Team passion at seretod-emnlp 2022: End-to-end task-oriented dialog system with improved prompting scheme.

Zhijian Ou, Junlan Feng, Juanzi Li, Yakun Li, Hong Liu, Hao Peng, Yi Huang, and Jiangjiang Zhao. 2022. A challenge on semi-supervised and reinforced task-oriented dialog systems. *arXiv preprint arXiv:2207.02657*.

Zhijian Ou and Yunfu Song. 2020. Joint stochastic approximation and its application to learning discrete latent variable models. In *Conference on Uncertainty in Artificial Intelligence*, pages 929–938. PMLR.

Zhijian Ou et al. 2024. Energy-based models with applications to speech and language processing. *Foundations and Trends® in Signal Processing*, 18(1-2):1–199.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2021. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. In *International Conference on Learning Representations*.

Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. 2019. Global autoregressive models for data-efficient sequence learning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 900–909.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15933–15946.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Bin Wang and Zhijian Ou. 2017. Language modeling with neural trans-dimensional random fields. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 294–300. IEEE.

Bin Wang and Zhijian Ou. 2018a. Improved training of neural trans-dimensional random field language models with dynamic noise-contrastive estimation. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 70–76.

Bin Wang and Zhijian Ou. 2018b. Learning neural trans-dimensional random field language models with noise-contrastive estimation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Bin Wang, Zhijian Ou, and Zhiqiang Tan. 2015. Trans-dimensional random fields for language modeling. In *Proceedings of the 53rd Annual Meeting of the*

Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 785–794.

Bin Wang, Zhijian Ou, and Zhiqiang Tan. 2017. Learning trans-dimensional random fields with applications to language modeling. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):876–890.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548.

Zhitong Yang, Xing Ma, Anqi Liu, and Zheyu Zhang. 2022. Discovering customer-service dialog system with semi-supervised learning and coarse-to-fine intent detection. In *EMNLP 2022 SereTOD Workshop*.

Weihao Zeng, Keqing He, Zechen Wang, Dayuan Fu, Guanting Dong, Ruotong Geng, Pei Wang, Jingang Wang, Chaobo Sun, Wei Wu, et al. 2022. Semi-supervised knowledge-grounded pre-training for task-oriented dialog systems. In *EMNLP 2022 SereTOD Workshop*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

A The Details of the JSA Algorithm in Semi-Supervised Dialog Systems

The detailed procedure of semi-supervised training using the JSA algorithm is summarized in Algorithm 2, which consists of two stages. First, supervised pre-training is conducted on the retrieval model p_θ^{ret}, the generation model p_θ^{gen}, and the inference model q_ϕ on labeled data. After supervised pre-training, the retrieval parameters θ^{ret} are frozen in the second stage of training over unlabeled data. Note that in unlabeled data, the knowledge pieces used in the dialogs are not annotated and the knowledge base (KB) is often not available. This presents a significant challenge for the training of both the traditional retriever and the Entriever over unlabeled data. Investigating the training of the retrieval parameters θ^{ret} over unlabeled data and unavailable KB is interesting future work.

In the second stage, supervised and unsupervised mini-batches are randomly drawn from labeled and unlabeled data. For labeled dialogs, the latent knowledge ξ_t are given. For unlabeled dialogs, we apply the recursive turn-level MIS sampler based on Eq. (6) to sample the latent states ξ_t and treat them as if being given. The gradient calculation and parameter updating are then the same for the labeled and unlabeled dialogs.

Algorithm 2 JSA algorithm for training semi-supervised dialog systems

Require: A mix of labeled and unlabeled dialogs.

Run supervised pre-training of $\theta = (\theta^{\text{ret}}, \theta^{\text{gen}})$ and ϕ on labeled dialogs;
Frozen the retriever parameters θ^{ret};
repeat
  Draw a dialog $(u_{1:T}, r_{1:T})$;
  if $(u_{1:T}, r_{1:T})$ is not labeled **then**
    Generate $\xi_{1:T}$ using the recursive turn-level MIS sampler;
  end if
  $J_{\theta^{\text{gen}}} = 0, J_\phi = 0$;
  for $i = 1, \cdots, T$ **do**
    $J_{\theta^{\text{gen}}} += \log p_\theta^{\text{gen}}(r_t \mid c_t, u_t, \xi_t)$;
    $J_\phi += \log q_\phi(\xi_t \mid c_t, u_t, r_t)$;
  end for
  Update θ^{gen} by ascending: $\nabla_{\theta^{\text{gen}}} J_{\theta^{\text{gen}}}$;
  Update ϕ by ascending: $\nabla_\phi J_\phi$;
until convergence
return θ and ϕ