DIALKI: Knowledge Identification in Conversational Systems through Dialogue-Document Contextualization

Zeqiu Wu ♠* **Bo-Ru Lu** ♠* **Hannaneh Hajishirzi** ♠◇ **Mari Ostendorf** ♠

♠University of Washington   ◇Allen Institute for AI

{zeqiuwu1,roylu,hannaneh,ostendor}@washington.edu

Abstract

Identifying relevant knowledge to be used in conversational systems that are grounded in long documents is critical to effective response generation. We introduce a knowledge identification model that leverages the document structure to provide dialogue-contextualized passage encodings and better locate knowledge relevant to the conversation. An auxiliary loss captures the history of dialogue-document connections. We demonstrate the effectiveness of our model on two document-grounded conversational datasets and provide analyses showing generalization to unseen documents and long dialogue contexts.

1 Introduction

Many conversational agent scenarios require knowledge-grounded response generation, where knowledge is represented in written documents. Most prior work has explored architectures for knowledge grounding in an end-to-end framework, optimizing a loss function targeting response generation (Ghazvininejad et al., 2018; Zhou et al., 2018; Yavuz et al., 2019). However, the knowledge needed at any one turn in the dialogue is typically localized in the document, and some studies have shown that directly optimizing for knowledge extraction helps resolve complex user queries (Feng et al., 2020) and increases user engagement in social chat (Dinan et al., 2019; Moghe et al., 2018). For long documents, explicit knowledge identification can also be useful for model interpretability and human-in-the-loop assistant scenarios.

Following (Feng et al., 2020), we define *knowledge identification* as the task of locating knowledge in a long document that is relevant to the current user query given the conversation context (Figure 1). Knowledge identification is similar to open question answering (Chen et al., 2017; Min
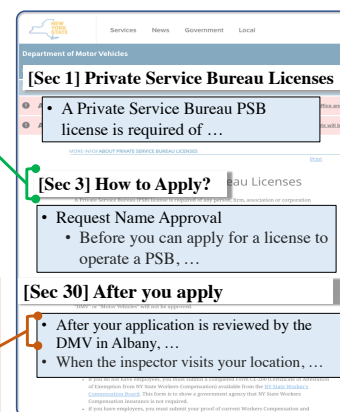
Dialogue Context   *Grounding document*

Figure 1: In a document-grounded conversation, *knowledge identification* targets to locate a knowledge string within a long document to assist the agent in addressing the current user query.

et al., 2019), the task of answering a factoid question given a large grounding, except that it is not an interactive setting like dialogues. With the assumption of a long grounding document, our task differs from prior work in conversational question answering (Choi et al., 2018; Reddy et al., 2019), which focuses on answering a sequence of factoid questions about a short text snippet. Additionally, real user information needs can involve conversations with diverse forms of user queries and dialogue acts (e.g., asking for user preference, etc) as shown in Figure 1. Previous work in knowledge identification encode the grounding document as a single string (Feng et al., 2020), or splitting it into isolated sentences (Dinan et al., 2019; Kim et al., 2020; Zheng et al., 2020) which potentially loses important discourse context.

In this paper, we introduce DIALKI to address knowledge identification in conversational systems with long grounding documents. In contrast to previous work, DIALKI extends multi-passage reader models in open question answering (Karpukhin et al., 2020; Cheng et al., 2020) to obtain dense

---
*Equal contribution

encodings of different spans in multiple passages in the grounding document, and it contextualizes them with the dialogue history. Specifically, DI-ALKI extracts knowledge given a long document by dividing it into paragraphs or sections and individually contextualizes them with dialogue context. It then extracts knowledge by first selecting the most relevant passage to the dialogue context and then selecting the final knowledge string within the selected passage. Processing each passage rather than the full document greatly shortens the knowledge context, while preserving enough discourse context for reasoning. DIALKI also uses a multi-task objective to identify knowledge for the next turn, as well as used knowledge for previous turns that helps improve the learning of both dialogue and document representations by capturing interdependencies between the next agent utterance, previous utterances, and the grounding document.

Our model significantly improves over baselines on two conversational datasets, particularly on previously unseen documents or topics. Analyses show good generalization to longer grounding documents and longer dialogue context, as well as improvements in response generation. Model robustness can be further improved with an f-divergence based posterior regularization.

Our contributions are summarized as follows. First, we propose a *knowledge identification* model to address the problem of locating relevant information from long documents in a conversational context. Second, we introduce a multi-task learning framework that models the dialogue-document interactions via an auxiliary task of history knowledge prediction and a knowledge contextualization mechanism. Lastly, our model advances the state of the art in knowledge identification tasks for two conversational datasets, with more than 60% and 20% gains over previous work.[1]

2 Related Work

Conversational Question Answering. Existing conversational question answering tasks (Choi et al., 2018; Reddy et al., 2019) are generally defined as the task of reading a short text passage and answering a series of interconnected questions in a conversation. ShARC (Saeidi et al., 2018) is a conversational machine reading dataset to address under-specified questions by requiring agents to

ask follow-up questions grounded in a short text snippet that are answerable with "yes/no" answers. These datasets focus on restricted dialogue act types (questions and answers) and short grounding texts. In contrast, our work focuses on more natural conversations that have a wider range of dialogue act types and are grounded in longer documents. In principle, models developed for short contexts could be applied to longer documents, but most have been based on pretrained language models (e.g., BERT (Devlin et al., 2019) or RoBerTa (Liu et al., 2019)) with an input that is the concatenation of a short document and dialogue context. Such language models usually have a limited maximum input length which constrains the implementation for long documents.

Knowledge-Grounded Dialogues. Most previous work in knowledge grounded dialogues (Ghazvininejad et al., 2018; Zhou et al., 2018; Zhao et al., 2020; Lin et al., 2020; Li et al., 2019) generally focus on optimizing a loss function that targets response generation. Our work focuses on the knowledge identification task instead of response generation in a document-grounded dialogue.

To the best of our knowledge, Wizard of Wikipedia (WoW) (Dinan et al., 2019), HollE (Moghe et al., 2018) and Doc2Dial (Feng et al., 2020) are the only conversational datasets that include the task of knowledge identification in long documents. Both WoW and HollE focus on social chat conversations. WoW covers a much wider range of open-domain topics, while HollE is restricted to movie discussions. Doc2Dial is a more recent goal-oriented conversational dataset in various social welfare domains.

Knowledge Identification Models for Dialogues. Only a few works (Dinan et al., 2019; Lian et al., 2019; Feng et al., 2020; Kim et al., 2020; Zheng et al., 2020) build models for and evaluate on the knowledge identification task. Feng et al. (2020) encodes each long document as a single string without leveraging document structures, while Dinan et al. (2019); Lian et al. (2019); Kim et al. (2020); Zheng et al. (2020) separately encode sentences in documents, which may have strong contextual dependencies among each other. Our model leverages document structures and divides each document into multiple passages to process. Similar to our model, Kim et al. (2020); Zheng et al. (2020) incorporate previously used knowledge, but they

use a single vector to sequentially track the state of the used knowledge. Instead, we apply a multi-task learning framework to model relations between grounding documents and history turns.

3 Method

Problem Definition   *Knowledge identification* in a document-grounded dialogue is defined as follows. At inference, given the dialogue context consisting of a sequence of n previous utterances $(u_1, u_2, \ldots, u_n)$ and a grounding document $\mathcal{D}$, select a substring y in $\mathcal{D}$ that is relevant to the dialogue context and will be used in the next turn (i.e., utterance) in a conversation. We denote u_1 as the last (user) turn and u_n as the earliest turn in dialogue history to help our model explanation.

Each document consists of a sequence of passages $\mathcal{D} = \{p_1, p_2, \ldots, p_{|\mathcal{D}|}\}$ based on paragraphs or sections. Each passage p consists of a sequence of semantic units $p = (s_1, s_2, \ldots, s_l)$, where each semantic unit (SU) can be either a token or a span or a sentence depending on how the document is segmented. For simplicity, we use "span" as the semantic unit in this section to describe our model.

Method overview   In this section, we introduce DIALKI, a multi-task learning model for knowledge identification as illustrated in Figure 2. We first introduce how we obtain dialogue utterance and knowledge span representations from BERT (Devlin et al., 2019) and a span-level knowledge contextualization mechanism (Section 3.1). These representations are then used for knowledge identification in our multi-task learning framework, which includes the main task of next-turn knowledge identification and an auxiliary task of history knowledge prediction applied during training only (Section 3.2). Finally, we describe our joint training objective and inference details (Section 3.3).

3.1 Encoding Dialog Context and Knowledge

3.1.1 Multi-Passage Encoding

Here, we describe how we initially obtain dense vector representations of each passage in the grounding document as a set of span representations. Inspired by the recent open-domain question answering multi-passage reader models (Karpukhin et al., 2020; Cheng et al., 2020), we concatenate the dialogue context $(u_1, u_2, \ldots, u_n)$ with u_1 to be the latest user turn, the document title t and each passage p, and use a pretrained language model like BERT (Devlin et al., 2019) to encode

the concatenated sequence. More formally, the model input $\mathbf{X}$ for a passage p of length l becomes:

$$
\begin{aligned}
\mathbf{X} \quad = \quad & [\text{cls}][\text{usr}]\, u_1 [\text{agt}] u_2 \cdots [\text{usr}]\, u_n \\
& [\text{sep}]\, t\, [\text{cls}]\, s_1\, [\text{cls}]\, s_2 \cdots [\text{cls}]\, s_l\, [\text{sep}]
\end{aligned}
$$

where '[usr]' and '[agt]' are special tokens indicating the start of a user or agent turn. '[cls]' indicates the start of the whole sequence or each knowledge span. '[sep]' are separator tokens. Then we encode $\mathbf{X}$ and gather a sequence of pooled output vectors $\mathbf{H} = G(\text{BERT}(\mathbf{X}))$ where $G(.)$ gathers vectors of all '[cls]', '[usr]' and '[agt]' tokens. We decompose $\mathbf{H}$ as $[\mathbf{z}, \mathbf{u}_1, \ldots, \mathbf{u}_n, \mathbf{s}_1, \ldots, \mathbf{s}_l]$ where $\mathbf{z}, \mathbf{u}_i, \mathbf{s}_j$ are pooled representations of the whole input (the first '[cls]' token in $\mathbf{X}$), dialogue utterance u_i and span s_j respectively.

3.1.2 Knowledge Contextualization

We leverage the pooled global, utterance and span representations $\mathbf{z}, \mathbf{u}_i, \mathbf{s}_j$ obtained above for each passage p to further contextualize knowledge span representations. Inspired by how EntNet (Henaff et al., 2017) updates each entity representation based on an input sequence, for each span s_j we calculate an updated span embedding $\widehat{\mathbf{s}}_j$ contextualized with the sequence of previous *user* utterance representations as below. We use a gating function g which determines how much the span embedding should be updated based on $\mathbf{u}_i$ and $\mathbf{z}$.

$$
\begin{aligned}
\mathbf{a}_{i,j} &= \mathbf{W}_s\, \mathbf{s}_j + \mathbf{W}_z\, \mathbf{z} + \mathbf{W}_u\, \mathbf{u}_i\, , \; i \in \mathbf{C}_u \\
\widehat{\mathbf{s}}_j &= \upsilon\Big(\sum_{i \in \mathbf{C}_u} \big[\phi(\mathbf{a}_{i,j}) \odot g_{i,j} \big] + \mathbf{s}_j \Big) \\
g_{i,j} &= \sigma\Big(\mathbf{u}_i^\top \mathbf{z} + \mathbf{u}_i^\top \mathbf{s}_j \Big),
\end{aligned}
$$

where $\mathbf{W}_s, \mathbf{W}_z, \mathbf{W}_u \in \mathbb{R}^{d \times d}$ are model parameters, and $\mathbf{C}_u$ indexes the most recent user turns. $\upsilon(.)$ is the vector normalization operation. σ is the sigmoid function and ϕ can be any non-linear activation function. In our model, we use ReLU for ϕ. Similarly, we calculate $\widetilde{\mathbf{s}}_j$ with previous *agent* turns. Then new span embedding denotes as $\dot{\mathbf{s}}_j = [\mathbf{s}_j, \widehat{\mathbf{s}}_j, \widetilde{\mathbf{s}}_j]$. In our experiments, we set $\mathbf{C}_u$ to contain the most recent 2 user turns only, which leads to the best results.

3.2 DIALKI with Multi-Task Learning

3.2.1 Next-Turn Knowledge Identification

The main task in DIALKI is next-turn knowledge identification consisting of two parts: passage
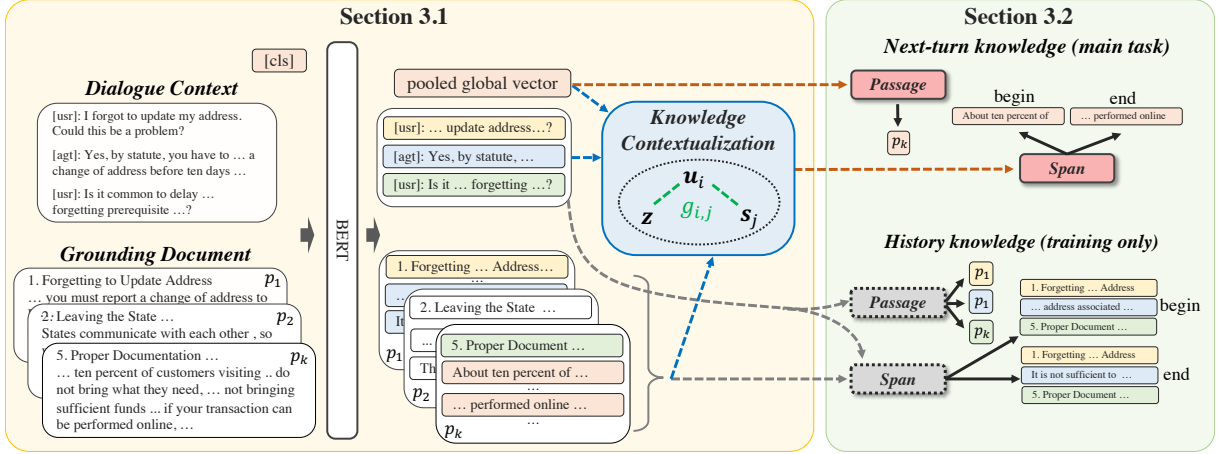
Figure 2: The overview of DIALKI. Each document is divided into passages. We apply BERT and a knowledge contextualization mechanism to obtain dialogue context and knowledge representations (left), for performing both next (main) and history (auxiliary) turn knowledge identification tasks (right). For each turn, DIALKI identifies knowledge by selecting the relevant passage as well as the begin/end spans in the passage.

prediction and knowledge span prediction (upper right in Figure 2). At inference, given a set of passages $\{p_1, p_2, \ldots, p_{|\mathcal{D}|}\}$ and the dialogue context $(u_1, u_2, \ldots, u_n)$, our multi-passage knowledge identification model targets to predict the most relevant passage p_k based on the softmax probability in Eq. (1), as well as the most relevant knowledge string $y = (s_b \ldots s_e)$ in p_k to the next agent turn in the conversation, based on the begin and end span probabilities in Eq. (2-3)

We obtain the pooled global vector $\mathbf{z}$, each utterance and span representation $\mathbf{u}_i, \dot{\mathbf{s}}_j$ for each passage p described in Section 3.1. We denote $\mathbf{Z}$ as the matrix containing the pooled global vectors for all passages and $\mathbf{U}_i$ as utterance representations for u_i in all passages. $\dot{\mathbf{S}}$ denote the matrix with all span representations after knowledge contextualization.

Training Objectives Eq. (1-3) show objective functions of knowledge passage $\mathcal{L}_{\text{psg}}$, begin $\mathcal{L}_{\text{begin}}$ and end $\mathcal{L}_{\text{end}}$ span predictions. $q(.)_t$ denotes the t-th index of the vector resulting from the softmax function. The variables $\hat{k}$, $\hat{b}$ and $\hat{e}$ correspond to the gold passage, begin and end span indices.

$$\mathcal{L}_{\text{psg}} = -\log q(\mathbf{W}_p \mathbf{Z})_{\hat{k}} \quad (1)$$

$$\mathcal{L}_{\text{begin}} = -\log q(\mathbf{W}_b \dot{\mathbf{S}})_{\hat{b}} \quad (2)$$

$$\mathcal{L}_{\text{end}} = -\log q(\mathbf{W}_e \dot{\mathbf{S}})_{\hat{e}} \quad (3)$$

where $\mathbf{W}_p, \mathbf{W}_b, \mathbf{W}_e \in \mathbb{R}^d$. Therefore, the combined next turn knowledge selection loss function becomes:

$$\mathcal{L}_{\text{next}} = \mathcal{L}_{\text{psg}} + \mathcal{L}_{\text{begin}} + \mathcal{L}_{\text{end}}.$$

3.2.2 History Knowledge Identification

The auxiliary task during training is to predict previously used knowledge, with the intuition that used knowledge in documents by history turns would guide the search for the next knowledge to use.

Training Objectives With the same representations used in next-turn knowledge identification, similar to Eq. (1-3), we calculate both passage- and span-level prediction losses for each history turn u_i if the knowledge string for u_i can be found in $\mathcal{D}$. We calculate the passage prediction loss $\mathcal{L}_{\text{psg}}^h$ of previous turns as follows:

$$\mathcal{L}_{\text{psg}}^h = \frac{1}{\|\mathbf{U}^*\|} \sum_{u_i \in \mathbf{U}^*} -\log q\big(\mathbf{W}_p^h \; \phi(\mathbf{W}^h \, \mathbf{U}_i)\big)_{k_i}$$

where $\mathbf{W}^h \in \mathbb{R}^{d \times d}, \mathbf{W}_p^h \in \mathbb{R}^d$ are model parameters. $\mathbf{U}^*$ is the set of history turns that can find their knowldge strings in the document $\mathcal{D}$. k_i is the gold passage index for turn u_i. ϕ is a non-linear activation function, with ReLU used in our model.

Similarly, we calculate the average losses of predicting the gold begin $\mathcal{L}_{\text{begin}}^h$ and end $\mathcal{L}_{\text{end}}^h$ knowledge spans of all history turns in $\mathbf{U}^*$ in their gold passages. To do so, for each u_i, we compute the dot product between a linearly transformed $\mathbf{u}_i$ and each span embedding $\mathbf{s}_j$ in p_{k_i} and apply a cross-entropy loss for the begin or end span prediction. Therefore, the combined knowledge selection loss function of history turns becomes:

$$\mathcal{L}_{\text{hist}} = \mathcal{L}_{\text{psg}}^h + \mathcal{L}_{\text{begin}}^h + \mathcal{L}_{\text{end}}^h.$$

3.3 Training and Inference

Posterior Regularization   During training, we incorporate a posterior regularization mechanism (Cheng et al., 2021) to enhance the model's robustness to domain shift. Specifically, we add an additional adversarial training loss as below. Div is some f-divergence.[2] Let x be the encoded **X** (defined in Section 3.1.1) after the BERT word embedding layer, DIALKI outputs $f_{psg}(x)$, $f_{begin}(x)$ and $f_{end}(x)$ as the next turn passage, begin and end knowledge span logits (the results before the softmax function $q(.)$ in Eq. (1-3)) respectively.

$$\mathcal{L}_{\text{adv}} = \max_{\|\epsilon\| \leq a} \sum_{f \in \{f_{psg}, f_{begin}, f_{end}\}} \text{Div}\big(f(x) \| f(x+\epsilon)\big)$$

The above loss function essentially regularizes the g-based worst-case posterior difference between the clean and noisy input (with norm of the added noise no larger than some scalar a) using an inner loop to search for the most adversarial direction.

Joint Training Objective   Combining all the above components, our final model optimizes the joint objective $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{next}} + \alpha\mathcal{L}_{\text{hist}} + \beta\mathcal{L}_{\text{adv}} \tag{4}$$

where α and β are tunable hyperparameters.

Inference   During inference, we perform next-turn knowledge prediction only. We first predict the passage with the highest probability and enumerate over all possible knowledge span sequences in the selected passage with a maximum length. Then we select the span sequence with the highest score to be the final knowledge string. The score of each span sequence is calculated as the sum of begin and end span probabilities.

4 Experiment

4.1 Datasets and Evaluation Metrics

Datasets   We use two datasets for our experiments: *Doc2Dial* (Feng et al., 2020) and *Wizard of Wikipedia (WoW)* (Dinan et al., 2019). Doc2Dial contains about 4.8k goal-oriented dialogues in 4 social-welfare domains, with an additional Covid-19 domain in the blind held-out test set. Each dialogue has an average of 14 turns, grounded on a long document with more than 1k tokens on average. Each user or agent turn is grounded in a

sequence of knowledge spans as annotated in the dataset.[3] WoW contains over 20k social chat conversations with an average of 9 turns on over 1k open-domain topics. For each agent turn, the agent (i.e., wizard) chose one or no grounding sentence from on average 7 Wikipedia passages provided by a pre-defined retriever based on the dialogue history for composing the response. Each passage contains 10 sentences. The original data has its dev/test set split to two subsets, which contain conversations about topics seen or unseen in training.

Evaluation Metrics   For evaluation, we use exact match (EM) and token-level F1 score as originally used in Feng et al. (2020); Dinan et al. (2019).

4.2 Implementation Details

Doc2Dial   Documents are automatically split into passages by parsing the html files into smaller sections with different titles, resulting in an average of 8 passages per document. All passages from a single html file are used throughout each conversation. Knowledge SUs are spans as segmented in the data. During inference, we set the maximum knowledge length to be 5 spans based on the dev set distribution.[4]

WoW   Passage segmentation is automated in the WoW pre-defined retriever and the dataset provides 7 passages on average for each agent turn in the dataset. Since agents are allowed to select no grounding sentence, we add an additional passage with only one single sentence "no passages used" following the original data processing script[5]. Knowledge SUs are sentences. We set the maximum knowledge length to be 1 sentence during inference. Passages may differ for each agent turn in the same conversation. Thus, during training, we only calculate history loss for previous agent turns whose ground truth knowledge can be found in next turn passages, which on average cover over 70% of history agent turns.

Experimental Setup   We initialize and finetune on BERT (Devlin et al., 2019). We use the uncased base BERT in most of our experiments, and

---

[2]We use Jensen-Shannon divergence in all experiments.

[3]Note that the knowledge identification task only targets to predict grounded knowledge for agent turns.

[4]We use data, baselines, preprocessing & evaluation scripts at: `https://github.com/doc2dial/sharedtask-dialdoc2021`

[5]`https://github.com/facebookresearch/ParlAI`

set $3e^{-5}$ as the learning rates and 1000 as warm-up steps. For each experiment, we search the weights in Eq. (4) on the dev set in the ranges of $\alpha = \{0.5, 1, 2\}$, $\beta = \{0.5, 2.5, 5\}$. We do not observe much difference with different weight combinations, but the best result is achieved when $\beta = 5$ for both datasets, and $\alpha = 1$ for Doc2Dial and $\alpha = 0.5$ for WoW. Models are selected based on the best dev set EM score. The maximum length of dialogue context is 128. The maximum lengths of model input are 512 and 384 for Doc2Dial and WoW respectively. In training, we provide multiple passages from the grounding document as the input, where only one of them is the gold passage. We find that learning benefits from having more negative passage examples, and the number used is constrained by memory consumption (up to 10 for the models with posterior regularization and up to 20 otherwise). For inference, up to 20 passages from the target document are considered. For longer documents, the first 20 passages are used. Further details are in Appendix A.

4.3 Compared Systems

BERTQA-Token: The original baseline (Feng et al., 2020) and the best published model on Doc2Dial. It uses BERTQA (Devlin et al., 2019) with each dialogue context as the question and sliding windows to process each document, and predicts the start and end tokens in the document.

BERTQA-Span: Similar to BERTQA-Token, but predicts the start and end knowledge spans instead of tokens. Instead of using sliding windows, we increase the number of position embeddings to be 2048, initialized with 512 position embeddings in BERT repeated 4 times, following (Beltagy et al., 2020). We observe better results with this operation than when using sliding windows.

Transformer MemNet: The original baseline (Dinan et al., 2019) of WoW, which uses a vanilla Transformer (Vaswani et al., 2017) to encode all knowledge sentences separately and a memory network for sentence selection. Another model version includes pre-training on Reddit conversations.

SLKS: The state-of-the-art model (Kim et al., 2020) on WoW that encodes all knowledge sentences and dialogue turns separately with BERT (or RNN). It uses two GRUs to update the states of dialog history and previously selected sentences.

Method	Overall	
	EM	F1
BERTQA-Token	34.6	53.2
BERTQA-Token (our version)	35.8	52.6
DIALKI ($\mathcal{L}_{next}$ only)	51.2	64.7
DIALKI	59.5	71.0
DIALKI (BERT-large)	**61.8**	**73.1**

Table 1: Evaluation results on the Doc2Dial test set.

DiffKS: This model (Zheng et al., 2020) is similar to SLKS. Additionally, it computes the representation difference between each candidate knowledge sentence and the state of previously used knowledge for in the final decision function.

Multi-Sentence: This baseline is designed to be similar to DIALKI, but divides documents into sentences instead of passages. It calculates the next knowledge prediction loss $\mathcal{L}_{next}$ only. For Doc2Dial, we use subsections, mostly single sentences, as segmented in documents. Knowledge strings rarely exceed the subsection boundaries.

DIALKI (Ours): Our multi-passage knowledge identification model with the next turn knowledge prediction loss $\mathcal{L}_{next}$, history knowledge prediction loss $\mathcal{L}_{hist}$, contextualization mechanism (*know-ctx*) and posterior regularization loss $\mathcal{L}_{adv}$.

4.4 Quantitative Results

Doc2Dial Table 1 reports the results of different systems in the blind held-out test set with an unseen Covid-19 domain. All models are based on the BERT-base model except the last one that uses BERT-large. The full model of DIALKI achieves best results, demonstrating the effectiveness of combining all components of the system described in Section 3. The significant improvement from DIALKI ($\mathcal{L}_{next}$ only) over BERTQA-Token, which takes the full document as a single string, shows the benefit of our multi-passage framework. BERT-large helps further improve the overall results.[6]

WoW Results on both test sets are presented in Table 2, containing conversations on seen and unseen topics in training. DIALKI significantly outperforms all other systems, which encode knowledge sentences disjointly. This again confirms the

---

[6]After ensemble with other large language models of RoBerTa and ELECTRA (Liu et al., 2019; Clark et al., 2020), our model achieves EM / F1 scores as 67.09 / 76.34, achieving the best scoring system on the leaderboard outperforming beating the second-best scores 63.53 / 75.94.

Method	Seen		Unseen	
	EM	F1	EM	F1
Transformer MemNet	22.5	33.2	12.2	19.8
Transformer MemNet + Pretrain	24.5	36.4	23.7	35.8
DiffKS (RNN)	25.5	–	19.7	–
SLKS (RNN)	23.4	–	14.7	–
SLKS (BERT-base)	26.8	–	18.3	–
Multi-Sentence (BERT-base)	30.4	37.7	27.6	35.4
DIALKI (BERT-base)	**32.9**	**40.7**	**35.5**	**43.4**

Table 2: Evaluation results of WoW test sets.

advantage of our multi-passage framework and the modeling of dialogue-document relations.

Surprisingly, DIALKI achieves even higher results on the unseen test set while others observe performance drops. One potential reason is that blindly dividing passages into disjoint sentences to process may hurt the model's reasoning ability and generalization. In addition, Transformer Mem-Net, DiffKS and SLKS decouple the encoding of dialogue history and grounding sentences, which prevents the model to effectively reason over their relations. With more investigation into the data, we found that two thirds of the grounding passages used in those conversations about unseen topics actually appear in the training set based on title matching.[7] Moreover, all topics in the dataset are similar lifestyle topics collected from persona description sentences in (Zhang et al., 2018). These observations support the possibility of better performance on the original unseen set.

Impact of different components of DIALKI. Table 3 reports results of ablating different components of our system on the dev set of both datasets. Although Doc2Dial does not provide separate seen and unseen sets as WoW does, we split the dev set into examples that have grounding documents seen or unseen in the training set. Note that "seen" and "unseen" refer to documents and topics for Doc2Dial and WoW respectively.

We observe that DIALKI consistently beat baseline models that either process the full document as a single string or isolated sentences. Our framework leads a to smaller performance gap between seen and unseen examples. Adding the auxiliary history knowledge prediction loss ($\mathcal{L}_{hist}$) leads to further improvements on both datasets. Adding *know-ctx* helps enhance the performance on Doc2Dial while does not appear to be effec-
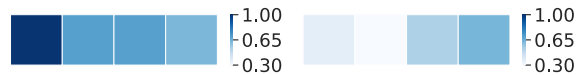

Figure 3: Similarity between $\mathbf{z}$ and the latest 4 history turn representations (i.e., $\mathbf{u}_1 \ldots \mathbf{u}_4$ from left to right) on Doc2Dial (left) and WoW (right).

tive on WoW, as explored further below. Adding posterior regularization ($\mathcal{L}_{adv}$) is effective on both datasets while Doc2Dial gets more advantage from it especially on the unseen subset. Combining all model components yields the best results.

4.5 Analysis

Impact on Response Generation We apply BART (Lewis et al., 2020) to decode agent responses given the concatenated dialogue context and grounding knowledge (e.g., document or predicted knowledge string) as the input. BART is also used as the baseline for the agent response generation task on Doc2Dial (Feng et al., 2020),[8] where the model is given the dialogue history concatenated with the full document to decode the next agent response. We conduct experiments on the same model architecture, with the knowledge input being the predicted knowledge string or passage. Without changing the model at all, using knowledge predicted by DIALKI leads to almost 3 points in the sacrebleu score (Post, 2018), as shown in Table 4. Examples of generated responses are shown in Table 5.

Passage Identification Accuracy We map predicted knowledge strings back to the passages and calculate the passage-level accuracy. Table 6 shows that DIALKI outperforms baseline models in locating the passage containing the knowledge string. Notably, our models generalize well in passage prediction to unseen documents or dialogue topics.

Similarity Between Global and History Turn Representations The dot product between $\mathbf{z}$ (the encoding of the whole input sequence) and each history utterance representation $\mathbf{u}_i$ (sigmoid normalized) is used in the gating function g (defined in Section 3.1.2) that gates the effect of each utterance u_i in calculating span embeddings. Figure 3 shows such normalized dot product scores between $\mathbf{z}$ and the latest four history turn vectors. In Doc2Dial,

---

[7]Only 1.6% of agent turns in the WoW unseen (topic) test set have all of their retrieved passages not seen in training.

[8]https://github.com/doc2dial/sharedtask-dialdoc2021

Method	Doc2Dial						WoW					
	Overall		Seen		Unseen		Overall		Seen		Unseen	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
BERTQA-Token	42.2	58.1	48.3	61.1	37.0	55.6	–	–	–	–	–	–
BERTQA-Span	46.3	59.3	54.4	63.5	39.4	55.6	–	–	–	–	–	–
Multi-Sentence	59.5	68.8	63.6	71.6	56.0	66.4	29.2	37.0	32.4	39.7	26.1	34.3
DIALKI ($\mathcal{L}_{next}$ only)	60.4	71.2	64.2	72.3	57.1	70.2	31.5	39.7	33.3	41.1	29.8	38.3
+$\mathcal{L}_{hist}$	63.0	72.6	66.5	73.9	59.9	71.9	33.6	41.6	35.1	42.7	32.2	40.5
+$\mathcal{L}_{hist}$, *know-ctx*	63.8	73.4	**67.7**	74.8	60.5	72.3	33.6	41.5	35.2	42.8	32.1	40.3
+$\mathcal{L}_{adv}$	64.4	73.8	66.2	73.9	62.8	73.7	32.9	40.8	34.6	42.2	31.1	39.5
+$\mathcal{L}_{hist}$, $\mathcal{L}_{adv}$, *know-ctx*	**65.9**	**74.8**	67.6	**74.9**	**64.4**	**74.7**	**34.2**	**42.1**	**35.9**	**43.5**	**32.6**	**40.7**

Table 3: Ablation results on Doc2Dial and WoW dev sets.
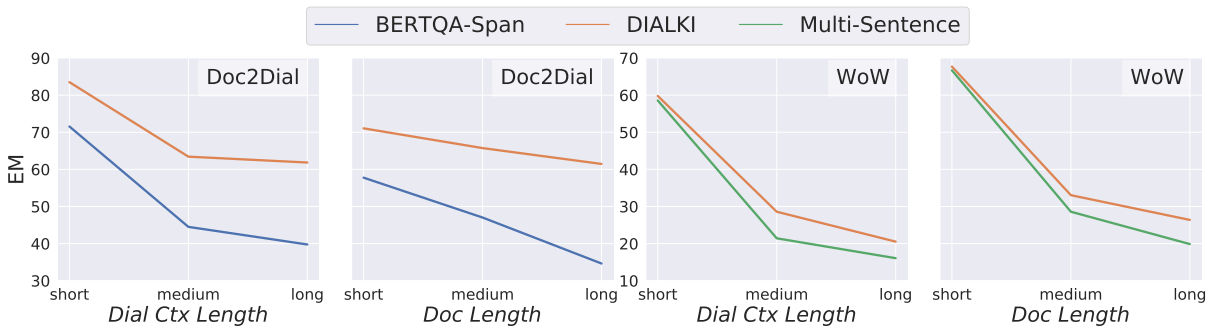

Figure 4: EM versus the length of dialogue context (# previous turns) or document (# tokens).

KI Model	Knowledge Input	sacrebleu
–	full doc	22.84
BERTQA-Token	pred span	21.42
DIALKI	pred span	25.16
DIALKI	pred span & passage	**25.84**

Table 4: Response generation results on Doc2Dial dev set. KI stands for Knowledge Identification.

the score is relatively high for the more recent user turn and decreases for earlier turns. Such patterns are not observed in WoW. One potential reason is that each agent in a Doc2Dial conversation has a clear goal to directly address user queries, while WoW conversations are more like social chat. This distinction may explain why *know-ctx* does not work well on WoW. The reason for even higher similarities with earlier turns in WoW could be that knowledge is related to people being referred to with pronouns with names being introduced earlier.

Availability of History Knowledge Labels In Doc2Dial, we calculate the history knowledge prediction loss ($\mathcal{L}_{hist}$) for all history turns, since all the labels are available in the training set. In practice, it might not be feasible to annotate all history turns, particularly user turns. Hence, we conduct exper-

iments comparing scenarios where history knowledge labels are given for all turns, agent turns only, random 50% agent turns, or no turns. We get EM scores of 63.0, 62.7, 62.4 and 60.4 respectively, finding that removing user turn labels and half of agent turn labels do not affect the results much.

Impact of Dialogue / Document Length Figure 4 shows the average EM scores vs. the dialogue context and grounding document length on the Doc2Dial and WoW dev set. Dialogue context lengths are grouped into 0-2 (short), 3-5 (medium) and $\geq$ 6 (long) history turns. In Doc2Dial, documents are categorized as short, medium and long: 0-500, 501-1000 and 1000+ tokens, respectively. In WoW, documents are categorized as short, medium and long: 0-800, 801-1600 and 1600+ tokens, respectively. DIALKI shows less performance drop as the two input lengths increase compared with baselines that do not leverage the multi-passage structure of grounding documents.

Span Prediction Error Analysis We randomly select and analyze 50 examples from both Doc2Dial and WoW where DIALKI makes wrong predictions (EM=0). Since DIALKI can achieve relatively high passage-level prediction accuracy as shown in Table 6, we focus on analyzing pre-

Dialogue Context	*User*: I want to trade in my license for a New York one. *Agent*: You have to exchange your-out-state driver license within 30 days of … *User*: What if I need my license for when I go back to my other country? … *Agent*: Are you studying in New York State? *User*: No	*User*: I heard something about co-op training pro- gram. Could you tell me about it, please? *Agent*: A co-op training program refers to … *User*: Sounds awesome. What should I do to get that? *Agent*: Well, for that, lets do a little fact check, shall we? Are you using VA educational assistance? *User*: Yes, I'm using that.
Baseline	Do you need a New York State driver license?	Good. In that case, you may be able to get money for books, tuition and housing.
Ours	Are you a driver from another country?	Are you enrolled at an approved institution of Higher Learning?
Gold Response	Are you a licensed driver from another country?	Good. Now, are you enrolled at an approved institu- tion of Higher Learning or IHL?

Table 5: Sample generated responses from BART with the full grounding document (baseline) or the predicted grounding span and passage by DIALKI (ours) as the additional input to the dialogue context.

Method	Doc2Dial		WoW	
	Seen	Unseen	Seen	Unseen
BERTQA-Span	76.9	72.7	–	–
Multi-Sentence	85.3	81.6	68.0	57.8
DIALKI ($\mathcal{L}_{\text{next}}$ only)	86.6	84.4	72.9	69.0
DIALKI	**88.5**	**87.5**	**73.4**	**69.7**

Table 6: Passage prediction accuracy on dev sets.

diction errors where the passage is predicted correctly. Among the 50 examples from Doc2Dial, we find 6 error types of predicted knowledge: (1) overlapped with the ground truth knowledge string while including redundant or missing relevant details (38%); (2) relevant to the user query but at an incorrect granularity level (18%); (3) completely irrelevant (16%); (4) relevant to history user queries instead of the current one (14%); (5) contains keywords of the last user utterance that are irrelevant (10%); (6) wrong gold labels (4%).

For WoW, the conversation style is different and the predicted knowledge is a single sentence instead of multiple spans, so prediction errors fall into different classes: (1) open-ended situations where the predicted knowledge is appropriate (48%); (2) unnatural for use in the next response (30%); (3) the predicted knowledge is more appropriate to use than the ground truth (12%); (4) irrelevant knowledge that does not answer the user's questions (10%). The high percentage of open-ended examples explains the relatively low evaluation scores of knowledge identification on WoW.

5 Conclusion

In summary, we introduce DIALKI to address knowledge identification in conversational systems with long grounding documents, taking advantage of document structure to contextualize document passages together with the dialogue history. DI- ALKI uses a multi-task objective that identifies knowledge for the next turn and used knowledge for previous turns, which captures interconnections between the dialogue and the document. Additional posterior regularization in learning further improves results. The model gives state of the art performance for this task on Doc2Dial and Wizard of Wikipedia, respectively. We show that improvements in knowledge selection transfer to response generation with a baseline generator.

The current study is limited by the static nature of the available data. Further work is needed to assess performance in an interactive setting. It would also be of interest to consider scenarios where document information is irrelevant or users change their minds about what information they want.

Acknowledgments

Ethical Considerations

Our work is primarily intended to address and encourage future work on the task of knowledge identification in document-grounded dialogue systems. We believe that locating relevant knowledge to be used in the following conversation can be useful

for improving the interpretability of response generation models. Knowledge identification can also play an important role in human-in-the-loop assistant scenarios. This places greater control into the hands of human agents instead of automatic response generation models, which tend to suffer from ethical issues like generating hallucinated (Zellers et al., 2019; Wu et al., 2021) or toxic content (Pavlopoulos et al., 2020).

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

H. Cheng, X. Liu, L. Pereira, Y. Yu, and J. Gao. 2021. Posterior differential regularization with f-divergence for improving model robustness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Hao Cheng, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2020. Probabilistic assumptions matter: Improved models for distantly-supervised document-level question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5657–5667, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,

pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 5110–5117.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *International Conference on Learning Representations*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. 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, Online. Association for Computational Linguistics.

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

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

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International*

Joint Conference on Artificial Intelligence, IJCAI-19, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 41–52, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Knowledge guided text retrieval and reading for open domain question answering. *CoRR*, abs/1911.03868.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

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

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. DeepCopy: Grounded response generation with hierarchical pointer networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 122–132, Stockholm, Sweden. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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.

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, Online. Association for Computational Linguistics.

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

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

A Experimental Setup Details

We initialize and finetune on BERT (Devlin et al., 2019) downloaded from Huggingface Transformers (Wolf et al., 2020).[9] We use the uncased base model of BERT in most of our experiments, and set $3e^{-5}$ as the learning rates and 1000 as warm-up steps with linear decay. For each experiment, we search the weights in Eq. 4 on the dev set in the ranges of $\alpha = \{0.5, 1, 2\}$, $\beta = \{0.5, 2.5, 5\}$. We eventually use $\beta = 5$ for all experiments, and $\alpha = 1$ for Doc2Dial and $\alpha = 0.5$ for WoW. We search for fewer than 5 hyperparameter trials for each experiment. All models are trained for 20 and 10 epochs for Doc2Dial and WoW respectively. Models are selected based on the best dev set EM score. The maximum length of dialogue context is 128. The maximum lengths of model input for each passage are 512 and 384 for Doc2Dial and WoW respectively, due to the larger variation in passage length in Doc2Dial. During training, we feed up to 20 passages into the models without posterior regularization. Otherwise the number of passages is reduced to 8 or 10 depending on the memory consumption. For inference, we feed up to 20 passages into the model for all experiment settings. Each training process is run on 2 NVIDIA Quadro Q6000 GPUs. It takes about 18 and 10 hours to train with or without posterior regularization for both datasets. The inference time takes less than 1 minutes per 1000 examples for all experiment settings on 2 GPUs. All our models based on uncased BERT base model contains between 110 to 115 million parameters.

B Dataset Details

We follow all original preprocessing and evaluation scripts for data processing of the following datasets.[10] These original data preprocessing scripts contain the step of downloading data.

Doc2Dial (Feng et al., 2020) contains about 4.8k English goal-oriented dialogues in 4 social-welfare domains, with an additional Covid-19 domain in the blind held-out test set. Each dialogue has an average of 14 turns, grounded on a long document with more than 1k tokens on average. Each user or agent turn is grounded in a sequence of knowledge spans as annotated in the dataset. In terms of the number of agent turns, there are about 20k / 4k examples in the train / dev set. The blind held-out test set contains 800 examples.

WoW (Dinan et al., 2019) contains over 20k English social chat conversations with an average of 9 turns on over 1k open-domain topics. For each agent turn, the agent (i.e., wizard) chose one or no grounding sentence from on average 7 Wikipedia passages retrieved by a pre-defined retriever based on the dialogue history for composing the response. Each passage contains 10 sentences. The original data has its dev/test set split to two subsets, which contain conversations about topics seen or unseen in training. It contains about 18k dialogues for training, 2k dialogues for validation and 2k dialogues for test. The test set is split into two subsets, Test Seen and Test Unseen. Test Seen contains 965 dialogues on the topics overlapped with the training set, while Test Unseen contains 968 dialogues on the topics never seen before in training and validation set.

---

[9] https://github.com/huggingface/transformers

[10] Doc2Dial: https://github.com/doc2dial/sharedtask-dialdoc2021; WoW: https://github.com/facebookresearch/ParlAI/tree/master/parlai/tasks/wizard_of_wikipedia