doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset

Song Feng Hui Wan Chulaka Gunasekara Siva Sankalp Patel
Sachindra Joshi Luis A. Lastras
IBM Research AI
{sfeng@us, hwan@us, chulaka.gunasekara@}.ibm.com
{siva.sankalp.patel@, jsachind@in, lastrasl@us}.ibm.com

Abstract

We introduce **doc2dial**, a new dataset of goal-oriented dialogues that are grounded in the associated documents. Inspired by how the authors compose documents for guiding end users, we first construct dialogue flows based on the content elements that corresponds to higher-level relations across text sections as well as lower-level relations between discourse units within a section. Then we present these dialogue flows to crowd contributors to create conversational utterances. The dataset includes over 4500 annotated conversations with an average of 14 turns that are grounded in over 450 documents from four domains. Compared to the prior document-grounded dialogue datasets, this dataset covers a variety of dialogue scenes in information-seeking conversations. For evaluating the versatility of the dataset, we introduce multiple dialogue modeling tasks and present baseline approaches.

1 Introduction

The task of reading documents and responding to queries has been the trigger of many recent research advances. On top of the development of contextual question answering QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019), more recent work MANtIS (Penha et al., 2019) and DoQA (Campos et al., 2020) included more kinds of user intents for querying over documents; while ShARC (Saeidi et al., 2018) added follow-up questions from agents and binary answers from users for the inference over documents. These exciting works confirm the importance of modeling document-grounded dialogue. Yet, it involves more complex scenes in practice, which requires better understanding of the inter-relations between conversations and documents. Thus, we aim to investigate how to create the training instances to further approach real-world applications of document-grounded dialogue for information seeking tasks.

In this work, we propose a new dataset of goal-oriented document-grounded dialogue. Figure 1 shows sample utterances from dialogues D1, D2 and D3 between an assisting agent and a user, and an example document in the middle. D1 and D2 are grounded in the given document, while D3 is irrelevant to the document. It illustrates two different types of contexts that we aim to capture: (1) *dialogue*-based context, where a query could be formed by a single or multiple turns, and (2) *document*-based context, which corresponds to varied forms of knowledge represented in the document. More specifically, *dialogue*-based context of a query could be initiated by a user (e.g., U1 in D1) or an agent (e.g., A3 in D1), and carried out through multiple turns by both roles (e.g., all turns in D2). *Document*-based context could involve structural elements in documents, such as the headers T1 and T2 or list items of m1 and m2, as well as textual discourse units, such as clauses (e.g, "If your clothing has been damaged").

For creating such dataset, we consider the document contents for social welfare websites,such as ssa.gov and va.gov, which guide users to access various forms of information. We develop a pipeline approach for dialogue data construction. Inspired by how human authors compose user-facing web content, we utilize both the high-level hierarchical relations between document components, as well as the low-level semantic relations between discourse units (Stede et al., 2019) to dynamically create outlines of dialogues, or we call dialogue flows. A ***dialogue flow*** is a sequence of interactions between an assisting agent and a user. Each turn contains a ***dialogue scene*** that is defined by a dialogue act, a role (user or agent) and a piece of grounding content from a document. Then we present these dialogue flows to crowd contributors to create conversational utterances. Such approach helps to avoid additional noise from the post-hoc

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8118–8128,
November 16–20, 2020.

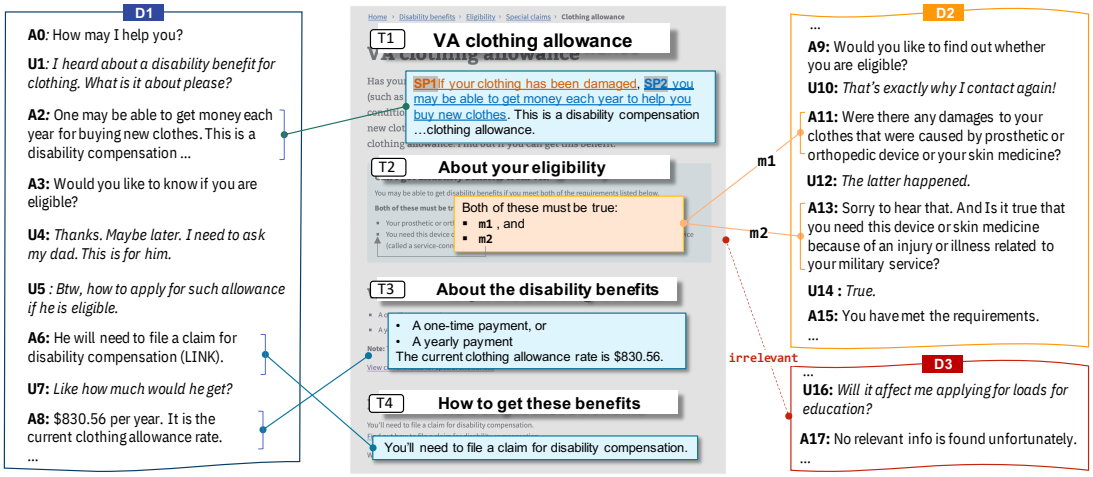

Figure 1: Sample segments of conversations (D1, D2 and D3) with various dialogue scenes that are grounded in a webpage (middle) from va.gov. The relevant content elements, such as hierarchical headers, list-items and spans, are highlighted. A / U indicates Agent / User role.

human annotations of dialogue data, which is a known challenge (Geertzen and Bunt, 2009).

The dataset contains about 4500 annotated conversations with an average of 14 turns per dialogue. The utterances are grounded in over 450 documents from four domains. Unlike the previous work on document-grounded question answering or dialogues (Choi et al., 2018; Reddy et al., 2019; Saeidi et al., 2018) that are based on a short text snippet, our dialogues are grounded in a much wider span of context in the associated documents.

For evaluation, we propose three tasks that are related to identifying and generating responses with grounding content in documents: (1) user utterance understanding; (2) agent response generation; and (3) relevant document identification. For each task, we present baseline approaches and evaluation results. Our goal is to elicit further research efforts on building document-grounded dialogue models that can incorporate deeper contexts for tackling goal-oriented information-seeking tasks. We summarize our main contributions as follows:

- We introduce a novel dataset for modeling dialogues that are grounded in documents from multiple domains. The dataset is available at http://doc2dial.github.io/.
- We develop a pipeline approach for dialogue data collection, which has been adapted and evaluated for varied domains.
- We propose multiple dialogue modeling tasks that are supported by our dataset, and present the baseline approaches.

2 Doc2Dial

We introduce **doc2dial**, a new dataset that includes (1) a set of documents; and (2) conversations between an assisting agent and an end user, which are grounded in the associated documents. Figure 1 presents sample utterances from different dialogues along with a sample document from va.gov in the middle. It illustrates some prominent features in our dataset, such as the cases where a conversation involves multiple interconnected sub-tasks under a general inquiry (e.g., D1); or the cases where a conversation involves multiple interactions to verify the conditional contexts for one query (e.g., D2).

Recent work, such as Saeidi et al. (2018), has started to address the challenge of modeling complex contexts by allowing follow-up questions from agents based on natural language inference rules extracted from the relevant documents. However, it also simplified the task by using only restricted forms of questions and binary answers. In our work, we not only encourage free-form utterances, but also aim to include various dialogue scenes that provoke inquires with different *document*-based and *dialog*-based contexts. A user query can be formed in single-turn or multiple-turn manners: (1) the user explicitly states a context that is associated with a text-span that contains a solution to the query, e.g., U5 on T4; (2) the user describes an implicitly stated context associated with a solution, e.g., U7; (3) the user accepts or rejects a piece of agent-stated context that is associated with a solution, e.g., U4 (rejection), and U12 & U14 (acceptance). An agent response, on the other hand, either

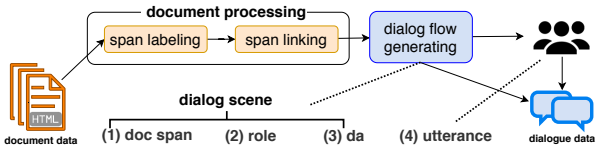

Figure 2: The overview of the process for constructing and annotating **doc2dial** dataset.

provides a solution or poses a query depending on the context of a given user query: (1) whether the query is irrelevant to the grounding document, e.g., A17; (2) whether the query is under-specified, if so, the agent will suggest associated context, e.g., A11 and A13; (3) whether a relevant answer is identified in the grounding document, e.g, A6, A8 and A15.

2.1 Data Collection

For collecting document-grounded dialogue data, we propose a pipeline approach derived from the framework proposed by Feng et al. (2020). As shown in Figure 2, it includes the components for: (1) processing the document contents; (2) generating dynamic dialogue flows; (3) crowdsourcing the dialogue utterances.

2.1.1 Data Construction Approach

Processing document contents We first select documents that contain the context-indicative elements, such as hierarchical headers and explicit discourse relations (Prasad et al., 2008, 2019), since those document contents could provoke more diversified dialogue flows. Then we extract text-spans to create a graph with the spans as nodes and semantic relations as edges. Some spans in the graph correspond to a piece of information for solving user problems, while some correspond to the conditional context of those solutions, such as `SP2` and `SP1` in Figure 1 respectively. The semantic relations are largely determined by the heuristics derived from the document structures (Mukherjee et al., 2003) and semantic connectives (Das et al., 2018) between discourse units or clauses. Both spans and semantic relations are labeled automatically via our tool. The labels can be reviewed and annotated via crowdsourcing platforms, which is also supported by our tool.

Generating dynamic dialogue flows Each flow consists of a sequence of dialogue scenes. A ***dialogue scene*** is described with (1) role, either a user or an agent; (2) a selected span as the grounding content from the given document; (3) a dialogue act that determines how to describe the selected span in the given role. Thus, each turn is inherently annotated with the dialogue act and a reference to the document contents. The dynamics of the dialogue flows are introduced by varying the three factors that are constrained by the relations from the semantic graph and dialogue history. In principle, we randomly select content from a candidate pool of spans of conditional contexts and solutions. The pool is updated after every turn is generated based on the status of the previously selected span. The general rule for updating the candidate pool is to avoid re-selecting any spans with an established status. In addition, the dialogue flow is principally aligned with common practice of dialogue management, for instance, after an agent asks a user a question, we expect the next turn would be the user answering the question.

Collecting human utterances Finally, we present the sequences of dialogue scenes to crowdsourced contributors to convert them into conversational utterances.

2.1.2 Crowdsourcing Setup

Our data collection task asks the crowd contributors to focus on one turn at a time so that they can carefully review the given dialogue scene and the dialogue history. Since the crowd generally prefers to work on tasks in batches, we try different settings to combine the tasks: (1) each writer plays the same role but for different dialogues per batch; or (2) each writer plays both agent and user role and completes entire dialogue in order, as inspired by Byrne et al. (2019). We also find that the conversations by the second setting tend to be more coherent and less time consuming. Many writers would make efforts to differentiate their writing styles for different roles. Therefore, our tasks were completed based on the second setting by about 70 qualified contributors from `appen.com`. We pay $1.5-$2 per conversation.

2.2 Document Data

For document contents, we consider the public government service websites that are designated to provide information to a vast group of users. We collect web contents from four domains and select about 450 documents for creating dialogue flows as shown in Table 1. Our dataset provides document contents in plain text and HTML, along with the meta information of titles and URLs. Each docu-

Domain	#Dials	#Docs	# per doc			
			tk	sp	p	sec
ssa.gov	860	86	758	66	16	5
va.gov	1340	138	823	70	20	9
dmv.gov	1420	149	955	77	18	10
cdc.gov	850	85	1251	94	16	9
all	4470	458	947	77	18	8

Table 1: The breakdown count of the dialogues, documents and average number of content elements per document by domain.

Role	DA	#Turns	#Tokens/Turn
user	request/query	25719	13
agent	request/query	8574	13
user	respond/yesOrNo	9254	7
agent	respond/reply	26273	24
total	all	69820	14

Table 2: The total # of turns and the average # of tokens per turn, aggregated on dialogue act category.

ment is also represented as a sequence of spans, for which we provide indexes to the plain text and the HTML respectively.

Content elements To characterize the document contents, we examine the HTML source to extract the content elements with different scopes such as, tokens (**tk**), spans (**sp**), paragraphs (**p**) and titled sections (**sec**). Some of the spans within one sentence, such as `SP1` and `SP2` in Figure 1, are extracted via constituency parsers (Joshi et al., 2018). The paragraphs and sections are determined using HTML markups. The average counts of these elements per document in Table 1 show the rich structures that are employed across domains. While this work starts to explore the simpler semi-structured information such as D2 in Figure 1; we are yet to explore various semantics from complex list structures, tables and other multi-modal contents in the webpages for future work.

2.3 Dialogue Data

Given a grounding document, we create about 10 unique dialogue flows with an average of 14 turns for this dataset. All dialogues are created based on a unique dialogue flow. In total, there are about 4500 conversations with close to 70,000 turns from four domains as shown in Table 1. Each dialogue utterance is annotated with a dialogue scene, i.e., role, dialogue act and the grounding span. As it is a known challenge to annotate conversation turns for the dialogue scenes (Geertzen and Bunt, 2009), our pipeline approach for data collection helps avoid

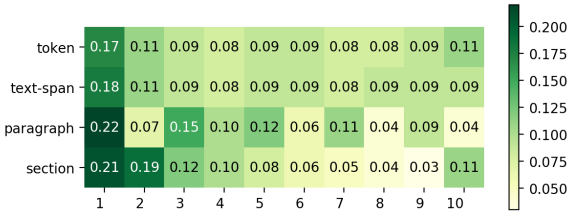

Figure 3: An illustration of the indexes of the relevant grounding contents in the documents.

the cost and the noise from the additional human annotations. Next we further describe it from different perspectives regarding the dialogue scene.

Dialogue acts We adopt the hierarchical dialogue act scheme by Pareti and Lando (2018) with a focus on the ones most essential to the information-seeking tasks. We describe those dialogue acts to the crowdsourced contributors pertaining to the selected grounding content and the assigned role (detailed descriptions in Appendix A). For future work, we plan to extend current dialogue scenes with other actions such as elucidations (Azzopardi et al., 2018) and social acts (Klüwer, 2011). To examine the dialogue distributions, we aggregate the hierarchical dialogue acts and list the total of turns, and the average length per turn under each category in Table 2. For example, "agent — request/query" corresponds to the queries based on document-guided dialogue management turns via an agent role; "user — respond/yesOrNo" corresponds to the scene where a user responds to an agent's query. Since we encourage the crowd to express "yes" or "no" in natural and creative writings, such as U10 in D2 in Figure 1, the average length of "respond/yesOrNo" is 7 tokens.

Grounding content We aim to include the contents that are associated with varied conditional contexts based on the aforementioned span graph without introducing strong bias on certain index position in the document as discussed in Geva and Berant (2018). Therefore, we examine the coverage of the document contents from the generated dialogue flows. As illustrated in Figure 3, we create index of all the selected grounding contents to different document segments such as tokens, spans, paragraphs and titled sections (y-axis). The x-axis (numbered 1-10) indicates the position where 1 is closest to the beginning and 10 is closest to the end of a document. The numbers in the cells indicate the percentage distribution among all the ground-

feedback on rejected dialogue scene	%
The selected-text is not a contextual condition.	74.3
The selected-text is not a solution to the query.	10.5
Cannot write a turn to be coherent with the chat history.	10.1
There is not enough information in the selected (or adjacent) text.	2.4
The selected-text is not Comprehensible.	1.8
Other.	0.9

Table 3: Feedback on the reasons for rejecting a dialogue scene by crowdsourced annotators.

ing contents. The heatmap shows some degree of coverage on all parts of the documents, with a higher density at the beginning as we do include the scenarios of under-specified queries that typically correspond to the intro of a document.

Dialogue flows For assessing the quality of the dialogue flows, we also ask the contributors to reject a dialogue turn when it is considered as infeasible to write a coherent utterance. We also solicit feedback via multiple choices on the reason as shown in Table 3. Out of 700 sampled dialogue flows, annotators reject about 4% of the turns. Among the rejected turns, 70% is due to not being able to interpret the selected span as applicable conditional context for user requests. In this dataset, we exclude the (sub)dialogues with rejected turns accordingly. However, we also observe certain "false positive" cases, where the crowd would rather try to adjust their writing for a less desirable dialogue scene rather than rejecting the turn, for which they get paid the same either way.

2.4 Data Recomposition

One benefit of constructing the dialogue data via our pipeline approach is that it provides a convenient and cost-effective way to reshape the existing dialogue data based on their dialogue flows. For instance, to ensure the quality, we can recollect or remove certain turns from the dialogues if they are rejected by the crowd contributors or affected by the changes in the grounding documents. In addition, for obtaining the training instances to identify the irrelevant queries, we modify an existing dialogue by inserting sub-dialogues created for another document or domain, for instance, adding `D3` to `D1` as *irrelevant* for `va.org` in Figure 1. Similarly, for creating dialogues that are grounded in multiple documents, we select sub-dialogues based on different documents and combine them into one.

3 Tasks and Baselines

For evaluation, we propose three tasks related to identifying the grounding content for a given dialogue: (1) user utterance understanding; (2) agent response prediction; (3) relevant document identification. In our tasks, we also aim to detect the cases that are irrelevant to the associated documents, for which we modify dialogues to include *irrelevant* (`Irr`) queries via data re-composition as described in Section 2.4. We split the dialogues into train/dev/test sets as 70%, 15%, 15% with half of the dev/test set grounded in "unseen" documents (not in training set). Experiment results are on test set unless otherwise stated. Numbers in the form of "mean $\pm$ stdev" are computed out of 3 random seeds.

3.1 User Utterance Understanding

One of our main goals for creating this dataset is to broaden the coverage of different user queries for various task goals. Thus, our first task is interpreting a user utterance based on the dialogue history and the grounding document content. It aims to identify the associated dialogue scene, i.e., (1) grounding span in the document and (2) dialogue act, as described in the following two sections respectively.

3.1.1 User Utterance Grounding

In our dataset, all turns are associated with a dialogue scene that includes the grounding span. Interpreting the user utterance could be quite challenging, because in some cases, it would completely depend on the dialogue history such as U12 and U14; while some cases, such as U1 and U16, depend more on the user utterance itself. For the input of this task, it takes a user utterance along with (1) the dialogue history and 2) the document content with simplified document structure. The output is a span in the document as the text reference of the given user utterance. Each grounded user turn is considered a training instance, so a dialogue with n grounded user turns is considered as n instances, with overlapping dialogue context.

Baseline Approach We formulate the problem as span selection, inspired by extractive question answering tasks such as SQuAD task (Rajpurkar et al., 2016, 2018). As a baseline, we adopt the extractive question answering model with transformers encoder by (Devlin et al., 2019). More specifically, we follow the QA example from Hugging-

Face Transformers (Wolf et al., 2019). Pretrained bert-large-uncased-whole-word-masking model is used as encoder, and is fine-tuned during training.

The document content serves as the context input of the model. The query input is the dialogue context, for which we experiment different settings of utilizing the dialogue history: (1) "last two turns", i.e., the input user utterance for which we want to identify the dialogue scene, and the agent utterance before the given user utterance; (2) "all prev", i.e., the input user utterance and all the utterances before it; (3) "all prev w/DA", i.e., context in (2) along with the corresponding dialogue acts. The dialogue context is concatenated in reversed time order where the latest user utterance appears first.

Often the grounding document is longer than the maximum sequence length of transformers. In such cases, we truncate the documents in sliding windows with a stride. The dialogue context and each document trunk form one instance to be fed in batch into the encoder. The sequence of the encoded embeddings is then sent to a linear layer, which maps each embedding in the sequence into two logits, representing the probability of the corresponding position being the start and end position of the span. During training, we apply the Cross Entropy loss function to compute the loss. If the ground truth span does not fall in the document trunk, the start and end positions are both considered to be the beginning of the sequence. During decoding, the start-position and end-position logits from all document trunks are considered together to find the span most favored by the model.

Evaluation Metrics For evaluation we use Exact Match score and token-level F1 score, as in the evaluation script 2.0 of SQuAD . In addition, since our data comes with predefined spans in each document, we map predicted span to the closest predefined span start index and span end index, and evaluate the mapped span with Exact Match score, as "ts EM" in Table 4 and Table 7.

Experiment Results The experiment results are summarized in Table 4. Generally, the model performance improves with more information added to the dialogue context. It indicates that the queries in our datasets are highly conversational contextual and our dataset could serve as a valuable source for evaluating dialogue models' capability of learning from deeper context. We also conduct an experiment using the `w/Irr` data with the "all prev" dialogue context. Table 5 summarizes the results in comparison with `wo/Irr` data. `Irr` turns impose noise in understanding the context, reduce the model accuracy from 54.3 to 53.3 on the original turns that are grounded to the document. However, the `Irr` turns themselves are easy to identify and achieve a high score of 99.1. As a result, the overall score including the `Irr` turns is increased to 62.7.

dial-ctxt	text EM	text F1	ts EM
last two turns	52.6 ± 0.3	64.3 ± 0.3	52.5 ± 0.4
all prev	54.3 ± 0.5	66.2 ± 0.3	54.4 ± 0.2
all prev w/ DA	55.1 ± 0.4	66.3 ± 0.3	55.2 ± 0.4

Table 4: Results for user utterance grounding.

all prev	text EM	text F1
`wo/Irr`	54.3 ± 0.5	66.2 ± 0.3
`w/Irr`	62.7 ± 0.4	70.1 ± 0.6
`has_ans` turns	53.3 ± 0.4	62.7 ± 0.8
`Irr` turns	99.1 ± 0.3	99.1 ± 0.3

Table 5: Comparison of `w/Irr` and `wo/Irr` settings for user utterance grounding.

3.1.2 User Dialogue Act Identification

Dialogue act prediction using dialogue context as input is an important task in dialogue systems modeling (Liu et al., 2017; Tran et al., 2017). We indentify the dialogue act of each user turn considering three different cases of dialogue context as input: (1) U: only the input user utterance, (2) U + A: the input user utterance and previous agent utterance, and (3) U+A(w. da): inputs in (2) along with agent turn's dialogue act. We use the hidden state of the tokens as the representation of the dialogue context, and further process it by a linear layer to identify the probability distribution over the total number of user dialogue acts. There are 7 dialogue acts for `w/Irr`, 6 dialogue acts for `wo/Irr`. We use the common metrics of accuracy (Acc), recall (R) and precision (P) for evaluation.

Baselines and Experiment Results As a baseline we adopted BertForSequenceClassification model, a multi-class sequence classifier popular for GLUE tasks (Wang et al., 2018). We use pretrained bert-base-uncased model as the encoder and fine-tune during training.

The results in Table 6 indicate much room for improvement, e.g., by adding document context, or building a joint model with the user utterance grounding task. The macro-averaged P and R are much lower than the micro-averaged Acc because

dial-ctxt	w/Irr			wo/Irr		
	Acc.	**R.**	**P.**	**Acc.**	**R.**	**P.**
U	60.2	34.5	35.6	77.2	45.2	47.2
U+A	72.7	51.8	53.7	79.3	50.8	48.6
U+A(w. da)	76.4	53.8	55.7	80.6	50.9	55.2
-	53.3	14.3	7.6	67.4	16.7	11.2

Table 6: Results for user dialogue act identification by BERT. The last row is by majority vote. Acc. is micro-averaged, while R. and P. are macro-averaged.

of the imbalanced DA distribution as shown in Table 2. The results also reflect the challenges effectively posed by the introduction of `Irr` turns, which was intended by our task design.

3.2 Agent Response Prediction

For this task, we aim at predicting the next agent turn with a focus on identifying the reference to the grounding document for the response. Such task can be a very important step towards building explainable conversational systems with practicality.

3.2.1 Agent Response Grounding Prediction

This task takes as input 1) the dialogue context; and 2) the document content with simplified document structure, and predicts a span in the document that grounds the next agent response. This task looks very similar to the user-turn grounding text prediction task in Section 3.1.1 in that they both take dialogue context and document context as input and perform a span selection inside the document. However, they are essentially different: the user-turn grounding text prediction is to understand what the user has already said, whereas this task is to predict what the agent would want to respond.

Baseline Approach and Evaluation Metrics As opposed to investigating this task from the aspect of dialogue management and planning, as a first attempt, we continue with our focus on identifying the associated grounding content in the document. Thus, we treat this as a span selection task, and adopt the same evaluation metrics and baseline approach as in Section 3.1.1. Note that with the same input dialogue context and text context, the model output in Section 3.1.1 is the dialogue scene corresponding to the given user utterance, while the model output of this task is the dialogue scene predicted for the next agent response.

Experiment Results The experiment results are summarized in Table 7. The scores are much lower than the ones from our previous task in Table 4 due to the challenging nature of the task. However, we do see a significant improvement after including dialogue act information, which directs our further work on dialogue management to further improve the performance. Table 8 compares the experiment result in the `w/Irr` and `wo/Irr` settings, where we see a similar trend as in Table 5 unsurprisingly.

dial-ctxt	**text EM**	**text F1**	**ts EM**
last two turns	33.4 ± 0.4	49.6 ± 0.8	34.7 ± 0.5
all prev	34.3 ± 0.2	50.0 ± 0.8	35.9 ± 0.2
all prev w/ DA	36.2 ± 0.4	52.6 ± 1.0	37.6 ± 0.7

Table 7: Results for agent response grounding prediction.

all prev	**text EM**	**text F1**
`wo/Irr`	34.3 ± 0.2	50.0 ± 0.8
`w/Irr`	47.3 ± 0.2	57.6 ± 0.6
`has_ans` turns	33.8 ± 0.3	46.7 ± 0.7
`Irr` turns	98.8 ± 0.3	98.8 ± 0.3

Table 8: Comparison of `w/Irr` and `wo/Irr` settings for agent response grounding prediction.

3.2.2 Agent Response Generation

Next we evaluate the dataset via the task of generating agent response. One primary goal of our task is to enable document-guided agent response, which overlaps with the primary goal of ShARC (Saeidi et al., 2018). However, our dataset includes more types of dialogue scenes and sets no restriction on the natural language forms of queries and responses. Thus, we investigate how one of the best performing end-to-end approaches to-date for ShARC works on our dataset. Compared to ShARC, our user queries do not come with the scenario description but are annotated with the grounding span, and the grounding documents is much longer. Therefore, we truncate the document into sub-documents with a size of 200 tokens. We try different ways to truncate the text: (1) only at the end of a span (ts); (2) only at the end of a paragraph (p).

Baseline Approach and Experiment Results We adopt the model from Zhong and Zettlemoyer (2019). The input is the user query with dialogue history of up to last 4 turns as well as their grounding spans and the document content; the output is the agent utterance. The model learns to extract the relevant spans implicitly that are entailed by the dialogue-based and document-based contexts, and then edit them to generate the agent response.

The BLEU scores are reported in Table 9. We ob-

doc-ctxt	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ours (ts)	40.45	34.65	31.84	29.98
ours (p)	58.12	54.26	52.53	51.51
(ShARC)	(67.14)	(60.59)	(56.46)	(53.67)

Table 9: Results for agent turn generation (dev set).

serve better results with the preprocessing method that maintains the original document structure at larger scale. Compared to the results by the same model reported on ShARC dev dataset, our BLEU scores are significantly lower. This is related to the more dynamic forms of agent responses in our dataset; another factor is the length of relevant document context, even when truncated, ours is 4 times longer.

3.3 Relevant Document Identification

Given that the goal-oriented dialogues could correspond to different tasks from a same document or multiple documents, in order to facilitate the understanding of such challenge, we experiment with two settings for the task on retrieving the grounding document(s): (1) the dialogues that are grounded in a single document; (2) the dialogues that are grounded in multiple documents.

3.3.1 Single-Document Retrieval

This task is to identify the relevant grounding document given limited dialogue history information. Thus, the input is certain dialogue context and a pool of 594 documents from all four domains.

Baselines and Experiment Results We consider two different baselines for this task: (1) BM25 (Robertson and Zaragoza, 2009) based Information Retrieval method, and (2) A multi-class sequence classifier based on BertForMultipleChoice, using pretrained bert-base-uncased model as the encoder (Zellers et al., 2018).

BM-25 method takes the full document into account to create the index and match them against the provided dialogue contexts. BERT model takes the dialogue context d and a document y together as a sequence. We use 512 tokens and feed BERT with the 256 tokens each from d and y. For each dialogue context, we create a set of triples: one triple containing the correct document (labeled with 1), and m triples containing incorrect documents sampled randomly from the set of all documents (labeled with 0). Table 10 corresponds to the setting $m = 4$. During evaluation, we evaluate a given dialogue context against the set of all documents. The task is evaluated with the commonly used recall ($R@k$) metric in retrieval tasks, which measures the fraction of times the correct document is found in the top-k predictions.

n	BM-25			BERT		
	R@1	R@5	R@10	R@1	R@5	R@10
1	26.1	44.8	53.5	32.4	59.6	67.3
2	49.3	74.2	78.8	50.5	77.8	85.1
3	49.4	73.9	79.0	51.7	83.7	88.8
4	56.0	80.4	84.9	57.6	84.3	89.4
5	59.3	80.7	86.0	60.2	85.6	90.7

Table 10: Results for single-document retrieval with n previous turns as input.

Domain	R@1	R@5	R@10
va.org	52.3	78.3	86.4
dmv.org	50.5	76.4	86.4
ssa.org	33.6	74.2	86.1
cdc.org	46.8	74.1	83.2
Weighted Average	47.5	76.1	85.9

Table 11: Results for multi-document retrieval in single domain.

As shown in Table 10, DL-based approach shows better performance consistently. From the perspective of examining the quality of our dataset, we also see the numbers confirms that as more turns are included, the better the dialogue is grounded to the relevant document.

3.3.2 Multi-Document Retrieval

We construct the dialogues that are grounded in multiple documents as described in Section 2.4. To make the tasks more challenging and closer to real-life applications, the segments of a dialogue are all grounded in the documents from the same domain. This dataset contains 2051 conversations, out of which 1640, 206 and 205 conversations were used in the train, dev and test sets respectively.

Baselines and Experiment Results A baseline similar to Section 3.3.1 was constructed for this task using BertForMultipleChoice. At each user turn, we predict which document it should be grounded to, given the user utterance, previous agent utterance and the domain.

This task is essentially related to conversational search task Penha et al. (2019), which predicts a link to the relevant document given a dialogue. Even though our document pool is not large compared to IR tasks, each document is quite long. The results in Table 11 show much room for improvement, and our dataset could be valuable resource for further deep document modeling.

4 Related Work

Our work is mainly focused on modeling dialogues that are grounded in documents. It is generally inspired by the recent substantial interests on the challenges of machine reading comprehension and conversational QA, such as CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018) and DoQA (Campos et al., 2020). Those tasks aim to support conversational question answering, which involves understanding a text passage and answering a series of interconnected questions that appear in a conversation. These tasks add the complexity of coreference resolution and contextual reasoning to the reading comprehension challenges such as SQuAD (Rajpurkar et al., 2016, 2018), yet aim at identifying a solution from a given list of candidates by reasoning over spans from a document. Our task shares those challenges and additionally introduces the dialogue scenes where the agent asks questions when the user query is identified as under-specified or additional verification required for a resolute solution.

Another recent work Kim et al. (2020) extends MultiWOZ (Budzianowski et al., 2018) by adding turns that are grounded in the FAQ knowledge for certain entity and domain. The document-based knowledge used in our work is beyond FAQs with entity as context but whole documents with more complex contexts. In addition, ours is also largely related to conversational search tasks, such as MANtIS (Penha et al., 2019). Similarly, it also provides multi-turn conversations with varied user intents that are grounded in documents from Stack Exchange website. In addition to the domain difference, one major distinction is that the grounding in MANtIS is determined by the hyperlinks to a document. Our grounding is defined at at a much finer level in addition to the link to a document.

To the best of our knowledge, the closest related work to ours is ShARC (Saeidi et al., 2018) with dialogues that are grounded to a span of a given text snippet. It also proposes to address under-specified questions by requiring follow-up questions that are answerable with yes/no answers in similar domains. Our dataset goes beyond ShARC in several aspects nonetheless: we exploit not only paragraph-level structure but also higher-level document structure, we create conversations over much longer span of document content, where utterances are free-formed, as opposed to yes/no answers.

5 Conclusion

We have introduced **doc2dial**, a new dialogue dataset for goal-oriented tasks that are grounded in documents from multiple domains. Compared to previous work, our dialogues cover a greater variety of dialogue scenes that correspond to a much wider span of document content. For evaluation, we investigated three types of dialogue tasks and proposed baseline approaches. We hope this work will inspire and assist both dialogue and document modeling for tackling more real-life dialogue tasks.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. We thank Vera Liao and Kshitij Fadnis for their advice during the early stage of this project. We also thank crowd contributors for their valuable inputs for building our tool and dataset.

References

Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *The Second International Workshop on Conversational Approaches to Information Retrieval*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, 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.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of english discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365.

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.

Song Feng, Kshitij Fadnis, Q Vera Liao, and Luis A Lastras. 2020. Doc2dial: a framework for dialogue composition grounded in documents. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Jeroen Geertzen and Harry Bunt. 2009. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 126–133. Association for Computational Linguistics.

Mor Geva and Jonathan Berant. 2018. Learning to search in long documents using document structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 161–176, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1199, Melbourne, Australia. Association for Computational Linguistics.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access. *arXiv preprint arXiv:2006.03533*.

Tina Klüwer. 2011. "i like your shirt"-dialogue acts for enabling social talk in conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 14–27. Springer.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Saikat Mukherjee, Guizhen Yang, Wenfang Tan, and IV Ramakrishnan. 2003. Automatic discovery of semantic structures in html documents. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 245–249. IEEE.

Silvia Pareti and Tatiana Lando. 2018. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing mantis: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. In *LDC2019T05*. Philadelphia: Linguistic Data Consortium.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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.

M Stede, T Scheffler, and A Mendes. 2019. Connective-lex: A web-based multilingual lexical resource for connectives. discours. *Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics,(24)*.

Quan Hung Tran, Gholamreza Haffari, and Ingrid Zukerman. 2017. A generative attentional neural network model for dialogue act classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 524–529.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

Victor Zhong and Luke Zettlemoyer. 2019. E3: Entailment-driven extracting and editing for conversational machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2310–2320, Florence, Italy. Association for Computational Linguistics.