Exploiting Dialogue Act for Knowledge Selection and Response Generation

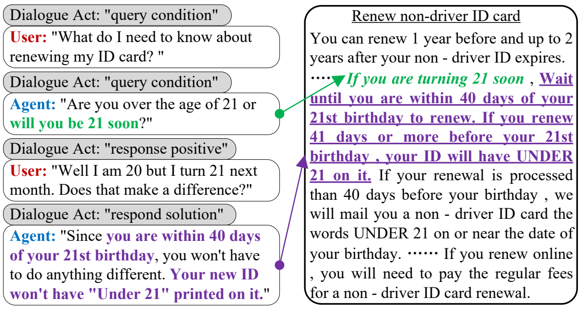
Anonymous ACL submission

Abstract

Dialogue act (DA) is the description of the intention or function of a dialogue utterance. In document-grounded dialogue, correctly understanding the dialogue context is crucial for models to select knowledge and inject knowledge into responses. Leveraging dialogue act can help to understand the dialogue context and consequently assist the utilization of document information. In this paper, we propose a novel framework leveraging two different kinds of DAs (model-annotated and human-annotated) for **Knowledge Selection** (KS) and **Response Generation** (RG). The framework consists of two modules: the prediction module is trained with multi-task learning and learns to select knowledge and predict the next DA; the generation module uses the selected knowledge and the predicted DA for the RG. Our model achieves new state-of-the-art performance on three public datasets and the results verify that leveraging DA can help KS and RG. Our code and data will be released on github.com.

1 Introduction

Neural conversation models aim to generate meaningful responses. However, it is widely observed that the generated responses lack sufficient information (Li et al., 2016; Ghazvininejad et al., 2018). Previous researchers proposed different methods to alleviate this issue, such as introducing external knowledge to generate informative responses. The external knowledge can be structured knowledge triples (Zhou et al., 2018a; Wu et al., 2019) or unstructured text (Ghazvininejad et al., 2018; Dinan et al., 2018). The document-grounded dialogue (DGD) (Zhou et al., 2018b; Moghe et al., 2018; Gopalakrishnan et al., 2019) belongs to the latter and uses a document as external knowledge. A document contains multiple logically related sentences, which together constitute a description of the topic of the document. Figure 1 shows an example of DGD in the Doc2Dial dataset (Feng et al., 2020).



Dialogue Act: "query condition"

User: "What do I need to know about renewing my ID card? "

Dialogue Act: "query condition"

Agent: "Are you over the age of 21 or **will you be 21 soon**?"

Dialogue Act: "response positive"

User: "Well I am 20 but I turn 21 next month. Does that make a difference?"

Dialogue Act: "respond solution"

Agent: "Since **you are within 40 days of your 21st birthday**, you won't have to do anything different. **Your new ID won't have "Under 21" printed on it.**"

Renew non-driver ID card

You can renew 1 year before and up to 2 years after your non - driver ID expires. ····· ***If you are turning 21 soon*** , **Wait until you are within 40 days of your 21st birthday to renew. If you renew 41 days or more before your 21st birthday , your ID will have UNDER 21 on it.** If your renewal is processed than 40 days before your birthday , we will mail you a non - driver ID card the words UNDER 21 on or near the date of your birthday. ······ If you renew online , you will need to pay the regular fees for a non - driver ID card renewal.

Figure 1: A DGD example in the Doc2Dial dataset.

The agent constructs the responses with the related document sentences as external knowledge.

Two main challenges in the DGD task are **knowledge selection** (KS) and **response generation** (RG). KS is to select relevant document information using dialogue context and RG is to use the selected information to generate a response. To utilize the document information, DGD models first need to correctly understand the dialogue context.

Some earlier work relied on the encoding ability of different kinds of encoders to capture the semantic information of dialogue (Zhou et al., 2018b; Moghe et al., 2018). These models performed the interaction between dialogue and document with attention operation (Meng et al., 2019; Qin et al., 2019), which was too simple to link the dialogue with related knowledge and extract salient information. Later work tried to capture the semantic information change between dialogue contexts Li et al. (2019) and capture the change of KS distribution with dialogue utterances Kim et al. (2020); Meng et al. (2020). However, these methods implicitly modeled the semantic information of dialogues. It is difficult for them to measure the understanding of dialogue intention.

Recently, some researchers (Hedayatnia et al., 2020; Feng et al., 2020) tried to utilize explicit information, such as Dialogue Act (DA), to assist the dialogues modeling in DGD. DA is long term studied (Bunt et al., 2010, 2020) in open-domain dialogue research and is defined as the description of the intention or function of an utterance (Kawano et al., 2019). Figure 1 shows an example where dialogue utterances are accompanied with DAs. The first utterance is labeled with a "query condition" DA, which means the user wants to acquire relevant document information. The second utterance is from the agent and also owns a DA of "query condition", which means the agent needs to clarify the age information of the user before answering. After getting a "positive" response, the agent can finally answer the first query with a DA of "respond solution". This example shows that DAs provide explicit guidance for utilizing the document information, in both KS and RG.

However, human-annotated DA is expensive and most DGD datasets (Moghe et al., 2018; Qin et al., 2019) do not have this kind of label. Recently, Hedayatnia et al. (2020) used an SVM tagger to automatically annotate DA on the Topical-Chat dataset (Gopalakrishnan et al., 2019). However, they only used these DAs for RG policy planning. Majumder et al. (2020) proposed a media interview dataset that labeled question types as DA. However, the DAs of the response utterances were not given. Most recently, Feng et al. (2020) introduced the Doc2Dial dataset with human-annotated DAs and their experiments showed the DA information was useful for KS but not helpful for RG.

In this paper, we exploit DA information for the DGD task and analyze two **research questions** (RQs): 1) Can we utilize DA information to improve both KS and RG in a DGD model? 2) There are two different DAs (human-annotated and model-annotated). Can the performances of the model-annotated DAs match the expensive human-annotated ones? We trained a DA tagger to annotate DA labels for DGD datasets. For RQ 1, we propose a framework that first selects knowledge and predicts the next DA, then uses the selected knowledge and the predicted DA to generate a response. For RQ 2, we test human/model-annotated DAs on the same dataset to compare their effectiveness. Our contributions are as follows:

(1) We propose a novel framework to leverage **D**ialogue **A**cts for **K**nowledge **S**election and response generation (DAKS) in DGD task.

(2) We train a BERT-based (Devlin et al., 2019) DA tagger with four public open-domain dialogue datasets[1] under the ISO DA standard. We use this

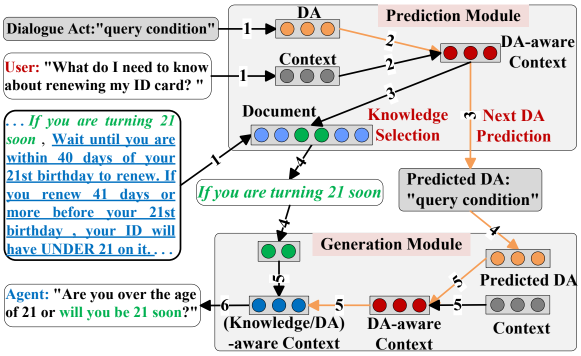

Dialogue Act:"query condition"
1
DA
Prediction Module
2
DA-aware Context
User: "What do I need to know about renewing my ID card? "
1
Context
2
3
Document
Knowledge Selection
3
Next DA Prediction
. . . If you are turning 21 soon , Wait until you are within 40 days of your 21st birthday to renew. If you renew 41 days or more before your 21st birthday , your ID will have UNDER 21 on it. . . .
1
4
If you are turning 21 soon
Predicted DA: "query condition"
4
4
Generation Module
5
Predicted DA
Agent: "Are you over the age of 21 or will you be 21 soon?"
6
5
5
(Knowledge/DA) -aware Context
DA-aware Context
Context

Figure 2: The architecture of the DAKS model.

well-trained tagger to annotate three public DGD datasets[2] for our experiments.

(3) We conduct extensive experiments and give a detailed analysis of the results. The experimental results show that: 1) DAKS owns better KS accuracy and RG quality than the state-of-the-art models on all three DGD datasets; 2) The model-annotated DAs have comparable effects with the human-annotated ones. These results can help the research of utilizing model-annotated DAs in open-domain dialogue research.

2 Our Proposed Model

2.1 Problem Statement

Given document $\mathbf{K} = [K_1,K_2,...,K_{|\mathbf{K}|}]$ with $|\mathbf{K}|$ sentences as external knowledge, a dialogue context $\mathbf{C} = [C_1,C_2,...,C_{|\mathbf{C}|}]$ with $|\mathbf{C}|$ turns and the response $R = [R_1,R_2,...,R_r]$ with r tokens, the DGD models learns to generate R with probability $P(R|\mathbf{K},\mathbf{C};\Theta)$, Θ is the model's parameters. We introduce DA information $\mathbf{S} = [S_1,S_2,...,S_{|\mathbf{C}|}]$ for $\mathbf{C}$ and S_R for R in this probability, then the generation model changes to $P(R,S_R|\mathbf{K},\mathbf{C},\mathbf{S};\Theta)$. The model learns to generate response and predict the DA of the response. When the ground truth knowledge is a sentence K_i, we can further separate $P(R,S_R|\mathbf{K},\mathbf{C},\mathbf{S};\Theta)$ into prediction module $P_{KS}(K_i,S_R|\mathbf{K},\mathbf{C},\mathbf{S};\Theta_{KS})$ and generation module $P_{RG}(R|K_i,\mathbf{C},S_R;\Theta_{RG})$, Θ_{KS} and Θ_{RG} are models' parameters. The prediction module predicts the next DA while selecting knowledge, then the predicted DA and selected knowledge are sent to the generation module to guide response generation.

2.2 Model Structure

The structure of DAKS is shown in Figure 2. We define P_{KS} and P_{RG} with BERT and GPT-2, re-

[1] They are DailyDialog (Li et al., 2017), Switchboard (Godfrey et al., 1992), AMI (Carletta et al., 2005), and Maptask (Anderson et al., 1991).

[2] They are WoW (Dinan et al., 2018), Holl-E (Moghe et al., 2018), and Doc2Dial.

spectively. The DA tagger we trained will be introduced in the Experimental Setup section.

2.2.1 Prediction module

When the dialogue context is three turns, the input to BERT model is a concatenated sequence [C_3;C_2;C_1;<ESP>;**K**], where [;] is the concatenation operation, <ESP> is a special token. Each DA description owns a specific trainable DA embedding. Hence each input word of the dialogue context is initialized with the sum of four embeddings: Word/DA/Positional (Vaswani et al., 2017)/Segment (Devlin et al., 2019). Words in **K** are similarly initialized except without the DA embedding. The multi-layer bidirectional attention mechanism in BERT allows the dialogue context **C** and the DA information **S** to sufficiently interact with each other, resulting in DA-aware Context representations, which is then used for KS and DA prediction.

For KS, we use the span extraction method in Machine Reading Comprehension (MRC) (Rajpurkar et al., 2016). This means the model learns to predict the start and end positions of a text span. We use a *Span Revision* method that forces the model to predict a whole sentence instead of random positions. Compared with the previous methods (Kim et al., 2020; Meng et al., 2020) of encoding candidate sentences into a vector representation, the MRC-based method can make better use of the semantic information between sentences. The reasons include: 1) The multi-head attention mechanism in BERT provides sufficient interaction between dialogue and knowledge sentences at the word level, so as to leverage the overall document information for selection; 2) The word-level interaction is consistent with the pre-training process of BERT, so as to fully leverage the ability of the model. A Cross-Entropy (CE) loss is calculated as the KS Loss:

$$\mathcal{L}_{KS} = -\frac{1}{N}\sum_{i=1}^{N}(\log P(y_i^s) + \log P(y_i^e)), \quad (1)$$

where N is the number of training samples, y_i^s and y_i^e are the ground-truth start and end positions of knowledge sentence, respectively. For the next DA S_R, we pass the last BERT layer's representation of the special token <ESP> into a MLP to predict the next DA. The DA prediction loss $\mathcal{L}_{DA}$ is a CE loss between predicted DA label y_i and ground-truth DA label $\bar{y}_i$:

$$\mathcal{L}_{DA} = -\frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i \log P(y_i)). \quad (2)$$

During training, the KS module needs to select the accurate knowledge sentence and predict the correct DA simultaneously. The $\mathcal{L}_{DA}$ is easier to be trained since the DA only has a few categories. We test different weight coefficient to balance the two objects: $\mathcal{L}_{BERT}$ = $\alpha\mathcal{L}_{DA}$+$(1-\alpha)\mathcal{L}_{KS}$. The empirical value is α = 0.25.

2.2.2 Generation module

GPT-2 is a Transformer-based (Vaswani et al., 2017) language model with a stack of masked multi-head self-attention layers which are suitable for language generation tasks. Following Rashkin et al. (2021), we treat DA descriptions as control tokens and prepend DAs to the input sequence of the GPT-2 model. The input to generation module is [S_R;C_3;C_2;C_1;K_i], words are initialized with the sum of Word/Positional embeddings. Dialogue context first interacts with S_R to get new DA-aware Context representations, which consequently interacts with K_i to get (Knowledge/DA)-aware representations. The final interaction results are used for generation. The RG Loss is as follows:

$$\mathcal{L}_{RG} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{r}(\log P(R_i^t)), \quad (3)$$

where R_i^t is the t-th word of the i-th response. The BERT and GPT-2 models are fine-tuned separately and then combined into a DGD framework.

3 Experimental Setup

3.1 Datasets

We choose three public datasets in DGD research: WoW, Holl-E, and Doc2Dial. They all have ground-truth knowledge sentence labels that can be used to test the KS accuracy. Doc2Dial has human-annotated DAs while WoW and Holl-E do not have. After using DA tagger to annotated DA on them, we can compare model-annotated DAs with human-annotated ones on Doc2Dial and verify the effect of using DA information on all three datasets. We use the 1.0.1 version of Doc2Dial and use the validation set for testing since we do not have the access to the test set. Tabel 1 shows the data statistics[3].

[3] There are Test seen/unseen sets in WoW according to whether including topics not seen in the training set and there are Test single/multi-references sets in Holl-E according to one/multiple ground-truth responses to a dialogue context.

Datasets	Dialogues (Train/Validation/Test)	T.s/Dialog	W./T.	External Source	W./Source	C.K.S./T.
WoW	22,311 (18,430 / 1,948 / 1,933)	9.1	17.2	1,356,509 (sentences)	30.7	61.2
Holl-E	9,071 (7,228 / 930 / 913)	10.0	15.3	921 (documents)	727.8	57.6
Doc2Dial	4,135 (3,474 / 661 / –)	15.6	14.0	458 (documents)	947.0	73.1

Table 1: Statistics of Datasets. "W./T./C.K.S." is "Words/Turn/Candidate Knowledge Sentences", respectively.

WoW / Holl-E / Doc2Dial	Doc2Dial
Model DA (Utterances)	Human DA (Utterances)
inform(159.6K/76.8K/20.8K)	query condition(6.8K)
question(41.7K/14.3K/5.3K)	respond solution(18.6K)
directive(664/320/167)	respond solution pos(511)
commissive(39/10/18)	respond solution neg(407)

Table 2: Statistics of the model/human-annotated DAs.

3.2 DA Tagger

We train a DA tagger with a BERT-base model. Following DailyDialog (Li et al., 2017), we select four DA labels in ISO standard, which are "inform/question/directive/commissive". Then following Mezza et al. (2018), we choose three commonly used public dialogue datasets (Switchboard, AMI, and Maptask) and map the human-annotated DAs in these datasets to the four ISO labels. By keeping the utterances longer than 2 words, we get 208,718 utterances from DailyDialog/Switchboard/AMI/Maptask, the label distribution is {'inform': 136,406, 'question': 39,085, 'directive': 23,283, 'commissive': 9,944}, the average utterance length is 12.5. These data are real human conversations coming from different domains. We randomly split these data into train and validation (9:1) and train the DA tagger to learn the common dialogue patterns in these utterances. The input to the DA tagger is utterances and the output is the corresponding DA labels[4]. To test the capability of the DA tagger, we randomly select 200 utterances from WoW and 100 utterances from Holl-E. We mix these utterances and manually annotate them with the four ISO DA labels. After training, the tagger achieves 93.7% accuracy on these 300 utterances. We also test the tagger on Doc2Dial. There are 4 different DAs for agent in Doc2Dial: "query condition"/"respond solution"/"respond solution positive"/"respond solution negative". We select the 6,785 utterances with "query conditions" DA for testing. The DA tagger assigns the "question" label to 90.9% of them. These testing results show that the DA tagger is well-trained and could be used for our experiment. We use the tagger to annotate WoW/Holl-E/Doc2Dial. The statistics of the model/human-annotated DAs are shown in Table 2. It is reasonable that the "inform/question" labels account for the vast majority because the DGD task is mainly to consult and provide information. More details about DA tagger are in Appendix B.

3.3 Baselines

We compare with the following baselines: (1) Transformer Memory Network (**TMN**) (Dinan et al., 2018) uses Transformer structure for KS and is introduced along with WoW dataset. (2) Sequential Knowledge Transformer (**SKT**) (Kim et al., 2020) uses BERT as encoder and selects knowledge with a sequential latent variable model. (3) Dual Knowledge Interaction Network (**DukeNet**) (Meng et al., 2020) is a state-of-the-art model which uses BERT as encoder and proposes a knowledge shifter and tracker module for KS. (4) **KnowledGPT** is a state-of-the-art DGD model proposed by Zhao et al. (2020). It uses BERT and GPT-2 to jointly optimize knowledge selection and response generation[5].

We choose these baselines for two reasons: 1) they are all explicit KS models and SKT/DukeNet/KnowledGPT/DAKS all employ BERT as encoder, so we can fairly compare the KS accuracy between them[6]; 2) KnowledGPT/DAKS both use GPT-2 as generation module, we can fairly compare the generation quality between them. We also present several different settings of DAKS. **DAKS(-DA)** is the model without DA as input, it only take **K**, **C** to predict K_i and generate response with K_i and **C**. **DAKS(-$\mathcal{L}_{DA}$)** is a prediction module that only leverages DA for KS but does not predict the next DA, its results can show the effect of $\mathcal{L}_{DA}$. **DAKS(GT)** is a GPT-2 model using dialogue history, ground truth knowledge sentence, and ground truth DA as input, it can be seen as the generation upper bound of our method.

3.4 Implementation Details

The setting of the baseline models follows the papers that proposed them, please refer to each paper for details. Our implementations of BERT-Base

[4] We use the last BERT layer's representation of the first word (a special token <CLS>) to predict the DA label.

[5] The code link is https://github.com/zhaoxlpku/KnowledGPT. We only use the WoW dataset to compare with this work since the authors only provide the evaluation code for WoW.

[6] RoBERTa (Liu et al., 2019a) or ELECTRA (Clark et al., 2020) can perform better than BERT in our experiments.

and GPT-2-medium are based on the public Pytorch implementation[7]. During fine-tuning, we truncated the length of the dialogue context to 60 tokens and maximum input length to 512 tokens. The maximum predicted span length is set to 90 words. In RG, the beam-search size is 5. All models are learned with Adam optimizer with β_1 = 0.9 and β_2 = 0.999. We use a single Tesla v100s GPU with 32gb memory to conduct experiments, the batch size is 4 for all datasets. The fine-tuning epochs are 4 for the prediction module and 3 for the generation module. The DA tagger is trained for 5 epochs.

3.5 Evaluation Metrics

We use the following automatic evaluation metrics employed by the baselines. For KS, we use Hits@1 (Dinan et al., 2018) to measure the KS and DA prediction accuracy. For RG, we use perplexity (PPL), unigram F1 (Dinan et al., 2018)[8], BLEU-4 (Papineni et al., 2002) of the ground-truth responses. Lower PPL and higher Hits@1/F1/BLEU-4 mean better performance. We recruit 3 professional researchers[9] for manual evaluation. We randomly select 50/50 dialogue samples from the WoW unseen/Doc2Dial validation sets, respectively. The generated responses to these samples are presented to the annotators accompanied with their corresponding dialogue history (3 turns) and external knowledge. The responses from different models are shuffled so the annotators do not know which model the response is coming from. Following Zhao et al. (2020), we only provides the ground-truth knowledge sentences and ask the annotators to judge the quality of the responses from three aspects: *Fluency*, *Context Coherence* and *Knowledge Relevance*. The annotators assign a score in {0:bad, 1:fair, 2:good} to each response for each aspect. The agreement among the annotators is measured via Fleiss' kappa (Fleiss and Joseph, 1971).

4 Experimental Results

4.1 Knowledge Selection (KS)

Table 3 shows the KS accuracy results of all models. Benefiting from the representation ability of BERT, the SKT and DukeNet have fairly close performance and both outperform the TMN. KnowledGPT also leverages BERT as an encoder and uses LSTM to sequentially select knowledge sentences. It outperforms DukeNet by 1.6/5.8 on WoW Test seen/unseen. Our DAKS model achieves the new state-of-the-art results, it outperforms the strong KnowledGPT 2.7/4.3 on WoW Test seen/unseen, respectively. On the Holl-E dataset, DAKS outperforms DukeNet 8.4/7.7 on Test single/multi-reference, respectively. On the Doc2Dial validation set, DAKS surpasses DukeNet around 7.6/9.7 with Model/Human-annotated DAs, respectively. Our method surpasses the strong BERT-based models and shows significant advantages in selecting knowledge sentences.

Models	Model-annotated		Model/Human
	WoW seen/unseen	Holl-E single/multi	Doc2Dial validation
TMN	21.6 / 12.1	22.7 / 32.2	43.1
SKT	26.8 / 18.3	29.2 / 38.3	49.5
Dk.Net	26.4 / 19.6	30.0 / 40.3	49.7
K.GPT	28.0 / 25.4	– – / – –	– –
DAKS	__30.7*/29.7*__	38.4*/48.0*	__57.3*/59.4*__
(-DA)	29.4*/29.0*	__39.1*/48.9*__	56.7*
(-$\mathcal{L}_{DA}$)	_31.6*/30.5*_	_38.7*/49.9*_	_58.3*/59.1*_

Table 3: KS results (Hits@1) on the WoW Test seen/unseen, Holl-E Test single/multi-reference and Doc2Dial validation sets. "K.GPT"/"Dk.Net" stands for "KnowledGPT"/"DukeNet", respectively. DukeNet model is the base model to do the significant test for our models (* means statistically significant with p<0.01).

4.1.1 Ablation Study in KS

In Table 3, DAKS(-DA) has a lower performance than DAKS on WoW and Doc2Dial, which shows that DAKS benefit from DA when performing KS. The reason why leveraging DA helps KS can also be explained by the dialogue example in Figure 1. When the prediction module predicts a "query condition" DA for the second turn, it means the model needs to clarify some pre-condition before it can choose a final answer for the first question. When the prediction module predicts a "respond solution" DA, it means the model is confident to predict the final answer for the first question. Different DAs can entail different KS results. The overall KS accuracy improvement verifies DAKS can leverage DA to find proper knowledge.

On the other hand, (-DA) outperforms DAKS on Holl-E. The reason lies in the dialogue mode. When constructing dialogues in Holl-E, the workers acting the role of agents were asked to only

[7] The link is https://github.com/huggingface/transformers. The hyper-parameters which are not introduced in this section follow the original implementation in the link.

[8] The code link is https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py.

[9] They all focus on natural language processing (NLP) and dialogue systems (DS) and have published papers at international conferences related to NLP and DS.

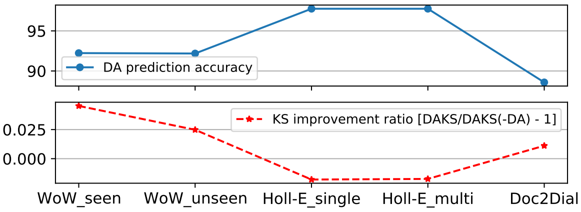


Figure 3: Comparison of DA prediction accuracy and KS improvement ratio when using model-annotated DA.

add a few words before or after the selected knowledge sentence to construct a response. Hence almost all agent turns are providing information and own "inform" DA labels, which entails a less natural dialogue mode compared to WoW/Doc2Dial[10]. Therefore, the guidance of model-annotated DAs is weaker on Holl-E. To verify our conjecture that model-annotated DA *works better on a more natural dialogue*, we compare the model-annotated DA prediction accuracy and KS improving rate (DAKS/(-DA) - 1) in Figure 3. We can see that the DA prediction accuracy of Holl-E is the highest. This means the DA mode in Holl-E is the simplest and easy to predict. However, the KS improving rate of Holl-E is the lowest, which means DAKS can not leverage DA to guide KS in this simple and unnatural dialogue mode. Similarly, Doc2Dial is constructed under pre-defined human-annotated DAs. The workers need to consider the pre-given DA when constructing the dialogue. This restriction makes the dialogues in Doc2Dial less natural than WoW. in Figure 3, the higher improvement ratio of WoW than Doc2Dial again verifies our conjecture.

The results of (-$\mathcal{L}_{DA}$) in Table 3 reflect how much reduction the multi-task learning schema caused on KS. (-$\mathcal{L}_{DA}$) outperforms DAKS on most data except Doc2Dial with human-annotated DA. The results show that adding DA (model-annotated) prediction loss reduces the KS accuracy. However, when leveraging real human-annotated DA, the multi-task learning schema improves the KS accuracy. This comparison shows that 1) the human-annotated DA is more powerful than model-annotated ones in multi-task learning schema; 2) it is possible to balance the multi-task learning schema and the KS performance by eliminating the gap between model and human annotated DAs.

Models	PPL	F1(%)	BLEU-4(%)
WoW Test seen/unseen (model-annotated DA)			
TMN	66.5/103.6	15.9/14.3	1.35/0.43
SKT	52.0/81.4	19.3/16.1	1.76/1.05
DukeNet	52.0/79.3	19.4/17.2	2.43/1.68
K.GPT	__19.2__/22.3	__22.0/20.5__	2.34/1.81
DAKS	20.5*/__20.0__*	20.6*/__20.5__*	__2.59__*/__2.40__*
(-DA)	20.9*/21.1*	19.6*/19.3*	2.53*/2.31
(GT)	_8.9*/ 9.4*_	_29.2*/ 28.8*_	_4.90*/ 4.80*_
Holl-E Test single/multi reference (model-annotated DA)			
TMN	66.5/90.1	15.9/14.1	6.77/8.98
SKT	48.9/28.5	29.8/36.5	17.81/24.69
DukeNet	48.2/27.8	29.3/36.4	19.15/26.83
DAKS	16.6*/__11.2__*	38.9/__45.2__*	__29.80__*/__36.61__*
(-DA)	__15.8__*/11.3	__39.4__*/45.0*	28.51*/35.38*
(GT)	_2.3*/2.3*_	_76.6*/76.6*_	_72.80*/72.80*_
Doc2Dial validation (model/human-annotated DA)			
DukeNet	30.6	39.6	19.45
DAKS	__5.1__*/__4.9__*	__45.9__*/__46.3__*	__25.21*/25.20*__
(-DA)	5.3*	44.5*	24.65
(GT)	_3.8*/3.1*_	_51.1*/53.9*_	_28.01*/30.80*_

Table 4: RG experimental results on the WoW Test seen/unseen, Holl-E Test single/multi-reference, and Doc2Dial validation sets. DukeNet is the base model to do the significant test for our models, values with * mean statistically significant with p<0.01. "K.GPT" is short for "KnowledGPT".

4.1.2 Human/Model-annotated DA in KS

In Table 3, DAKS using model-annotated DAs has a comparable performance with using human-annotated DAs. However, we still notice that there is a 2.1 KS accuracy gap between using human and model annotated DAs. Eliminating this gap requires further research. Since we only annotate 4 classes of DA and the label distribution is unbalanced. Future research could pay attention to more balanced/kinds of DAs, and multi-label DAs. More details about the KS are in Appendix C, D, and E.

4.2 Response Generation (RG)

Table 4 shows **automatic evaluation** results of RG[11]. For baseline models, benefiting from the knowledge selection accuracy, SKT and DuckNet outperform the TMN. KnowledGPT is the state-of-the-art model in RG and outperforms SKT and DukeNet on most metrics. The reason includes: 1) KnowledGPT is more accurate in KS than SKT/DukeNet; 2) The knowledge packed in the parameters of GPT-2 helps the generation. Our DAKS model achieves new state-of-the-art performance on most metrics in all datasets. Especially when comparing DAKS to DukeNet on Holl-E and Doc2Dial. However, DAKS has a lower F1 compared to KnowledGPT on WoW. It shows that the

[10] In Table 2, the "inform" label ratio is 79.0%/79.1%/84.0% for WoW/Doc2Dial/Holl-E, respectively. The higher ratio of "inform" means more unbalanced DA labels and more simple and unnatural dialogue mode. It is another evidence that the naturalness of three datasets is (WoW/Doc2Dial)>Holl-E.

[11] More automatic metrics about the RG are in Appendix F.

Models	Flu.	Coh.	Rel.	Kappa
WoW Test unseen (model-annotated DA)				
K.GPT	1.67	1.50	1.61	0.66
DAKS	**1.68**	**1.61**	**1.63**	**0.71**
(-DA)	1.66	1.55	1.60	**0.71**
Doc2Dial validation (model/human-annotated DA)				
Dk.Net	1.63	1.42	1.53	0.62
DAKS	**1.68/1.72**	**1.60/1.63**	**1.65/1.66**	**0.68/0.70**
(-DA)	1.67	1.57	1.59	0.67

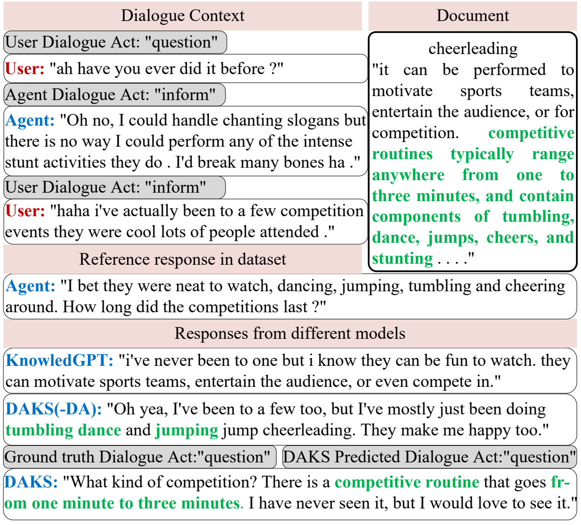
Table 5: Manual evaluation on the WoW Test unseen and Doc2Dial validation sets. "Flu."/"Coh."/"Rel."/ "K.GPT"/"Dk.Net" means "Fluency"/"Context Coherence"/"Knowledge Relevance"/"KnowledGPT"/"DukeNet", respectively.

advantage of DAKS on response generation is not as obvious as it achieved on KS accuracy. This indicates 1) the RG is a harder task than KS and has more influencing factors that need to be considered, higher KS accuracy does not necessarily guarantee a better performance; 2) the two separately trained modules of DAKS are inferior to the joint training methods in KnowledGPT. We take jointly training as future work; 3) the automatic evaluation metrics alone may not be sufficient to reflect the dialogue quality, so manual evaluations are needed.

Table 5 shows **manual evaluation** results of RG. We compare DukeNet and KnowledGPT with DAKS. All models are compared on Fluency / Context Coherence / Knowledge Relevance. DAKS is better than KnowledGPT on WoW Test unseen set and better than DukeNet on Doc2Dial validation set. The results are consistent with automatic evaluations. The overall inter-rater agreement measured by Fliess' Kappa ranges from 0.62 to 0.71, indicating substantial agreement among the annotators. The manual evaluation further verifies that DAKS is a new state-of-the-art DGD model.

4.2.1 Ablation Study in RG

Table 4 and 5 also show the ablation Study results in RG. Both automatic and manual evaluation show that DAKS outperforms DAKS(-DA) on WoW and Doc2Dial. Comparing PPL/F1/BLEU-4 of DAKS to that of DAKS(-DA), the improving ratio on the WoW Test unseen is 5.2%/6.2%/3.9%, the improving ratio on the Doc2Dial validation is 3.8%/3.1%/2.3%. In the analysis of KS, we outlined that the dialogue in WoW is more natural than Doc2Dial. The comparison between DAKS and DAKS(-DA) in RG further confirms that DA information is more helpful when the dialogue is more natural. On the other hand, the big gap between DAKS and DAKS(GT) in Table 4 indicates that we are far from exploiting the full potential of current pre-trained models in the DGD task.


Dialogue Context

User Dialogue Act: "question"

User: "ah have you ever did it before ?"

Agent Dialogue Act: "inform"

Agent: "Oh no, I could handle chanting slogans but there is no way I could perform any of the intense stunt activities they do . I'd break many bones ha ."

User Dialogue Act: "inform"

User: "haha i've actually been to a few competition events they were cool lots of people attended ."

Document

cheerleading

"it can be performed to motivate sports teams, entertain the audience, or for competition. **competitive routines typically range anywhere from one to three minutes, and contain components of tumbling, dance, jumps, cheers, and stunting** . . . ."

Reference response in dataset

Agent: "I bet they were neat to watch, dancing, jumping, tumbling and cheering around. How long did the competitions last ?"

Responses from different models

KnowledGPT: "i've never been to one but i know they can be fun to watch. they can motivate sports teams, entertain the audience, or even compete in."

DAKS(-DA): "Oh yea, I've been to a few too, but I've mostly just been doing **tumbling dance** and **jumping** jump cheerleading. They make me happy too."

Ground truth Dialogue Act:"question"

DAKS Predicted Dialogue Act:"question"

DAKS: "What kind of competition? There is a **competitive routine** that goes **from one minute to three minutes.** I have never seen it, but I would love to see it."

Figure 4: Dialogue case from WoW Test seen.

4.2.2 Human/Model-annotated DA in RG

Table 4 and 5 report the results when DAKS using model/human-annotated DAs on Doc2Dial validation set. In general, leveraging the human-annotated DAs is better than using model-annotated DAs. However, their performances are very close. Similar to the analysis in KS, the generation experiments again verify that the model-annotated DAs can play a comparable role to the human-annotated ones in our model.

4.2.3 Case Study

In Figure 4, we randomly select a dialogue case in WoW Test seen set. The dialogue context, corresponding DAs, the golden response in the dataset, and part of the documents are presented. We show the responses of the three best performance models: KnowledGPT, DAKS(-DA), and DAKS. The ground-truth knowledge sentence in the document is bold and green. KnowledGPT fails to select the correct knowledge sentence while both DAKS(-DA) and DAKS succeed. However, the response from DAKS(-DA) is not context coherence, which shows that the RG is a challenging task even when the model selects the correct knowledge sentence. In contrast, with the assistant of the correctly predicted DA "question", the DAKS not only generates a fluent response with strong context coherence but also gives an informative response starting with a context-related question. More cases and analysis are shown in Appendix A.

5 Related Work

5.1 Document-grounded Dialogue (DGD)

The **knowledge selection** (KS) (Ren et al., 2019; Meng et al., 2020; Kim et al., 2020) in DGD task is to select dialogue-related information from the given document. In terms of the sampling mechanism that selects the most relevant text fragments, KS can be categorized into **implicit selection** and **explicit selection**. Early implicit KS models(Zhou et al., 2018b; Moghe et al., 2018) usually employed the attentional Seq-to-Seq memory network to encode the dialogue and document respectively into a vector or a sequence of vectors as model memory. Then they used the decoder hidden state as a query to attentively read the memory. Some later work employed matching operations between dialogue and document before constructing the memory. Many of them borrowed idea from the MRC task (Meng et al., 2019; Ren et al., 2019). Instead of predicting a span, they took advantage of cross attention and matching matrix to generate a document-length memory for KS. However, the implicit methods are difficult to trace the knowledge they used. As a consequence, some scoring and sampling mechanisms were proposed to select fragments (usually a sentence) from the document, this process can be defined as the explicit selection mechanism and they can measure the accuracy of KS (Kim et al., 2020; Zhao et al., 2020) if the ground-truth labels exist. The scoring methods (dot-product attention (Lian et al., 2019), TF-IDF similarity (Gopalakrishnan et al., 2019), K-Nearest-Neighbors (Fan et al., 2020), etc.) in these models aim to match dialogue context with each pre-segmented text piece respectively and generate a preference distribution over them (Meng et al., 2020; Ahn et al., 2020). Besides dialogue context, some researchers attempted to utilize supplementary information to facilitate explicit KS. Liu et al. (2019b) aligned each knowledge sentence with a vertex in a knowledge graph and used Reinforcement Learning to train the reasoning policy over the graph. Zheng et al. (2020) argued that the difference between the knowledge sentence selected at different dialogue turns provided potential clues for KS. Our KS method is based on an MRC (Devlin et al., 2019) model and belongs to the explicit KS method with DA as supplementary information.

After KS, the DGD models use the selected knowledge for informative **response generation** (RG). Li et al. (2019) used a deliberation decoder to improve context coherence and knowledge correctness. Wang et al. (2019) investigated three different approaches (Concatenate, Alternate and Interleave) to combine context and knowledge encodings into a Transformer type decoder. Prabhumoye et al. (2021) added cross-attention layers into a BART (Lewis et al., 2020) model to integrate context and document information into decoder and achieved state-of-the-art performance. Rashkin et al. (2021) added control tokens to the input sequence of the GPT-2 (Radford et al., 2019) model for certain semantic features. Our RG module is also a GPT-2 model which leverages DA as the control signal.

5.2 Dialogue Act (DA)

Dialogue Act is often used rather loosely in the sense of ‘speech act used in dialogue’ (Bunt et al., 2010) or ’the intention or the function of an utterance in dialogues’ (Kawano et al., 2019). Researchers have set up an ISO standard for DA in open-domain dialogue (Bunt et al., 2010, 2020) and this standard has been applied in many studies of dialogue systems (Li et al., 2017; Hedayatnia et al., 2020). However, only a few DGD datasets contain human-annotated DAs (Majumder et al., 2020). Feng et al. (2020) showed human-annotated DAs were useful for KS. Hedayatnia et al. (2020) used an automatic tagger to annotate DA but only used DA for RG policy. In this paper, we try to expand the research in this field. First, we train a DA tagger following the ISO standard. Second, we leverage DA for both KS and RG and compare the human/model-annotated DAs in our model.

6 Conclusion

We propose a DAKS framework that first selects knowledge and predicts the next DA then uses the selected sentence and the predicted DA for RG. We trained an ISO standard DA tagger and annotated three public DGD datasets with the tagger for our experiments. Experimental results show that: 1) leveraging DA can improve KS accuracy and RG quality, especially in a natural dialogue such as WoW; 2) using model-annotated DAs is comparable with using the expansive human-annotated ones. Our findings have a positive effect on exploiting DA information in dialogue research. In the future, we would like to study how to eliminate the gap between model-annotated and human-annotated DAs. Specifically, we focus on the unbalanced/multiple DA labels problems in dialogue data.

References

Yeonchan Ahn, Sang-Goo Lee, and Jaehui Park. 2020. Exploiting text matching techniques for knowledge-grounded conversation. *IEEE Access*, 8:126201–126214.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Kôiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David R. Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Chengyu Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 549–558. European Language Resources Association.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*. OpenReview.net.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2020. Augmenting transformers with knn-based composite memory for dialogue. *CoRR*, abs/2004.12744.

Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8118–8128. Association for Computational Linguistics.

Fleiss and L. Joseph. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

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 *AAAI*, pages 5110–5117. AAAI Press.

John J. Godfrey, Edward Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '92, San Francisco, California, USA, March 23-26, 1992*, pages 517–520. IEEE Computer Society.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. pages 1891–1895. ISCA.

Behnam Hedayatnia, Seokhwan Kim, Yang Liu, Karthik Gopalakrishnan, Mihail Eric, and Dilek Hakkani-Tür. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems. *CoRR*, abs/2005.12529.

Seiya Kawano, Koichiro Yoshino, and Satoshi Nakamura. 2019. Neural conversation model controllable by given dialogue act based on adversarial learning and label-aware objective. In *INLG*, pages 198–207. Association for Computational Linguistics.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In

HLT-NAACL, pages 110–119. The Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *ACL (1)*, pages 12–21. 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 *IJCAI*, pages 5081–5087. ijcai.org.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019b. Knowledge aware conversation generation with reasoning on augmented graph. *CoRR*, abs/1903.10245.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian J. McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8129–8141. Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Refnet: A reference-aware network for background based conversation. *CoRR*, abs/1908.06449.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1151–1160. ACM.

Stefano Mezza, Alessandra Cervone, Evgeny A. Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. Iso-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3539–3551. Association for Computational Linguistics.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *EMNLP*, pages 2322–2332. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W. Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4274–4287. Association for Computational Linguistics.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *ACL (1)*, pages 5427–5436. Association for Computational Linguistics.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392. The Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 704–718. Association for Computational Linguistics.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. *CoRR*, abs/1908.09528.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Xinyi Wang, Jason Weston, Michael Auli, and Yacine Jernite. 2019. Improving conditioning in context-aware sequence to sequence models. *CoRR*, abs/1911.09728.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7281–7288. AAAI Press.

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 2020, Online, November 16-20, 2020*, pages 3377–3390. Association for Computational Linguistics.

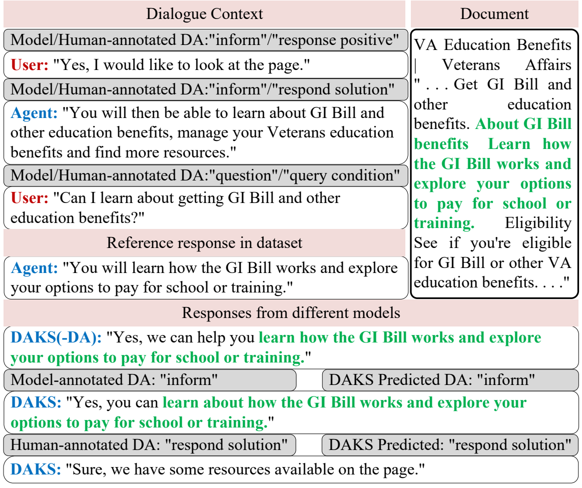
Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 115–125. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629. ijcai.org.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018b. A dataset for document grounded conversations. In *EMNLP*, pages 708–713. Association for Computational Linguistics.

A Case Study Appendix

We randomly select two dialogue cases in the Doc2Dial Validation set to compare leveraging model/human-annotated DAs. In Figure 5, DAKS(-DA) and DAKS (using model-annotated DA) predict the correct knowledge sentence. They generate very similar responses to the reference response in the dataset. DAKS (using human-annotated DA) selects the wrong knowledge and entails a less informative reply. This case verifies that using model-annotated or human-annotated DA could have a different impact on KS and RG. Although experiments show that human-annotated DA is more powerful, there are still cases that model-annotated DA works better. In Figure 6, DAKS(-DA) fails in KS while DAKS with both model/human-annotated DAs succeed. However, both responses from DAKS have logistic mistakes since they use only the KS results as external knowledge and ignore the pre-condition knowledge "If you don t want to transfer the registration" in the document. In contrast, DAKS(-DA) gives a reasonable response even without the ground-truth knowledge. This case shows that although adding DA information help the KS, the small inconsistency between the ground-truth knowledge and the reference response in the dataset could still harm the RG performance of our model. In conclusion, the case studies verify that DA information can help to perform KS and utilize the selected information for response generation. However, the quality of response is easily influenced by a number of factors, such as the bias



Dialogue Context

Model/Human-annotated DA:"inform"/"response positive"

User: "Yes, I would like to look at the page."

Model/Human-annotated DA:"inform"/"respond solution"

Agent: "You will then be able to learn about GI Bill and other education benefits, manage your Veterans education benefits and find more resources."

Model/Human-annotated DA:"question"/"query condition"

User: "Can I learn about getting GI Bill and other education benefits?"

Reference response in dataset

Agent: "You will learn how the GI Bill works and explore your options to pay for school or training."

Document

VA Education Benefits | Veterans Affairs " . . . Get GI Bill and other education benefits. **About GI Bill benefits Learn how the GI Bill works and explore your options to pay for school or training.** Eligibility See if you're eligible for GI Bill or other VA education benefits. . . ."

Responses from different models

DAKS(-DA): "Yes, we can help you **learn how the GI Bill works and explore your options to pay for school or training.**"

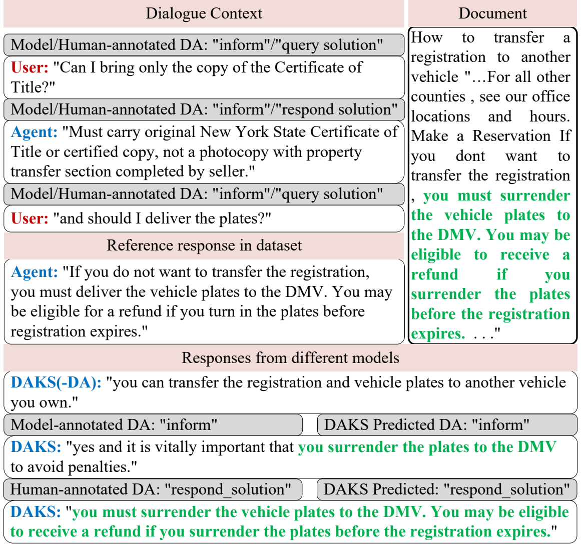
Model-annotated DA: "inform" | DAKS Predicted DA: "inform"

DAKS: "Yes, you can **learn about how the GI Bill works and explore your options to pay for school or training.**"

Human-annotated DA: "respond solution" | DAKS Predicted: "respond solution"

DAKS: "Sure, we have some resources available on the page."

Figure 5: The 1st case from Doc2Dial Validation set.



Dialogue Context

Model/Human-annotated DA: "inform"/"query solution"

User: "Can I bring only the copy of the Certificate of Title?"

Model/Human-annotated DA: "inform"/"respond solution"

Agent: "Must carry original New York State Certificate of Title or certified copy, not a photocopy with property transfer section completed by seller."

Model/Human-annotated DA: "inform"/"query solution"

User: "and should I deliver the plates?"

Reference response in dataset

Agent: "If you do not want to transfer the registration, you must deliver the vehicle plates to the DMV. You may be eligible for a refund if you turn in the plates before registration expires."

Document

How to transfer a registration to another vehicle "...For all other counties , see our office locations and hours. Make a Reservation If you dont want to transfer the registration , **you must surrender the vehicle plates to the DMV. You may be eligible to receive a refund if you surrender the plates before the registration expires.** . . ."

Responses from different models

DAKS(-DA): "you can transfer the registration and vehicle plates to another vehicle you own."

Model-annotated DA: "inform" | DAKS Predicted DA: "inform"

DAKS: "yes and it is vitally important that **you surrender the plates to the DMV** to avoid penalties."

Human-annotated DA: "respond_solution" | DAKS Predicted: "respond_solution"

DAKS: "**you must surrender the vehicle plates to the DMV. You may be eligible to receive a refund if you surrender the plates before the registration expires.**"

Figure 6: The 2nd case from Doc2Dial Validation set.

Dataset	'inform' DA	'question' DA	'directive' DA	'commissive' DA	total	A.L.U.
DailyDialog	45,469	28,994	17,267	9,296	101,026	13.8
Switchboard	82,176	9,097	708	99	92,080	10.6
AMI	7,231	0	3,187	549	10,967	16.2
Maptask	1,530	994	2,121	0	4,645	10.7
Total-for-training	136,406	39,085	23,283	9,944	208,718	12.5
WoW-for-testing	176	15	8	1	200	16.8
Holl-E-for-testing	81	16	2	1	100	15.4
Total-for-testing	257	31	10	2	300	12.5

Table 6: Dataset statistics with human-annotated DAs for training and testing the DA tagger. "A.L.U." is short for average length per utterance. The testing data are manually annotated by us.

of the data itself.

B DA Tagger Appendix

We use four datasets to construct the training data for the DA tagger. The statistic of the training data is shown in Table 6. The DailyDialog is dialogues in daily communication way and covers various topics about daily life. The Switchboard corpus is a dataset of transcribed open-domain telephone conversations. The AMI contains transcriptions of meeting recordings of the European-funded AMI project, a consortium dedicated to the research and development of technology. Maptask is dialogues involving two participants, one with a route marked map which must instruct the other to draw the same route on an empty map. These datasets contain human-annotated DAs related to ISO standards. We map the DAs in these datasets to 'inform'/'question'/'directive'/'commissive'. For example, utterances with 'suggest'/'request' DAs are mapped to 'directive' DA. Please refer to Mezza et al. (2018) for more details. We process the data based on the code released by Mezza et al. (2018). Noticed that we only keep the utterances longer than 2 words[12]. Table 6 also presents the testing data we manually annotated from WoW and Holl-E. The DA tagger achieves 88.7%/100%/100%/50%/93.7% accuracy for "inform"/"question"/"directive"/"commissive"/"total" of the testing data, respectively. The overall accuracy of 93.7% shows that the DA tagger is well-trained and can be used for our experiments.

C Span-Revision Appendix

The prediction module in DAKS selects a span in a document. When the start and end positions are within or across sentences, we expand, move, or truncate a span into a whole sentence. We illustrate the Span Revision method in Figure 7, all 4 cases select sentence 2 as the knowledge after revision.

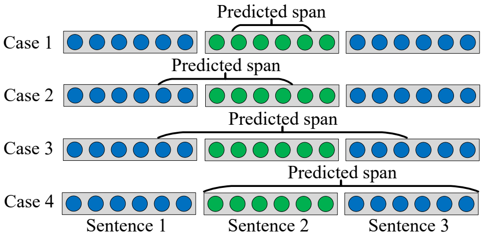


Figure 7: Dialogue case from WoW Test unseen.

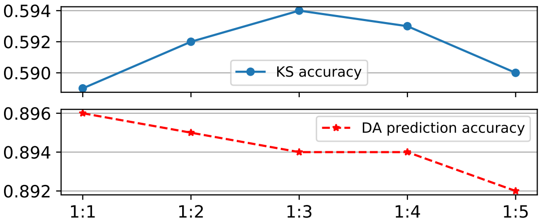


Figure 8: Weight coefficient α in the prediction module.

D Weight-coefficient Appendix

There is a weight coefficient α in the prediction module to balance the $\mathcal{L}_{DA}$ and $\mathcal{L}_{KS}$. Figure 8 shows the experiments on Doc2Dial validation set to determine α. The abscissa represents the values of α:(1-α). The ordinate shows the KS accuracy and the DA prediction accuracy, we choose α = 0.25 (1:3) when the KS accuracy is the highest. A similar trend of α is observed on WoW and Holl-E.

E Language Model for KS Appendix

We use BERT-base model as the prediction module for a fair comparison with the baselines. When using RoBERTa-base or ELECTRA-base instead of BERT-base, we can get a higher performance, Table 7 shows the KS results on the Doc2Dial validation set. It shows that leveraging DA is useful with different pre-trained models and RoBERTa is the strongest among the three models.

[12] For instance, there are 223,607 utterances in Switchboard, we only keep 98,264 of them.

Models	Model/Human-annotated DA
DAKS(BERT)	57.3 / 59.4
DAKS(RoBERTa)	**59.4*** / **61.3***
DAKS(ELECTRA)	58.8* / 60.5*

Table 7: Knowledge selection results (Hits@1) on the Doc2Dial Validation set. We take the DAKS(BERT) model as the base model to do the significant test, values with * mean statistically significant with p<0.01.

Models	ROUGE-L	Dist-1(%)	Dist-2(%)
WoW Test seen/unseen (model-annotated DA)			
TMN	15.7/14.4	3.4/2.6	10.5/6.8
SKT	17.6/16.1	7.6/3.1	27.3/16.1
DukeNet	18.5/17.1	8.6/5.0	28.4/18.0
K.GPT	18.8/17.5	10.0/9.5	30.2/24.5
DAKS	**18.9*/17.8***	**10.5*/9.9***	**31.6*/26.8***
(-DA)	18.7*/17.6*	10.2*/9.6	30.5*/25.9*

Table 8: ROUGE-L and Distinct results on the WoW Test seen/unseen sets. We take the DukeNet model as the base model to do the significant test, values with * mean statistically significant with p<0.05. "K.GPT" is short for "KnowledGPT".

F More Metrics for RG Appendix

We provide ROUGE-L (Lin, 2004) and Distinct-(1/2) (Li et al., 2016) for WoW in Table 8. ROUGE is based on the calculation of the recall rate of the common sub-sequence of generating response and the real one. Distinct measures the diversity of responses by calculating the proportion of distinct n-grams in the total number of n-grams. The higher values of them mean a better generation quality. Table 8 shows that DAKS outperforms baseline models on these two metrics, which further verifies the superior of our model.

G Ethical Statement Appendix

The datasets we used in this paper are all English data from previously published papers and are all publicly available. We did not make any changes to the data so there are no data ethical problems in this research.