

Goal Oriented Dialogue Systems

Saley Vishal Vivek

Indian Institute of Technology, Delhi

csz208845@cse.iitd.ac.in

March 4, 2022

Overview

- ① A Brief Introduction to Task-Oriented Dialogue Systems
- ② Transformer based models for Modular TODS
 - SimpleTOD
 - SOLOIST
- ③ Chit-chat and Task-Oriented Dialogue
 - ACCENTOR

Modular Design for TOD

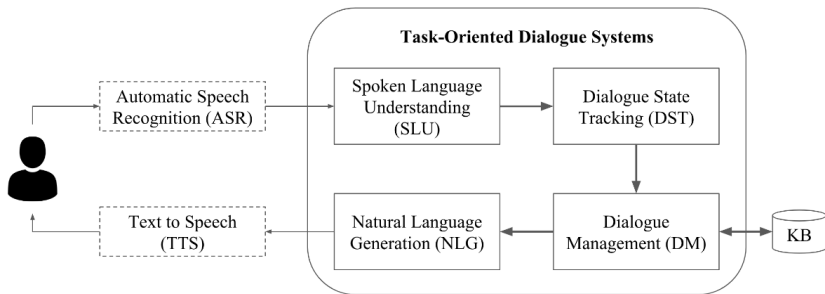


Figure: Image credits: [Wu19]

An Annotated Example (NLU)

U: i'm looking for an **expensive** restaurant in the **center** of town.

S: there are 33 restaurants [...]. would you like to narrow your search by type of food?

U: i **do not care** about food. surprise me.

S: **fitzbillies restaurant** serves british food, [...]. would you like to hear about any others? [...].

U: **that sounds fine**. can you book it please and get me the reference number?

S: sure , what day and time would you like and how many people ?

U: i would like a table for **5** at **11:30** on **tuesday**[...]

S: okay, the booking was successful. [...]. is there anything else i can help you with?

U: i'm also looking for a place to stay. it needs [...]**free wifi** and [be] in the **same area as the restaurant**.

Turn	Domain-slot pair	Value	Type	Coreference
0	restaurant-pricerange	expensive	span	
0	restaurant-area	center	span	
1	restaurant-food	<dontcare>	(dontcare)	
2	restaurant-name	fitzbillies	informed	
3	restaurant-people	5	span	
3	restaurant-book_time	11:30	span	
3	restaurant-book_day	tuesday	span	
4	hotel-internet	<true>	(bool)	
5	hotel-area	center	coreference (multiturn)	restaurant-area

Figure: Image credits: [HvNL⁺20]

An Annotated Example (DST)

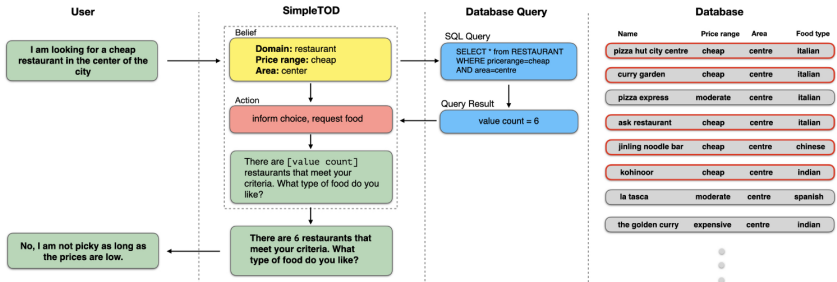


Figure: Image credits: [HAMW⁺20]

Tasks within TODS and Evaluation metrics

Traditionally, each sub-module is developed independently and thus has its own evaluation metrics.

Automatic Evaluations

- For NLU, accuracy is used for intent identification and F1 for slot filling.
- For DST, joint accuracy is used.
- For DM/NLG, inform and success rates, BLEU and Combined
$$= (\text{Inform} + \text{Success}) \times 0.5 + \text{BLEU}$$

Manual Evaluations

- Appropriateness: how useful are the responses for the given dialog context
- Naturalness: how human-like are the predicted responses.

Manual evaluation is time-consuming and is subjective but is critical for TODS.

Challenges in Modular approach and End-to-end TOD systems

Modular approach is popular for commercial deployments but faces following challenges[CLYT17]

- Requires very fine-grained hand-crafted labels and faces difficulties adapting to new domain.
- Information is not/wrongly propagated from one module to another.
- Updating one module may require updating all (inter-dependencies)

An alternate approach is an end-to-end system where model directly interacts with user utterance, database and produces response without using fine-grained labels.

Encoder-Decoder Design for TOD

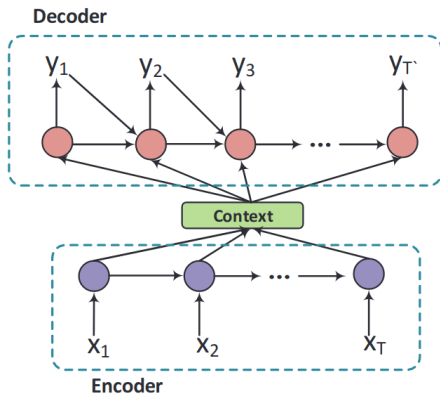


Figure: Image credits: [CLYT17]

A Note on "End-to-End"

"End-to-end" can have different interpretations

- A modular TOD system can be trained in end-to-end manner where signals propagate through all the modules
- A modular TOD system can be evaluated in end-to-end manner where each module consumes output from previous one during evaluations
- A TOD system can be trained end-to-end fashion without fine-grained annotations

We will clarify this point when we discuss some models.

A Simple Language Model for Task-Oriented Dialogue

Ehsan Hosseini-Asl
ehosseiniasl@salesforce.com
Salesforce Research

Bryan McCann
bmccann@salesforce.com
Salesforce Research

Chien-Sheng Wu
wu.jason@salesforce.com
Salesforce Research

Semih Yavuz
syavuz@salesforce.com
Salesforce Research

Richard Socher
rsocher@salesforce.com
Salesforce Research

Goals

SimpleTOD[HAMW⁺20] focuses on following areas

- Train modules in TODS (dialog state tracking, dialog policy and response generation) in unified manner
- Leverage capabilities of pre-trained language models for TODS

Following slides taken from

https://neurips.cc/virtual/2020/protected/poster_e946209592563be0f01c844ab2170f0c.html

SimpleTOD - Input Representation



- We propose recasting task-oriented dialogue as a simple, causal (unidirectional) language modeling task
- We show that such an approach can solve all the sub-tasks in a unified way using multi-task maximum likelihood training
- Dialogue context comprises all previous user/system responses context, $C_t = [U_0, S_0, \dots, U_t]$
- A single training sequence consists of the concatenation of context C_t , belief states B_t , database search results D_t , action decisions A_t , and system response S_t
- A schematic overview of each segment is shown below together with special tokens marking transition points.

$$x^t = [C_t; B_t; D_t; A_t; S_t]$$

C_t	Context	[context] [user] user input [system] system response ... [user] user input [endofcontext]
B_t	Belief State	[belief] domain slot_name value, domain slot_name value, ... [endofbelief]
D_t	DB Search	[db] #_matches, booking_status [endofdb]
A_t	Action	[action] domain action_type slot_name, domain action_type slot_name, ... [endofaction]
S_t	Response	[response] system delexicalized response [endofresponse]

SimpleTOD (Training)

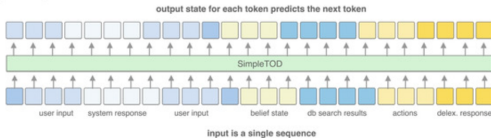


- A single training sequence consists of the concatenation $x^t = [C_t; B_t; D_t; A_t; S_t]$
- This allows us to model the joint probability $p(x)$ over the sequence x^t
- SimpleTOD is optimized by minimizing the negative log likelihood over the joint sequence

$$p(x) = \prod_{i=1}^n p(x_i | x_{<i})$$

$$\mathcal{L}(D) = - \sum_{t=1}^{|D|} \sum_{i=1}^{n_t} \log p_{\theta}(x_i^t | x_{<i}^t)$$

a) training



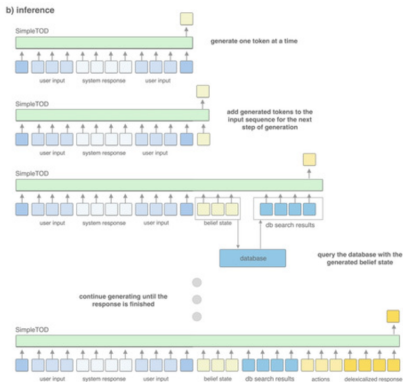
SimpleTOD (Inference)



$$B_t = \text{SimpleTOD}(C_t)$$

$$A_t = \text{SimpleTOD}([C_t, B_t, D_t])$$

$$S_t = \text{SimpleTOD}([C_t, B_t, D_t, A_t])$$



Dialogue State Tracking (DST)



Evaluation of Dialogue State Tracking (DST) on MultiWOZ 2.1 using joint accuracy metric.

- • **TRADE** proposes test label cleaning and recommended by MultiWOZ authors
- † **TripPy** uses label normalization and equivalent matching
- •• **DSTQA** uses the cleaning of TRADE model plus additional accounting for label variants

- SimpleTPDo no label-cleaning
- SimpleTOD* uses label-cleaning proposed by TRADE
- SimpleTOD+ performs cleaning of Type 2 and partial cleaning of Type 4 noisy annotations (proposed in our paper)

Model	Decoder	Context Encoder	Extra Supervision	Joint Accuracy
TRADE*	Generative + Classifier	Bidirectional	-	45.6
DSTQA**	Classifier	Bidirectional	knowledge graph	51.17
DST-Picklist*	Classifier	Bidirectional	-	53.3
SST*	Generative	Bidirectional	schema graph	55.23
TripPy†	Classifier	Bidirectional	action decision	55.3
SimpleTOD ^o	Generative	Unidirectional	-	55.72
SimpleTOD*	Generative	Unidirectional	-	55.76
SimpleTOD ⁺	Generative	Unidirectional	-	57.47

A list of discovered noisy annotations in MultiWOZ 2.1 alongside a cleaned version of the test set are provided

End-to-End Evaluation



Action and response generation uses three metrics.

inform and success rates: designed to capture how well the task was completed.

Inform rate: measures how often the entities provided by the system are correct.

Success rate: refers to how often the system is able to answer all the requested attributes by user.

BLUE score: is used to measure the fluency of the generated responses.

combined score: for action and response generation is computed as $(BLEU + 0.5 * (Inform + Success))$.

Model	Belief State	DB Search	Action	Inform	Success	BLEU	Combined
DAMD+augmentation	generated	oracle	generated	76.3	60.4	16.6	85
SimpleTOD (ours)	generated	oracle	generated	78.1	63.4	16.91	87.66
SimpleTOD (ours)	generated	dynamic	generated	81.4	69.7	16.11	91.66
SimpleTOD (ours)	generated	-	generated	84.4	70.1	15.01	92.26

Table 2: Action and response generation on MultiWOZ 2.0 reveals that SimpleTOD, a single, causal language model, is sufficient to surpass prior work.

Model	Belief State	DB Search	Action	Inform	Success	BLEU	Combined
DAMD+augmentation	oracle	oracle	oracle	95.4	87.2	27.3	118.5
PARG	oracle	oracle	oracle	91.1	78.9	18.8	103.8
SimpleTOD (ours)	oracle	oracle	oracle	93.4	83.2	17.78	106.08
SimpleTOD (ours)	oracle	-	oracle	92.3	85.8	18.61	107.66
HDSA	oracle	oracle	generated	82.9	68.9	23.6	99.5
DAMD+augmentation	oracle	oracle	generated	89.2	77.9	18.6	102.5
ARDM	oracle	oracle	-	87.4	72.8	20.6	100.7
LaRL	oracle	oracle	generated	82.78	79.2	12.8	93.79
SimpleTOD (ours)	oracle	oracle	generated	84	72.8	16.1	94.5
SimpleTOD (ours)	oracle	-	generated	88.9	67.1	16.9	94.9

Table 7: SimpleTOD results on MultiWOZ 2.0 using oracle information.

Comments

Pros

- Captures dependencies between TODS modules
- Leverages GPT-2 pre-trained model
- Robust to noisy samples

Cons/Possible Improvements

- GPT-2 is not trained on dialogue specific data. Pre-training on dialogue data can help.
- SimpleTOD computes likelihood of all the tokens in the sample, even context utterance tokens. This seems to be an overkill.
- No human evaluation (from official reviews)

Above issues are addressed in SOLOIST.

SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching

**Baolin Peng, Chunyuan Li, Jinchao Li
Shahin Shayandeh, Lars Liden, Jianfeng Gao**

Microsoft Research, Redmond, United States

{bapeng, chunyl, jincli, shahins, lars.liden, jfgao}@microsoft.com

Goals

- Few-shot fine-tuning of TODS model over new-domain
- Leverage machine teaching for resource constrained domains
- Building task-bots at scale

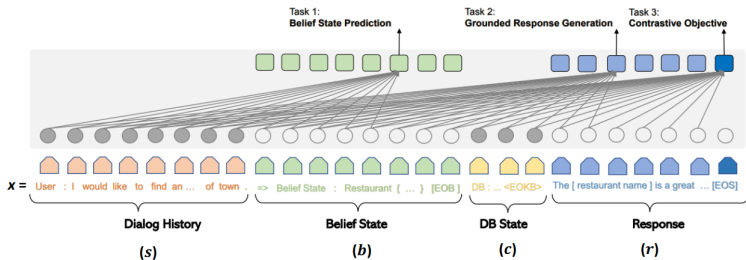
Following slides are taken directly from

<https://d3smihljt9218e.cloudfront.net/lecture/26055/slideshow/33ad01c435de7ea63e5d8273b281c041.pdf>

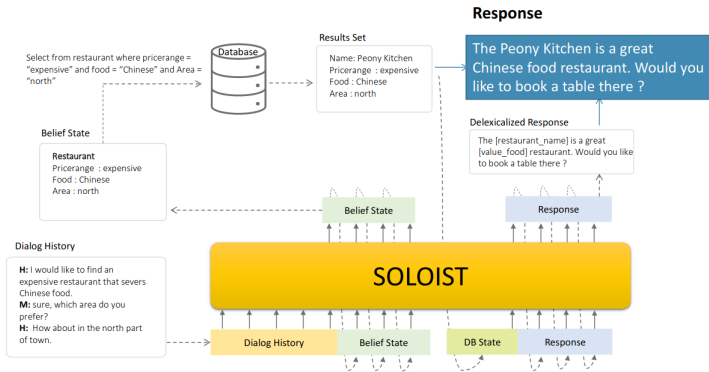
An auto-regressive model for dialog

- Each dialog turn: $\mathbf{x} = (\mathbf{s}, \mathbf{b}, \mathbf{c}, \mathbf{r})$
 - \mathbf{s} : dialog history
 - \mathbf{b} : belief state (user goal detected from \mathbf{s})
 - \mathbf{c} : DB state (retrieved from a DB using \mathbf{b} via APIs)
 - \mathbf{r} : system response
- Learn joint probability using an auto-regressive model
 - $p(\mathbf{x}) = p(\mathbf{r}, \mathbf{c}, \mathbf{b}, \mathbf{s}) = p(\mathbf{s}) p(\mathbf{b}|\mathbf{s}) p(\mathbf{c}|\mathbf{b}, \mathbf{s}) p(\mathbf{r}|\mathbf{c}, \mathbf{b}, \mathbf{s})$
where
 - $p(\mathbf{b}|\mathbf{s})$ -- belief prediction (detecting user goal)
 - $p(\mathbf{c}|\mathbf{b}, \mathbf{s}) = p(\mathbf{c}|\mathbf{b}) = 1$ -- deterministic database lookup
 - $p(\mathbf{r}|\mathbf{c}, \mathbf{b}, \mathbf{s})$ -- grounded response generation

Pre-train SOLOIST via multi-task learning



$$p(x) = p(r, c, b, s) = p(s) p(b|s) p(c|b, s) p(r|c, b, s)$$



Datasets

Name	#Dialog	#Utterance	Avg. Turn	#Domain
<i>task-grounded pre-training:</i>				
Schema	22,825	463,284	20.3	17
Taskmaster	13,215	303,066	22.9	6
<i>fine-tuning:</i>				
MultiWOZ2.0	10,420	71,410	6.9	7
CamRest676	676	2,744	4.1	1
Banking77	-	25,716	-	21
Restaurant-8k	-	8,198	-	1

← Task-grounded pre-training datasets

← Fine-tuning / downstream evaluation datasets

Table 1: Dialog corpora. The datasets in the upper block are used for task-grounded pre-training, and the datasets in the lower block are for fine-tuning.

End-to-End Evaluation Results

Model	Annotations		Evaluation Metrics			
	Belief State	Policy	Inform \uparrow	Success \uparrow	BLEU \uparrow	Combined \uparrow
Sequicity (Lei et al., 2018)	✓	✓	92.30	85.30	21.40	110.20
Sequicity (w/o RL)	✓	✓	94.00	83.40	23.40	112.10
GPT fine-tuning (Budzianowski and Vulić, 2019)			-	86.20	19.20	-
ARDM ¹ (Wu et al., 2019b)			-	87.10	25.20	-
SOLOIST	✓		94.70	87.10	25.50	116.40

¹ARDM is not fully E2E, as it requires a rule-based dialog state tracker.

Table 2: End-to-End evaluation on CamRest676. Results of existing methods are from Wu et al. (2019b).

Model	Annotations		Evaluation Metrics			
	Belief State	Policy	Inform \uparrow	Success \uparrow	BLEU \uparrow	Combined \uparrow
Sequicity (Lei et al., 2018)	✓	✓	66.41	45.32	15.54	71.41
HRED-TS (Peng et al., 2019)	✓	✓	70.00	58.00	17.50	81.50
Structured Fusion (Mehri et al., 2019b)	✓	✓	73.80	58.60	16.90	83.10
DSTC8 Track 1 Winner ¹ (Ham et al., 2020)	✓	✓	73.00	62.40	16.00	83.50
DAMD (Zhang et al., 2020b)	✓	✓	76.40	60.40	16.60	85.00
SOLOIST	✓		85.50	72.90	16.54	95.74

¹The result of DSTC8 Track 1 Winner is produced by adapting their code to our setting.

Table 3: End-to-end evaluation on MultiWOZ.

SOLOIST performs significantly better than competitors in the standard setting.

Adding Chit-Chat to Enhance Task-Oriented Dialogues

Kai Sun^{1*}, **Seungwhan Moon**², **Paul Crook**², **Stephen Roller**³, **Becka Silvert**²,
Bing Liu², **Zhiguang Wang**², **Honglei Liu**², **Eunjoon Cho**², and **Claire Cardie**¹

¹Cornell University

²Facebook, ³Facebook AI Research

✉ ks985@cornell.edu, shanemoon@fb.com

Goal

- To add chit-chat to enhance task-oriented dialogues for better user experience

The Data Problem

- Publicly available datasets either focus purely on chit-chat or on TODS.
- To alleviate this issue, Accentor uses a collaborative data collection strategy
- Accentor first generates (chit-chat) candidates using pre-trained versions of GPT-2 and BlenderBot.
- A model based filtering removes bad candidates.
- Human annotators then classify a candidate as good or bad

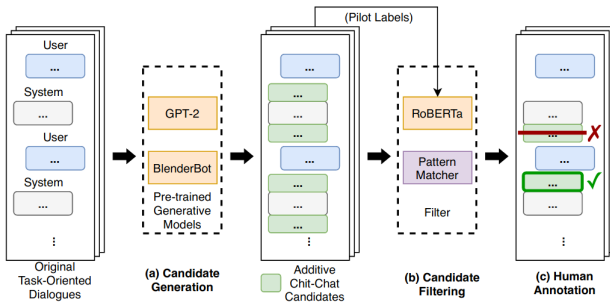


Figure: Image credits: [SMC⁺21]

Filtering Strategy

ACCENTOR collects pilot data for filtering.

- A RoBERTa based classifier
- A set of hand-crafted rules
- Candidate frequency
- Inter-candidate similarity
- Candidate response similarity

Candidates are ranked based on above filters and top-10 are selected for human annotations.

Annotation statistics

Metric	Value
# of candidates	228, 250
# of unique candidates	68, 406
vocabulary size	10, 005
# of distinct 2-grams	59, 259
# of distinct 3-grams	131, 989
# of distinct 4-grams	195, 508
# of distinct 5-grams	239, 278
average length (in tokens)	8.7
<hr/>	
# of good candidates (%)	94, 600 (41.4)
◇ social	86, 324 (37.8)
◇ useful	7, 681 (3.4)
◇ social & useful	577 (0.3)
◇ other (good)	18 (0.0)
<hr/>	
# of bad candidates (%)	133, 650 (58.6)
◇ inappropriate	127, 648 (55.9)
◇ misleading	5, 800 (2.5)
◇ inappropriate & misleading	164 (0.1)
◇ other (bad)	38 (0.0)

Figure: Credits: [SMC+21]

ACUTE-Eval for comparing dialogue models

Hi how are you today

Hi, I'm pretty good! Just listening to some aerosmith, they're my fave :) whatre you up to?

I am listening to some italian music

Italian music, nice! What do you do for work?

I'm a professional wrestler

Oh wow! I'm a construction worker, I've been doing it since I was 16.

What do you do for a living

I'm a construction worker haha, just told you! Whats your favorite type of food?

My favorite is garlic bread

Yum! I love mexican food, so good :) do you have any hobbies?

I like to workout and go to the gym

We're a bit different- I love watching nascar and ufc. They're so fun!

Hello there, how are you?

I am doing great. How are you?

I am great, I did something crazy for me and colored my hair blue!

I have a daughter and a son who also love blue colored balls. You should meet them

Well that neat, I got a new car my mother gave so maybe I could see them!

It is a beautiful city. And, I try to be... Just cannot afford a bigger house atm.

I am sorry to hear that, I feel bad going out of town for spring break now.

Ok. I going to school in the spring for casino manager

Well I turn 29 next week, I wonder if that is a good age to apply as one.

My grandmother just died from lung cancer, sucks

Who would you prefer to talk to for a long conversation?

- I would prefer to talk to **Speaker 1**
- I would prefer to talk to **Speaker 2**

Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

Data Quality using ACUTE-Eval

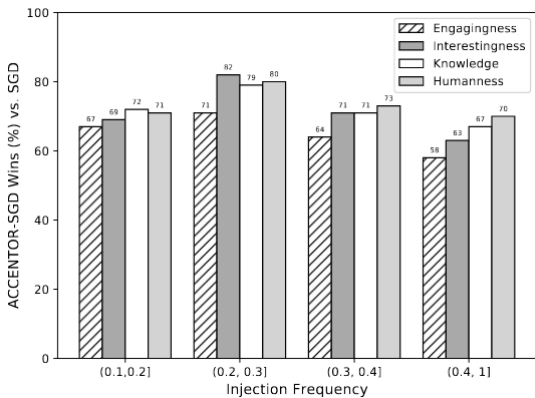


Figure: Image Credits: [SMC⁺21]

The Model Problem

ACCENTOR propose 3 models for fusing chit-chat with task-oriented dialogues

- 1 SimpleTOD(+)
- 2 Arranger
- 3 Rewriter

SimpleTOD and SimpleTOD+

- 1 SimpleTOD is as discussed before.
- 2 SimpleTOD+ additionally introduces *chit-chat* actions (add-before, add-after, do-not-add) to SimpleTOD

Arranger

- A RoBERTa based classifier is trained to decide if and where to add the chit-chat. Model is trained using good and bad candidates dataset.

Rewriter

- A GPT-2 based causal model which generates action and response based on dialogue history, generated belief states and chit-chat output

Both arranger and rewriter models use responses from off-the-shelf chit-chat and TOD models.

SimpleTOD is used as TOD model.

A BERT based chit-chat model fine-tuned on Accentor-SGD is used as chit-chat.

Results

		Win %			
		SimpleTOD	SimpleTOD+	Arranger	Rewriter
Loss %	SimpleTOD	-	63 **	76 **	91 **
	SimpleTOD+	37 **	-	46	50
	Arranger	24 **	54	-	39 *
	Rewriter	9 **	50	61 *	-

(a) Engagingness.

		Win %			
		SimpleTOD	SimpleTOD+	Arranger	Rewriter
Loss %	SimpleTOD	-	63 **	73 **	70 **
	SimpleTOD+	37 **	-	47	50
	Arranger	27 **	53	-	40 *
	Rewriter	30 **	50	60 *	-

(b) Interestingness.

		Win %			
		SimpleTOD	SimpleTOD+	Arranger	Rewriter
Loss %	SimpleTOD	-	64 **	77 **	81 **
	SimpleTOD+	36 **	-	47	55
	Arranger	23 **	53	-	45
	Rewriter	19 **	45	55	-

(c) Knowledge.

		Win %			
		SimpleTOD	SimpleTOD+	Arranger	Rewriter
Loss %	SimpleTOD	-	68 **	71 **	82 **
	SimpleTOD+	32 **	-	51	48
	Arranger	29 **	49	-	40 *
	Rewriter	18 **	52	60 *	-

(d) Humanness.

Figure: Credits: [SMC⁺21]

Comments

Pros

- Targets one of the essential areas for practical TOD systems
- Proposes a dataset based solution for TOD simple intuitive baselines
- We can extend this to multiple use-cases like domain mixing/transitions, multilingual dialogues and general knowledge grounded responses.

Cons

- Filtering results depend upon pilot data. This creates biases in the candidates.
- Does not support injection of chit-chat withing the response itself.

Relevance

- Personalized TODS
- Emotion infused models for sensitive domains like medical diagnosis and counselling
- Especially useful in conversational recommendation

Reviews: Advantages

- Models are simple and intuitive.
- Data generation and augmentation method is simple.
Filtering method is effective as performance is decent (40% on SGD and 30% on WoZ) with limited pilot data.
- Human annotation is considerate as it tries to remove strong opinions and misleading utterances.
- Code separation makes approach plug-n-play
- Good evaluation and results.

Reviews: Disadvantages I

- The ACUTE-Eval metrics are not exclusive and highly correlated.
- Although difficult, there should be a method of generating good chit-chat besides language models. The human annotators could be suggestive instead of just discriminative.
- Human annotators' involvement and human evaluation are extensive, which is usually not ideal.
- The user's preferences are not considered.
- The paper analyses chit-chat responses as being either prepended or appended in the SimpleTOD+ and Arranger models proposed.
- Source of Bias or inaccurate results: Authors annotate 1.7k candidates as good/bad.
- The focus should be on zero-shot/ few-shot methods of adding chit-chat, rather than creating new datasets.

Reviews: Disadvantages II

- For the initial filtering step, a better approach than the rule-based filtering would be to use crowd workers to produce annotations for this step as well.
- While discussing inappropriate behaviors of a candidate dialogue, the paper says that the “bot is not a person and should not pretend to have real life experience”. I disagree with this point, because the humanness of chatbots is an important factor.
- Does not handle 2 way chit chat, which may also be helpful in capturing the user’s sentiment towards the products/system.





Reviews: Extensions I

- Injection frequency should be a parameter that the user sets according to his/her own requirements.
- In Arranger there are 3 possibilities. However, if the user is chit-chatting, the agent could also just chit-chat. Therefore, there should be a 4th possibility of just chit-chatting. Rewriter could take care of this automatically though.
- Can further extend to other dialogue styles using style transfer - such as politeness, formal or informal etc which can be given as a parameter.
- Can focus on improving the candidate filtering mechanism - finding optimal settings (not done in the paper) as this can significantly improve the quality of the dataset.
- Extend dataset creation method so that chit-chat can be incorporated in between task-oriented-response.



Reviews: Extensions II

- Chit-chat may be subjective and this work can be extended to adjust the subsequent chit-chat according to how the user is responding to models chit-chat in the past. And Subjectivity of humans should also be taken into account in the evaluation process i.e. if possible the model responses should be subjective to the user.
- Generating rule based outputs using knowledge bases or databases like wikipedia, private knowledge bases etc can increase the number of responses of chit chats.
- Having a way to **measure and control the amount of engagement, interesting, knowledge and human-like** behavior of dialogue system may be useful.

References I

-  Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang, *A survey on dialogue systems: Recent advances and new frontiers*, ArXiv **abs/1711.01731** (2017).
-  Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher, *A simple language model for task-oriented dialogue*, ArXiv **abs/2005.00796** (2020).
-  Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishausser, Hsien-Chin Lin, Marco Moresi, and Milica Gavsic, *Trippy: A triple copy strategy for value independent neural dialog state tracking*, ArXiv **abs/2005.02877** (2020).
-  Margaret Li, Jason Weston, and Stephen Roller, *Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons*, ArXiv **abs/1909.03087** (2019).

References II

-  Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becca Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie, *Adding chit-chat to enhance task-oriented dialogues*, ArXiv **abs/2010.12757** (2021).
-  Chien-Sheng Wu, *Learning to memorize in neural task-oriented dialogue systems*, ArXiv **abs/1905.07687** (2019).

Thank You