

A first study on a dialogue system for an academic writing aid system

data adaption from restaurant reservation to writing aid and measurement of accuracy
of dialogue act prediction and slot filling

Ni Zihui

Graduate School of Information, Production and Systems,
Waseda University



nizihui@akane.waseda.jp

July 29, 2021

Outline

1 Introduction

2 Method

3 Experiment

4 Conclusion

5 Future work

Outline

1 Introduction

2 Method

3 Experiment

4 Conclusion

5 Future work

The goal: Realize a dialogue system for the writing aid task

- To build a dialogue dataset for the writing aid task.
Try to automatically modify the sentences and the annotations in the existing dialogue dataset.
- To finetune a dialogue model for the writing aid task.
In the existing dialogue framework, finetune the task with the modified data.

An example of the dialogue to the writing aid task

- **The system** (output): Hello, welcome to the writing aid system. You can do many kinds of modifications to the paragraph here. Do you want a grammar check first?
- **The user** (input): Yes, please.
- **The system** (output): Okay, do you want an error report or just modify the paragraph?
- **The user** (input): An error report please.

Introduction: The existing dialogue system

- The current task-based dialogue systems have been applied to many fields, such as *hotel reservations*, *ticket booking*, and *take-way ordering*.

Name of the system	Dialogue scalability	Natural language management	Natural language Understanding	Natural language generation
<i>OpenDial</i> [Lison and Kennington, 2016]	direct coding	plan	machine learning	template
<i>RASA</i> [Bocklisch et al., 2017]	direct coding	machine learning	machine learning	-
<i>ParIAI</i> [Miller et al., 2017]	direct coding	machine learning	machine learning	template
<i>DeepPavlov</i> [Burtsev et al., 2018]	direct coding	machine learning	machine learning	template
<i>OwlSpeak</i> [Heinroth et al., 2010]	universal plug and play	information state	rule	template
<i>PyDial</i> [Ultes et al., 2017]	direct coding	reinforcement learning	rule and machine learning	rule and machine learning
<i>DialogStudio</i> [Jung et al., 2008]	direct coding	example	machine learning	example
<i>RavenClaw</i> [Bohus and Rudnicky, 2009]	direct coding	plan	rule	template

Table: Some existing dialogue systems

Introduction: The existing dataset

- **The current dialogue datasets** have been applied to many fields, such as *restaurant reservation*, *ticket booking*, and *flight*.

Name of the dataset	Size	Source	Quality	Fields
<i>DailyDialog</i> [Li et al., 2017]	13,118	textbooks	auto-extracted	various topics
<i>Wizard-of-Wikipedia</i> [Dinan et al., 2019]	22,311	crowdsourcing	human-written	various topics
<i>Document-grounded</i> [Feng et al., 2020]	4,470	crowdsourcing	human-written	various topics
<i>Persona-Chat</i> [Zhang et al., 2018]	10,981	crowdsourcing	human-written	various topics
<i>Self-dialogue</i> [Fainberg et al., 2018]	24,165	crowdsourcing	human-written	various topics
<i>Cornell Movie Corpus</i> [Danescu-Niculescu-Mizil and Lee, 2011]	304,713	movie scripts	auto-extracted	movie
<i>Self-feeding chatbot</i> [Heinroth et al., 2010]	145,873	human-bot dialogues	human-written (half)	various topics
<i>Twitter corpus</i> [Cieliebak et al., 2017]	9,783	twitter posts/replies	auto-extracted	various topics
<i>Opensubtitles</i> [Lison and Tiedemann, 2016]	5,000,000	movie subtitles	auto-extracted	movie
<i>AirDialogue</i> [Wei et al., 2018]	301,427	human annotators	human-written	travel, flight
<i>MultiWOZ</i> [Budzianowski et al., 2018]	8,438	human behaviors	human-written	hotel, restaurant
<i>DSTC 2</i> [Henderson et al., 2014]	2235	human behaviors	Amazon Mechanical Turk	restaurant reservation

Table: Some existing datasets of the dialogue system

The goal: Realize a dialogue system for the writing aid task

- To build a dialogue dataset for the writing aid task.
Try to automatically modify the sentences and the annotations in the existing dialogue dataset.
- To finetune a dialogue model for the writing aid task.
In the existing dialogue framework, finetune the task with the modified data.

The DSTC 2 dataset

- The dataset we built is based on the DSTC 2 dataset [Henderson et al., 2014].
- DSTC 2 is a human-machine and multi-round dialogue dataset in the field of restaurant reservation, in which there are 1,612 training data, 506 validation data, and 1,117 test data.
- This is the annotations of the DSTC 2 dataset.

annotation	explanation
speaker	the system or the user
text	the dialogue text
goal	information about the goal/task for the system
slots	request parameter (slot list)
acts	the purpose of the dialogue sentence, such as inform, inquiry, etc
dialogue_acts	this is determined by what slots and acts is , directly determines how to generate answers and call the APIs

Table: The annotations of the DSTC 2 dataset

The DSTC 2 dataset

This is an example of the DSTC 2 dataset:

```
{
  "speaker": 2,
  "text": "Hello, welcome to the Cambridge
restaurant system. You can ask for
restaurants by area, price range or food
type. How may I help you?",
  "dialog_acts": [
    {
      "act": "welcomemsg",
      "slots": [
      ]
    }
  ]
}

{
  "speaker": 1,
  "text": "cheap restaurant",
  "goals": {
    "pricerange": "cheap"
  },
  "dialog_acts": [
    {
      "slots": [
        {
          "pricerange",
          "cheap"
        }
      ],
      "act": "inform"
    }
  ]
}
```

Figure: An example of the DSTC 2 dataset (Graph copied from [Henderson et al., 2014])

The introduction of the slot table

- **Slot** is parameters of the query of user.
- For example, in the keyword extraction task:
 - Slot: User needs to appoint **how many keywords** to be extracted.
 - `number_of_keywords = 5`

slot	informable	values
food	yes	91 possible values
name	yes	113 possible values
pricerange	yes	3 possible values
addr	no	-
phone	no	-
postcode	no	-
signature	no	-

Table: The slot table designed in DSTC 2 [Henderson et al., 2014]

The introduction of the slot table

```
"food": {  
  "caribbean": [  
    "carraibbean",  
    "carribean",  
    "caribbean"  
  ],  
  "kosher": [  
    "kosher"  
  ],  
  "tuscan": [  
    "tuscan"  
  ],  
  "french": [  
    "french"  
  ]  
}  
  
"this": {  
  "dontcare": [  
    "dont care",  
    "doesnt  
matter",  
    "any fine",  
    "any noise",  
    "any of town",  
    "noise  
anything",  
    "any type",  
    "anything",  
    "any thing",  
    "what ever",  
    "does not  
matter"  
  ]  
}
```

Figure: An example of the slot and the corresponding slot value in the DSTC 2 dataset (Graph copied from [Henderson et al., 2014])

The goal we need to do in the dataset

We need to modify the dataset from **dialogue sentences** and **dialogue annotations**.

The goal: Realize a dialogue system for the writing aid task

- To build a dialogue dataset for the writing aid task.
Try to automatically modify the sentences and the annotations in the existing dialogue dataset.
- To finetune a dialogue model for the writing aid task.
In the existing dialogue framework, finetune the task with the modified data.

Introduction: The structure of the existing dialogue system

- **User input:** User request.
- **System output:** System answer.

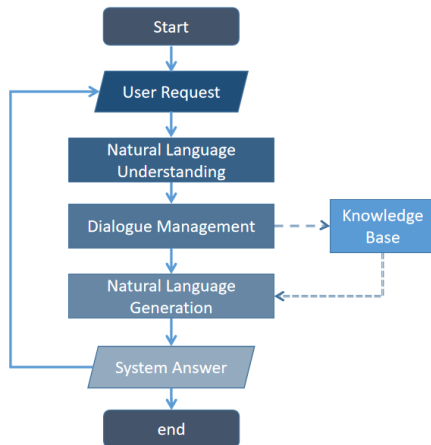


Figure: Task-oriented Pipeline Dialogue System

Introduction: The structure of the our dialogue system

- **User input:** User request and the paragraph.
- **System output:** System answer and the paragraph.

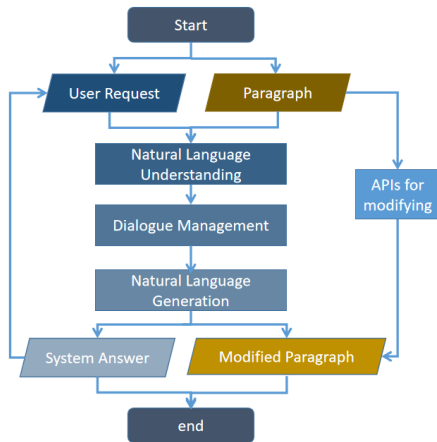


Figure: The flow chart of the pipeline dialogue system for writing aid task

Introduction: The dialogue system

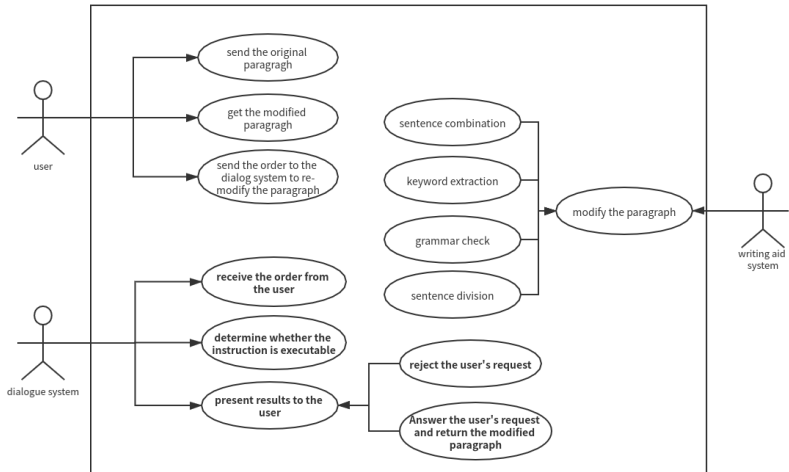


Figure: The user case for an academic writing aid system

Outline

1 Introduction

2 Method

3 Experiment

4 Conclusion

5 Future work

Method: Dataset

■ **Modify the annotations in the dataset.**

- Design the new slot and slot value in the slot table.
- Make the new slot and slot value appear in the new dialogue sentences and annotations.

■ **Modify the sentences in the dataset.**

- Prepare writing aid-related sentence dataset.
- Replace the all the words about restaurant reservation in DSTC 2 dataset.
- Use grammar check to correct the grammatical errors in the dialogue text.

Method: Dataset

(1) Replace the slot and the slot value from the original DSTC 2 dataset to the newly designed slot table.

slot	informable	numbers of values	values
type of sentence division terms	yes	3	and, but, so
type of grammar error correction	yes	2	error report, modified paragraph
number of keyword extraction	yes	5	1,2,3,4,5
type of text summarization	yes	2	opening sentence, abstract
number of sentences	no	-	-
average number of words	no	-	-
number of total words	no	-	-

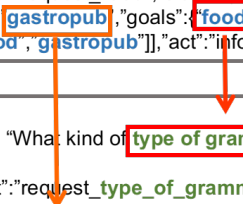
Table: The newly designed slot table

Method: Dataset

(2) Make the new slot and slot value appear in the new dialogue sentences.

DSTC 2 dataset:

```
{
  "speaker": 2,
  "text": "What kind of food would you like?",
  "dialog_acts": [
    {
      "act": "request_food",
      "slots": []
    }
  ]
},
{
  "speaker": 1,
  "text": "gastropub",
  "goals": [
    "food",
    "gastropub"
  ],
  "dialog_acts": [
    {
      "slots": [
        "food",
        "gastropub"
      ],
      "act": "inform"
    }
  ]
}
```



Modified dataset:

```
{
  "speaker": 2,
  "text": "What kind of type of grammar error correction would you like?",
  "dialog_acts": [
    {
      "act": "request_type_of_grammar_error_correction",
      "slots": []
    }
  ]
},
{
  "speaker": 1,
  "text": "error report",
  "goals": [
    "type of grammar error correction",
    "error report"
  ],
  "dialog_acts": [
    {
      "slots": [
        "type of grammar error correction",
        "error report"
      ],
      "act": "inform"
    }
  ]
}
```

Figure: The example of how to change the slot and slot value in the DSTC 2 dataset

Method: Dataset

(3) Prepare writing aid-related sentence dataset. We get **84** conversations about writing aid manually and added **2,494** sentences in **ACL-ARC dataset** [Bird et al., 2008]¹ as the dataset about writing aid.

Please demonstrate it with the result with the conjunction of the word "and".
I want to see the number of sentences.
I want to see the number of total words.
One possible weakness of discourse-based summarization techniques is that they rely greatly on the accuracy of the discourse parser they use .

Table: The example of the writing aid-related sentence dataset

¹use the ACL-ARC sentences extracted in the Mango system
:/files/Data/AcademicWritingAid/backup/cleaned_data

Method: Dataset

(4) We use pre-trained embedding BERT model [Devlin et al., 2019] to replace the extracted words of DSTC 2 with the words in the writing aid dataset. For each word W_i [Li et al., 2018] in DSTC 2, we calculate the probability $P_1(w_i)$ in DSTC 2 and $P_2(w_i)$ in writing aid dataset. And then we calculate

$$W_i = |\log(|P_1(w_i) - P_2(w_i)|)|$$

If W_i of word i is near 0, it should be substituted. The closer W_i is to 0, the greater the frequency difference between the two texts, and the more this word should be replaced.

Method: Dataset

(4) Use BERT model [Devlin et al., 2019] to replace extracted words of DSTC 2 to the words in the writing aid dataset. We set the threshold at 9.72. This is the part of the words that we sift out and need to be replaced.

word	result
euorpean	9.714
korean	9.714
burger	9.714
serve	9.714
mexican	9.714
Newnham	9.714

Table: The words are copied from the output of the dataset.

Method: Dataset

(5) Use **grammar check** to correct the grammatical errors in the dialogue text.

Method: The structure of the dialogue system

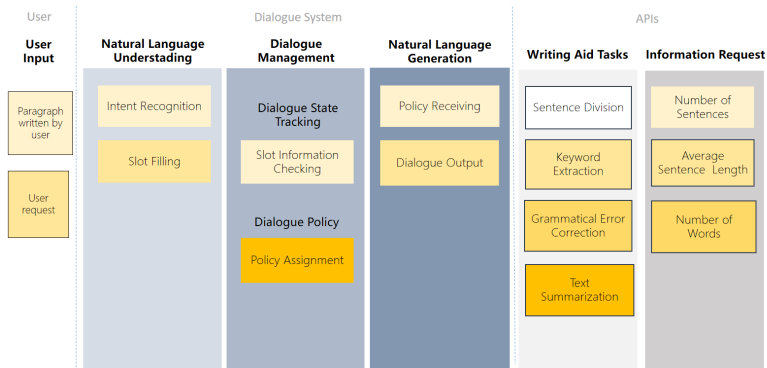


Figure: The structure of dialogue system

Method: The DeepPavlov framework

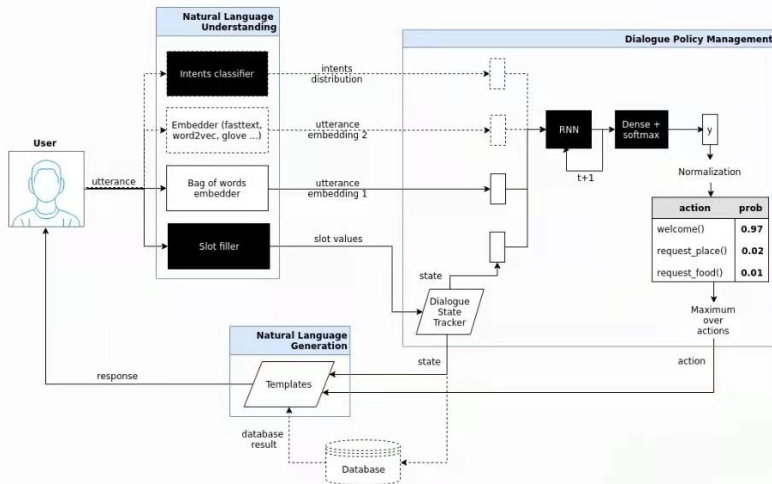


Figure: The structure of DeepPavlov framework (The graph is copied from [Burtsev et al., 2018])

Outline

1 Introduction

2 Method

3 Experiment

4 Conclusion

5 Future work

Experiment: Dataset

We generate **2,235** sets of dialogue data, containing **39,420** dialogue sentences with their annotations.

<i>'speaker' : 2, 'text' : ' Hello, welcome to the writing aid system. You can do many kinds of modification of the paragraph here. Do you want a grammar check first?'</i> , <i>'dialog_acts' : 'act' : ' welcomemsg', ' slots' : []</i>
<i>'speaker' : 1, 'text' : ' Yes, please give me the grammar check.', 'goals' : 'grammar check' : ' dontcare', 'dialog_acts' : ' slots' : ['type of text summarization', ' abstract'], ' act' : ' inform'</i>
<i>'speaker' : 2, 'text' : ' What kind of type of grammar error correction would you like?', 'dialog_acts' : ' act' : ' request_type_of_grammar_error_correction', ' slots' : []</i>
<i>'speaker' : 1, 'text' : ' any', 'goals' : 'type of grammar error correction' : ' dontcare', ' type of text summarization' : ' dontcare', 'dialog_acts' : ' slots' : [['this', ' dontcare']], ' act' : ' inform'</i>

Table: The example of our building writing aid dialogue dataset

Experiment: Baseline

- We use the gobot model under the DeepPavlov framework to calculate the accuracy of dialogue action in the dialogue state tracking (DST) part, the precision, recall, and F1 score in the slot filling (SF) task. We use the results of DSTC 2 [Henderson et al., 2014], MultiWOZ [Budzianowski et al., 2018] and AirDialogue [Wei et al., 2018] in these tasks as the baselines.
- We want to prove that the prediction effect of our new dataset under the same DeepPavlov framework can be basically the same as that of the mainstream dataset.

Experiment: Result

- Our dataset is modified from the DSTC 2 dataset, with **1,612** training set, **506** validation set, and **1,117** test set.

Dataset	accuracy (DST)	precision (SF)	recall (SF)	F1 (SF)
DSTC 2	47%	99%	97%	98%
MultiWOZ	47%	96%	96%	96%
AirDialogue	22%	59%	56%	55%
Ours	51%	98%	93%	96%

Table: The quantitative result of predicting bot answers on datasets.

The result of the dialogue system

Speaker 1: Hello, welcome to the writing aid system. You can do many kinds of modifications to the paragraph here. Do you want a grammar check first?
Speaker2: Yes, and please send me the error report.
Speaker1: Okay, I'll do it for you.

Table: The example of the chat in the writing aid task

Outline

1 Introduction

2 Method

3 Experiment

4 Conclusion

5 Future work

Conclusion

- We built a writing aid dialogue dataset based on DSTC 2.
- We finetuned our writing aid dialogue task on the DeepPavlov framework.

Outline

1 Introduction

2 Method

3 Experiment

4 Conclusion

5 Future work

Future work

- Test our dataset with different dialogue system frameworks.
- Analyze the reasons why our dataset have different results in different dialogue system frameworks.

Thank you for your listening.

Q&A

References I



Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008).

The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.

In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).



Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017).

Rasa: Open source language understanding and dialogue management.

ArXiv, abs/1712.05181.



Bohus, D. and Rudnicky, A. I. (2009).

The ravenclaw dialog management framework: Architecture and systems.

Comput. Speech Lang., 23:332–361.

References II



Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (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, Brussels, Belgium. Association for Computational Linguistics.



Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhreva, M., and Zaynutdinov, M. (2018).

DeepPavlov: Open-source library for dialogue systems.

In Proceedings of ACL 2018, System Demonstrations, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.

References III



Cieliebak, M., Deriu, J. M., Egger, D., and Uzdilli, F. (2017).
A Twitter corpus and benchmark resources for German
sentiment analysis.

*In Proceedings of the Fifth International Workshop on Natural
Language Processing for Social Media*, pages 45–51, Valencia,
Spain. Association for Computational Linguistics.



Danescu-Niculescu-Mizil, C. and Lee, L. (2011).
Chameleons in imagined conversations: A new approach to
understanding coordination of linguistic style in dialogs.

*In Proceedings of the Workshop on Cognitive Modeling and
Computational Linguistics, ACL 2011.*

References IV



Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.



Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019).

Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, abs/1811.01241.

References VI



Henderson, M., Thomson, B., and Williams, J. D. (2014).

The second dialog state tracking challenge.

In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.



Jung, S., Lee, C., and Lee, G. (2008).

Dialogstudio: A workbench for data-driven spoken dialog system development and management.

Speech Communication, 50:697–715.

References VII



Li, J., Jia, R., He, H., and Liang, P. (2018).

Delete, retrieve, generate: a simple approach to sentiment and style transfer.

In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.



Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017).

DailyDialog: A manually labelled multi-turn dialogue dataset.

In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

References VIII



Lison, P. and Kennington, C. (2016).

OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules.

In *Proceedings of ACL-2016 System Demonstrations*, pages 67–72, Berlin, Germany. Association for Computational Linguistics.



Lison, P. and Tiedemann, J. (2016).

OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles.

In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

References IX



Miller, A., Feng, W., Batra, D., Bordes, A., Fisch, A., Lu, J., Parikh, D., and Weston, J. (2017).

ParlAI: A dialog research software platform.

In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.



Ultes, S., Rojas-Barahona, L. M., Su, P.-H., Vandyke, D., Kim, D., Casanueva, I., Budzianowski, P., Mrkšić, N., Wen, T.-H., Gašić, M., and Young, S. (2017).

PyDial: A multi-domain statistical dialogue system toolkit.

In Proceedings of ACL 2017, System Demonstrations, pages 73–78, Vancouver, Canada. Association for Computational Linguistics.

References X



Wei, W., Le, Q., Dai, A., and Li, J. (2018).

AirDialogue: An environment for goal-oriented dialogue research.

In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3844–3854, Brussels, Belgium. Association for Computational Linguistics.



Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018).

Personalizing dialogue agents: I have a dog, do you have pets too?

In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.