Context Aware Module Selection in Modular Dialog Systems

Jan Nehring
German Research Center for
Artificial Intelligence (DFKI)
Alt-Moabit 91c
10559 Berlin, Germany
jan.nehring@dfki.de

René Marcel Berk, Stefan Hillmann
TU Berlin

Straße des 17. Juni 135
10623 Berlin, Germany
firstname.lastname@tu-berlin.de

Abstract

In modular dialog systems, a dialog system consists of multiple conversational agents. The task "module selection" selects the appropriate sub-dialog system for an incoming user utterance. Current models for module selection use features derived from the current user turn only, such as the utterances text or confidence values of the natural language understanding systems of the individual conversational agents, or they perform text classification on the user utterance. However, dialogs often span multiple turns, and turns are embedded into a context. Therefore, looking at the current user turn only is a source of error in certain situations. This work proposes four models for module selection that include the dialog history and the current user turn into module selection. We show that these models surpass the current state of the art in module selection.

1 Introduction

Dialog systems (DS) often consist of multiple sub-dialog systems or modules. There are multiple reasons for such a combination: The designer of a DS might want to combine several existing DS without a reimplementation. Sometimes a DS spans multiple departments and cannot be merged into a single, unified system. A hybrid system is a possible solution when a DS consists of multiple incompatible subsystems, e.g., a task-oriented DS and a question-answering system. Although this architecture is frequently used in practical applications, it is a gap in scientific research.

The modular dialog system (MDS) (Nehring and Ahmed, 2021) describes a framework to combine several dialog systems. In an MDS, a central component called "module selection" (MS) selects the appropriate sub-DS that generates the answer for an incoming user utterance (Nehring et al., 2023). MS is a classification task to choose one sub-DS from a list of sub-DS for a given user utterance.

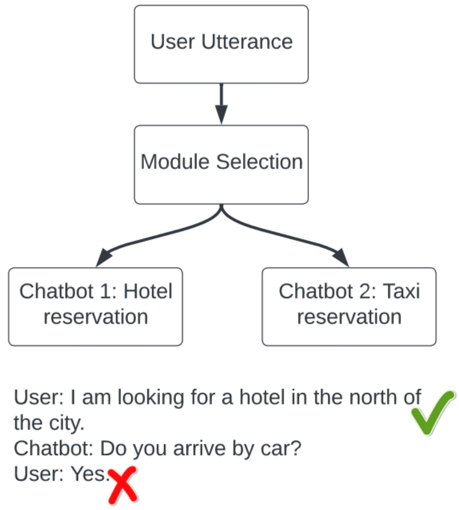

Figure 1: Example dialog between a user and a modular dialog system that consists of two chatbots. Based on the current utterance, the module selection can easily select the proper agent for the first user utterance. However, the module selection requires dialog context to classify the second user utterance correctly.

Current solutions for MS, such as Görzig et al. (2023) or Nehring et al. (2023), focus on the current user utterance only. They use the text of the user utterance, confidence values of the models NLUs, or additional features such as detected named entities for the models. However, both works showed that the text of the user utterance is the essential feature for high performance in MS (Görzig et al., 2023; Nehring et al., 2023). In some cases, more than the current user utterance and other derived features are needed to find the appropriate sub-DS.

Figure 1 shows an example MDS that consists of a hotel reservation bot and a taxi reservation bot. The MS can easily categorize the first user utterance of the example dialog "I am looking for a hotel in the north of the city". However, the second user utterance "yes" alone does not transport

Proceedings of Recent Advances in Natural Language Processing, pages 785–791
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_085

enough information for the MS. Therefore, we propose models to include the dialog history and the current user utterance into MS. We show that these models surpass state of the art in MS.

2 Background

In this work, we use task-oriented dialog systems which "use conversation with users to help complete tasks" (Jurafsky and Martin, 2009). Jurafsky and Martin (2009) define a turn as a "single contribution from one speaker of the dialog". The length of a turn is not fixed but can consist of a single utterance or up to multiple sentences. Let U_i be the ith turn of the user and S_i the ith turn of the system. A dialog is a sequence of alternating user and system turns U_1S_1 ... U_nS_n.

Jurafsky and Martin (2009) describe a typical architecture for task-oriented dialog systems: Each incoming user turn is first processed by *Natural Language Understanding* (NLU), which converts the unstructured textual information of the user turn into structured information. Most notable is intent detection, which classifies the user turn to a list of predefined intents. Another standard function of the NLU is slot filling, which extracts slots from the user turn. Slots are entities such as dates, names, or places. So, for example, for the user turn "I want to book a table for Friday, 8 pm" the NLU can detect the intent "book_table" and the slot "time = Friday 8 pm".

Dialog state tracking processes the results of the NLU and keeps track of the slot values across the dialog. So in the restaurant booking domain, we might define slots time and number of people. During the dialog, dialog state tracking fills these slots with values. A *dialog manager* keeps track of the various states of the dialog. Dialog managers can be hand-crafted or machine-learned. Finally, the *answer generation* generates the system turn, which is shown to the user.

MDS and MS are similar to multidomain dialog systems (MDDS) (see, e.g., (Ultes et al., 2017)), in which a dialog system encompasses different domains. The Multiwoz dataset was originally a dataset for MDDS. However, the essential difference between MDS and MDDS is the motivation: In MDDS, the goal is a dialog system with maximal performance, which can be implemented in a single, monolithic system. On the other hand, in MDS, we want to distribute the system across several DS, which often results in a decreased performance (Nehring et al., 2023).

3 Approach

3.1 Dataset Generation

We created a dataset for our application based on MultiWOZ dataset version 2.2 (Zang et al., 2020). MultiWOZ was first introduced by Budzianowski et al. (2018). It is "a large-scale multi-turn conversational corpus with dialogs spanning across several domains and topics. Each dialog is annotated with a sequence of dialog states and corresponding system dialog acts" (Budzianowski et al., 2018). It covers eight domains about the city of Cambridge in England: Attraction, general, hospital, hotel, police, restaurant, taxi, and train. Several improved versions of MultiWOZ add or correct the annotations. We chose MultiWOZ 2.2 because it improved intent annotation quality.

We deleted 3.452 dialogs from the dataset: 1) 1.639 dialogs cover multiple domains in a user turn. In our system a user utterance can be assigned to one single intent only, which is a common design choice in dialog systems, such as Rasa[1], Google Dialogflow[2] or IBM Watson Assistant[3]. 2) Some dialogs that missed the dialog act annotation in at least one turn. 3) We deleted dialogs with the domains hospital and police, because these domains were only present in the training partition of MultiWOZ and not in the valid or test partition.

Further, we preprocessed the dialogs: We lowercased all utterances, removed duplicate whitespaces, and normalized telephone numbers and postcodes. Also, we expanded contractions, such as "it's" to "it is" or "haven't" to "have not".

We kept the train, test, and valid partitioning from the original dataset, resulting in a dataset with 37.264 user turns in the training partition, 4.903 in the validation, and 4.991 user turns in the test partition.

Table 3 in the appendix shows an example dialog from the dataset that spans three domains. For better readability, we omitted the lowercasing of the text. Typically for this dataset, the dialog spans multiple domains and switches back and forth between them. The example shows that spelling and punctuation are not uniform: The user utterance in turn four starts with a lowercase "i". Names

[1] https://rasa.com
[2] https://cloud.google.com/dialogflow
[3] https://www.ibm.com/products/watson-assistant

such as Cambridge or London are not capitalized correctly.

3.2 Dataset characteristics

Figure 2 show the number of dialogs and user turns per domain and dataset partition. The dataset is imbalanced, with the taxi domain being the minority class. The domain general encompasses greetings and goodbye. Therefore it occurs in more dialogs than in the other classes. At the same time, conversations about the general domain are relatively short. Hence, the number of turns in the general domain is similar to that in other domains.

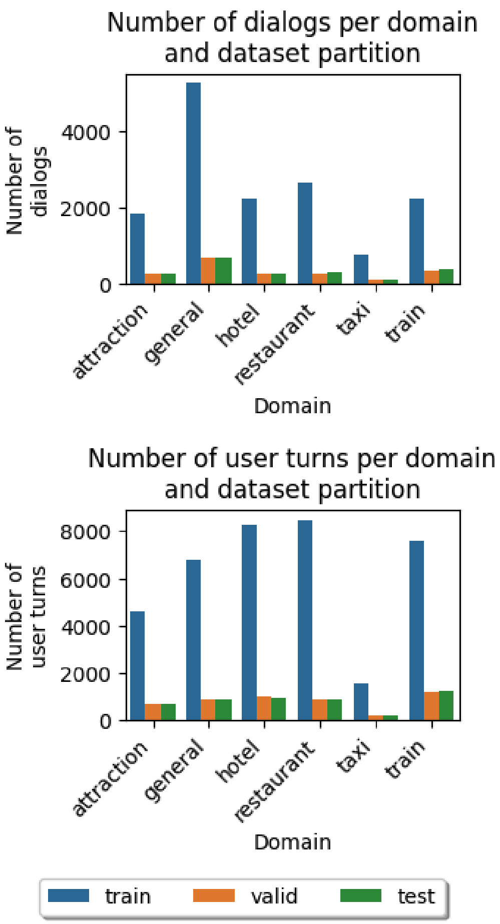

Figure 2: Number of dialogs and user turns per domain and data partition.

Figure 3 shows a boxplot of the length of the dialogs. The mean dialog length is 6.47, with a standard deviation of 2.32. The mean value for the number of domains per dialog is 2.61, with a standard deviation of 0.70. Only a few dialogs span a single domain, while most dialogs cover two or three domains.

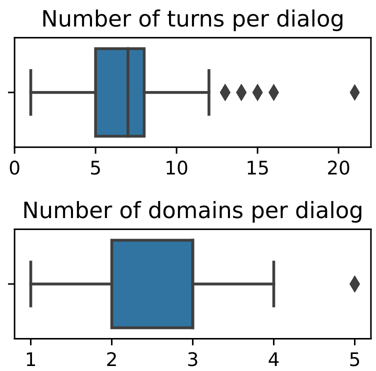

Figure 3: Lengths of the dialogs.

3.3 Experimental settings

We assigned the six domains to the dataset described in section 3.1 six agents in an MDS. However, we did not create individual dialog systems. We trained the MS only because this is enough for our experiments. Figure 4 shows the system.

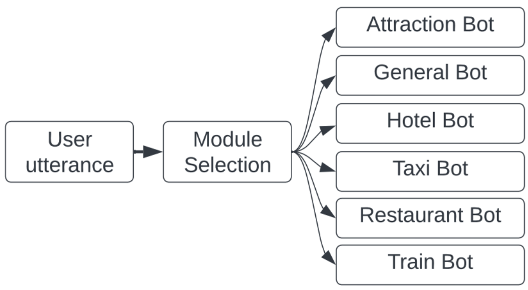

Figure 4: Architecture of the modular dialog system.

We trained the models described in section 3.4 on this dataset. We used a learning rate of 5×10^{-5} and a training batch size of 16 and three training epochs for all models. As an evaluation metric we used micro F1 scores.

3.4 Models

In our experiment, we use four different models for MS. The **baseline model** is a standard BERT model with a sequence classification head (Devlin et al., 2019), which was used for MS by Nehring et al. (2023). The baseline model classifies the current user utterance only.

We introduce three models that are aware of the history. They share the same architecture. Again we use the BERT for sequence classification architecture as in the baseline model. However, this time, we concatenate the texts of several previous user and system utterances. The **full history (FH)**

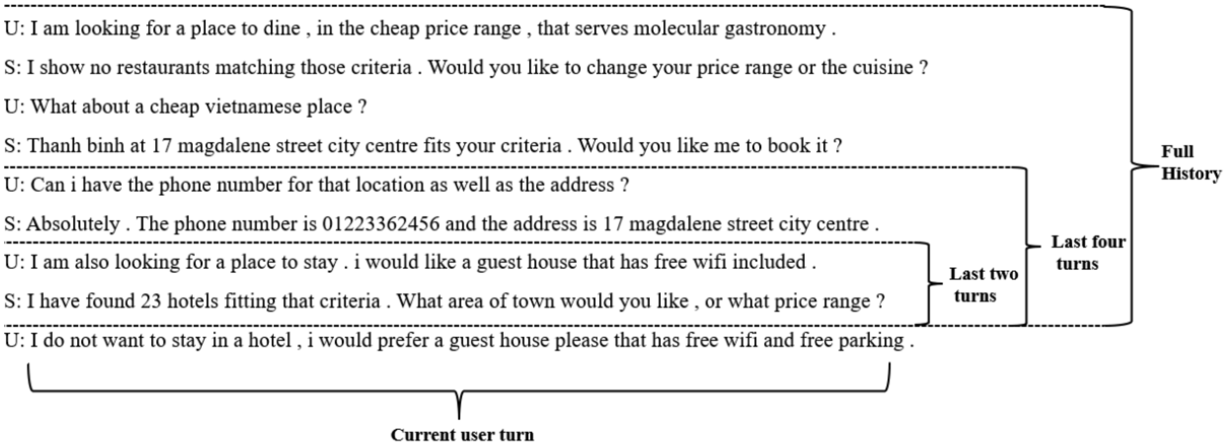

Figure 5: Depiction of the model context.

Model	F1-Score
Baseline	92.6%
FH	99.0%
L2T	98.7%
L4T	99.1%

Table 1: Performance of the models as micro F1-scores

Domain	Baseline	L2T	L4T	FH
Attraction	89.0%	99.1%	99.1%	99.2%
General	98.4%	98.5%	98.7%	98.3%
Hotel	90.1%	98.9%	99.2%	99.2%
Restaurant	88.0%	98.2%	98.9%	98.7%
Taxi	90.6%	96.9%	98.7%	98.0%
Train	96.1%	99.6%	99.6%	99.5%

Table 2: F1-scores of MS for each domain and model

model uses the entire dialog history. The **last two turns (L2T)** model uses the current user utterance, the last system utterance, and the user utterance before that. The **last four turns (L4T)** model concatenates the current user utterance and the last two system and user utterances. The input of BERT is limited to 512 tokens. So in case the input is longer than 512 tokens, we truncate the input by dropping the oldest input text so that the input length is 512 tokens. Figure 5 depicts the different contexts of the FH, L2T, and L4T models.

4 Results

Table 1 shows the results of the experiments. The three proposed models FH, L2T, and L4T produce high scores and surpass the baseline model. However, the FH, L2T, and L4T scores differ by 0.4%, which is very similar. This difference accounts for 20 wrongly classified samples out of the 4.991 test set samples.

Table 2 shows the F1-Scores per domain and model.

5 Discussion

All three proposed models surpass state of the art (see table 1). So we show that MS depends on the dialog history and that dialog history is an essential feature for MS.

At the same time, their results are very similar. We conclude that the most important contributions of the dialog history to the model's performance stem from the last turn (L2T model). Including longer parts of the history (models L4T and FH) improves the performance only marginally.

The F1-scores for the individual domains (table 2) are generally high. The general domain has the highest F1-scores. We hypothesize that the general domain encompasses greetings and goodbyes, which are relatively easy to detect, especially when the training data is large, with approximately 7k training samples. In section 3.2 we stated that the taxi domain is the minority class with much fewer training examples than the rest. Still, the f1-scores of the taxi domain are in the same range as the other domains. We argue that, although the taxi domain is the minority class, the amount of training samples is still rather high.

Generally, the amount of training data is huge compared to the small number of domains and the limited range of the domains. The amount of generated training data would be lower in a practical use case due to the cost of training data generation. Also, in a real-world scenario, the test data will be more diverse. So although we reached almost 100% F1-score in our experiments on this dataset, we do not believe that the task MS is solved.

6 Related works

Here we give an overview of MS. Other approaches used features derived from the current user utterance only; Nehring et al. (2023) and Nehring et al. (2021) used a text classification on the current user utterance, which serves as the baseline model in our paper. Görzig et al. (2023) compared various features for MS with each other, such as confidence values or slot values of the dialog systems NLU. However, these works do not utilize the dialog history for MS.

The scientific literature proposes several approaches to combine multiple dialog systems. Some authors (Planells et al., 2013; Banchs et al., 2013) use domain classification, which is similar to our framework, although their work stems from the MDDS tradition and not from the MDS tradition. Another strategy is to let every dialog system generate a response and rank them to find the most suitable response (Tanaka et al., 2019; Song et al., 2018). The very successful and feature-rich chatbot Xiaoice uses a framework based on Options over Markov Decision Processes to decide which of his modules can answer the user utterance (Zhou et al., 2020). The CLARA dialog system lets the user decide which of the sub-agents he wants to talk to (D'Haro et al., 2015).

7 Conclusion

We proposed new models for MS that include the dialog history as a feature. These models show a high performance in our experiments, surpassing the state of the art. Moreover, the performance is close to 100%, meaning that, at least on our dataset, the MS works almost perfectly. In the future, we plan to experiment more with our approach with less training data and on more diverse and challenging datasets.

8 Acknowledgments

This project received funding from the German Federal Ministry for Economic Affairs and Climate Action (BMWK) as part of the ToHyVe project, funding number 01MT22002C and from the German Federal Ministry of Education and Research (BMBF) as part of the USOS project, funding number 16DHBKI088.

References

Rafael E Banchs, Ridong Jiang, Seokhwan Kim, Arthur Niswar, and Kheng Hui Yeo. 2013. Aida: Artificial intelligent dialogue agent.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 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 (EMNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luis Fernando D'Haro, Seokhwan Kim, Kheng Hui Yeo, Ridong Jiang, Andreea I. Niculescu, Rafael E. Banchs, and Haizhou Li. 2015. Clara: A multifunctional virtual agent for conference support and touristic information.

Philine Görzig, Jan Nehring, Stefan Hillmann, and Sebastian Möller. 2023. A comparison of module selection strategies for modular dialog systems. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, ESSV'23.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Jan Nehring, Ahmed Ahmed, and Lena A. Jager. 2023. Module selection: A new task for dialog systems. In *Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology*, IWSDS'23.

Jan Nehring and Akhyar Ahmed. 2021. Normalisierungsmethoden für Intent Erkennung Modularer Dialogsysteme. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, pages 264–271. TUDpress, Dresden.

Jan Nehring, Nils Feldhus, Harleen Kaur, and Akhyar Ahmed. 2021. Combining open domain question answering with a task-oriented dialog system. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 38–45, Online. Association for Computational Linguistics.

Joaquin Planells, Lluís F Hurtado, Encarna Segarra, and Emilio Sanchis. 2013. A multi-domain dialog system to integrate heterogeneous spoken dialog systems. pages 1891–1895. ISCA.

Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. 2018. An ensemble of retrieval-based and generation-based human-computer conversation systems. pages 4382–4388. International Joint Conferences on Artificial Intelligence Organization.

Ryota Tanaka, Akihide Ozeki, Shugo Kato, and Akinobu Lee. 2019. An ensemble dialogue system for facts-based sentence generation.

Stefan Ultes, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 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.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46:53–93.

Turn	User utterance	Domain	System answer
1	I need to book a train to cambridge. I would like to arrive by 21:00.	train	There is a depature at london kings cross on friday which departs at 7:17.
2	I need to depart london liverpool street on tuesday to arrive by 21:00.	train	I have one that arrives at 19:07, would that be alright?
3	Sure! What is the total travel time for that one, and what's the train ID?	train	Train ID: TR1395 it an 88 min. ride.
4	i am also looking for a park or something in town.	attraction	There are 5 parks in town what area are you looking for?
5	Sure! Any area is fine. Can you provide me the phone number of the first park on the list?	attraction	Cambridge university botanic gardens phone number is 01223336265
6	Thanks. Can you book the train ticket for me?	train	Is it just you traveling or do you also have guests?
7	Just me thanks	train	You are booked on TR1395. The confirmation number is RRXHU8AY .
8	Okay, I'm all done. Thanks! Bye!	general	You're welcome have a great day.

Table 3: Example dialog from the dataset