Joint Goal Segmentation and Goal Success Prediction on Multi-Domain Conversations

Meiguo Wang **Benjamin Yao** **Bin Guo**
Xiaohu Liu **Yu Zhang** **Tuan-Hung Pham** **Chenlei Guo**
Amazon Alexa AI

Abstract

To evaluate the performance of a multi-domain goal-oriented Dialogue System (DS), it is important to understand what the users' goals are for the conversations and whether those goals are successfully achieved. The success rate of goals directly correlates with user satisfaction and perceived usefulness of the DS. In this paper, we propose a novel automatic dialogue evaluation framework that jointly performs two tasks: goal segmentation and goal success prediction. We extend the RoBERTa-IQ model (Gupta et al., 2021) by adding multi-task learning heads for goal segmentation and success prediction. Using an annotated dataset from a commercial DS, we demonstrate that our proposed model reaches an accuracy that is on-par with single-pass human annotation comparing to a three-pass gold annotation benchmark.

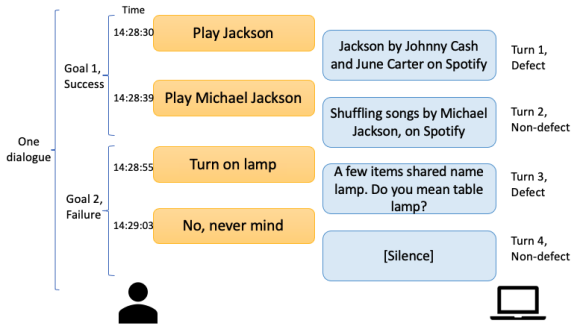

Figure 1: A dialogue that contains two goals: play-music and turn-on-device. Here, a turn is defined as one back-and-forth between the user and agent. A dialogue is a set of turns in quick succession as indicated by the timestamps. A goal represents an objective the user is trying to accomplish in a sequence of turns. Human annotation labels for turns (defect/non-defect) and goals (success/failure) are provided for reference.

1 Introduction

Today, commercial conversational AI assistants (e.g., Amazon Alexa, Apple Siri, and Google Assistant) are increasingly popular. However, it is challenging to establish reliable metrics to continuously measure the system performance in business reports and A/B experiments. The commonly used metrics in industry, e.g., monthly active users, dialogue count per user, and downstream impacts such as increased subscriptions and product sales etc., usually move slowly and are not suitable to measure the impact of functionality changes. On the other hand, the dissatisfaction metrics based on manual annotation, which can capture dissatisfying user experiences due to system errors, incomplete service coverage, or poor response quality etc., are sensitive to functionality changes, but not suitable for online monitoring and experimentation due to their offline nature. Automatic dialogue evaluation metrics (Schmitt and Ultes, 2015; Ling et al., 2020) aim to combine the benefits of the two types of metrics mentioned above by providing an online metric that measures user experience and is sensitive to functionality changes. The effectiveness of such metrics depends on how well they are correlated with the human annotations.

Automated dialogue evaluation metrics in the literature could be generally grouped into two categories (Deriu et al., 2020): 1) Interaction Quality (IQ) metrics (e.g., (Schmitt and Ultes, 2015; Ling et al., 2020; Gupta et al., 2021)) that measures user experiences at turn level; and 2) Dialogue Quality (DQ) metrics (e.g., (Sun et al., 2021)) that measures user experiences for the whole dialogue. However, we argue that, for multi-domain goal-oriented dialogue systems, there is a need to add a goal-level metric with granularity in between turn and dialogue. If we define a dialogue as a set of turns in quick succession, a user may switch her objective mid-dialogue without clear delineations. As shown by the example in Figure 1, the user started the dialogue with a playing music request but switched her objective halfway to turn on a smarthome device. Here, we propose a new granularity – goal, which

represents an objective the user is trying to accomplish in a sequence of turns. Goal segmentation is especially important for accurately evaluating the dialogue quality. In Figure 1 example, the user successfully achieved her first goal but failed on the second one. It would be ambiguous to evaluate the user experience at the whole dialogue level. To our best knowledge, there is no other work in the literature that performs goal segmentation and goal success prediction jointly. Prior work either assumes that a dialogue automatically terminates upon the completion of a goal e.g., (Bodigutla et al., 2020) or the goal boundaries are known beforehand e.g., (Walker et al., 1997).

To build a reliable automatic goal evaluation metric, we extend the RoBERTa-IQ model (Gupta et al., 2021) by adding multi-task learning heads for goal segmentation and success prediction. Even though RoBERTa-IQ model is a model built for turn level metrics, it encodes the turns before and after the reference turn (the turn on which the model will make prediction) to ingest dialogue context information. Besides, time difference between turns are encoded in the time bin embeddings and feed into the model together with the context embedding. Thus, such characteristics makes RoBERTa-IQ model a suitable starting point for goal level evaluation task. For goal segmentation, different from LSTM based method (Koshorek et al., 2018; Arnold et al., 2019), we leverage self-attention mechanism across turns. The practical benefit of this approach is that it improves inference efficiency on long token sequences, which is a common problem for RNN models. To demonstrate the effectiveness of our proposed methods, we evaluated our method against a golden dataset in which each dialogue is labeled by three annotators and the ground truth labels are determined by majority vote. We then compare the accuracy of the model predictions with that of a single-pass human annotation.

The proposed work has two main contributions: 1) A novel goal level dialogue evaluation framework that matches the real-world scenarios in multi-domain goal-oriented dialogue system; 2) A deep learning model that jointly learns goal segmentation and goal success prediction, with accuracy on-par with single-pass human annotation. The remainder of this paper is organized as the following: Section 2 reviews related work about goal segmentation and dialogue evaluation. Section 3 introduces the model architecture. Section 4 shows the experimental setup and discusses the performance of the model. We conclude the paper in Section 5.

2 Related Work

2.1 Topic Segmentation

To our best knowledge, our study is the first one to propose dialogue goal segmentation. Similar to the topic segmentation for generic text, dialogue goal segmentation aims to segment a dialogue into the goal-coherent units. Therefore, the previous approaches, which were originally proposed for generic text topic segmentation, are ready to be used for conversational corpora. Early topic segmentation approaches can be classified into two types: 1) lexical cohesion models and 2) content-oriented models. A well-known algorithm of lexical cohesion models is TextTiling (Hearst, 1997). Content-oriented models rely on the re-occurrence of patterns of topics, such as Bayesian Unsupervised Topic Segmentation (Eisenstein and Barzilay, 2008). More recently, neural network-based approaches (Koshorek et al., 2018; Arnold et al., 2019) are favored by researchers because of robust model performance and efficiency.

2.2 Dialogue Evaluation

Unsupervised methods There are some unsupervised evaluation methods, which provide a good assessment for open domain dialogues. RUBER (Tao et al., 2018) is a turn level metric that combines the relatedness between the turn level response and the previous issued query-response, respectively. Besides relying on semantics of the sentences, GRADE (Huang et al., 2020) proposes a method to leverage the graph embedded topic-level representation for turn level success evaluation. For dialogue level evaluation, MAUDE (Sinha et al., 2020) is a context aware model that measures the quality of a generated reply given the previous dialogue context.

Supervised methods The well known evaluation framework (Walker et al., 1997) based on user satisfaction is PARADISE (PARAdigm for DIalog System Evaluation) framework. In PARADISE, given a set of manually extracted input features and user ratings, a linear regression model is fitted to predict the user satisfaction. In contrast to evaluating the entire dialogue, there are various

approaches to evaluate the user satisfaction at turn level, such as Interaction Quality (IQ) (Schmitt and Ultes, 2015). To further generalized the model, IQ-NET (Ling et al., 2020) directly uses raw dialogue turn contents and system metadata without hand-crafted features. Meanwhile, by using post experience explicit user feedback as a proxy to user satisfaction, several joint turn and dialogue level evaluation methods (Bodigutla et al., 2020; Park et al., 2020) are proposed.

3 Methodology

We extend the RoBERTa-IQ (Gupta et al., 2021) model architecture by adding the goal segmentation task and goal success prediction task. As a brief recap, there are three key differences separating RoBERTa-IQ from the vanilla RoBERTa model, as illustrated in Figure 2:

1. RoBERTa-IQ introduces two special tokens [USER] and [AGENT] to delineate the beginning of each user request and agent response. This enables the model to encode the multi-turn dialogue as a flattened token sequence (see Table 1 for an example), which enables self-attention mechanism across different turns;

2. RoBERTa-IQ has a notion of "reference turn" versus "contextual turns". The reference turn is the target for the model prediction while contextual turns are context surrounding the reference turn. In this paper, we use two turns before and two after the reference turn as context (ablation study showed no statistical difference with larger context window);

3. Instead of the usual "position embedding" for RoBERTa model, RoBERTa-IQ has an embedding called "time bin embedding" that is calculated based on timestamp difference between a context turn and the reference turn. The time bin embedding has two uses: 1) it allows the model to understand the temporal relationship between the reference turn and contextual turns; 2) it enables the model to locate the reference turn – a special bin: BIN_0 is reserved for reference turn's tokens.

We add the goal segmentation and goal success prediction task heads on top of the [CLS] token embedding. For each turn in a dialogue, we mark it as the reference turn and let the model to learn from a binary classification label (B if the reference turn is at the beginning of a goal and I if inside a goal). Goal success prediction is modeled as a multi-class classifier with three possible outputs: success, failure, and unactionable.

[USER] Play Jackson
[AGENT] Jackson by Johnny Cash and June Carter on Spotify
[USER] Play Michael Jackson
[AGENT] Shuffling songs by Michael Jackson, on Spotify
[USER] turn on lamp
[AGENT] A few items shared name lamp. Do you mean table lamp?
[USER] No, never mind
[AGENT] {Silence}

Table 1: The example dialogue in Figure 1 in flattened token sequences form, considering turn-2 as the reference turn (bolded).

4 Experiments

4.1 Dataset

We used de-identified data for our experiments. The data is labeled by one annotator for two goal level tasks - goal segmentation and goal success prediction. The training dataset contains ~500K dialogues, randomly split into training (80%) and validation (20%) sets. 44% of those dialogues are single-turn dialogues. For all goals identified in the training dataset: 75% are success, 14% are unactionable and 11% are failure. An additional three-pass human annotation is applied to the evaluation dataset and the majority vote is used as golden labels. We call the evaluation dataset with golden label as golden dataset. The golden dataset is used to measure the performance of both human and the trained models. The single-pass annotation is considered as prediction of human. The golden dataset, which contains ~30K dialogues, has similar data distribution as the training dataset.

4.2 Experiment Setup

Implementation Details The model training and evaluation are implemented in PyTorch. We continue the training for 15 epochs and select the best model based on the performance on the validation dataset. For multi-task model, we add the two loss functions (cross-entropy loss) for both tasks with equal weights (i.e. 0.5). To compare with the model trained in the multi-task learning framework, we also train goal segmentation and goal success prediction models separately on the same dataset. Note, the model prediction score on different turns may be different. The goal success prediction is

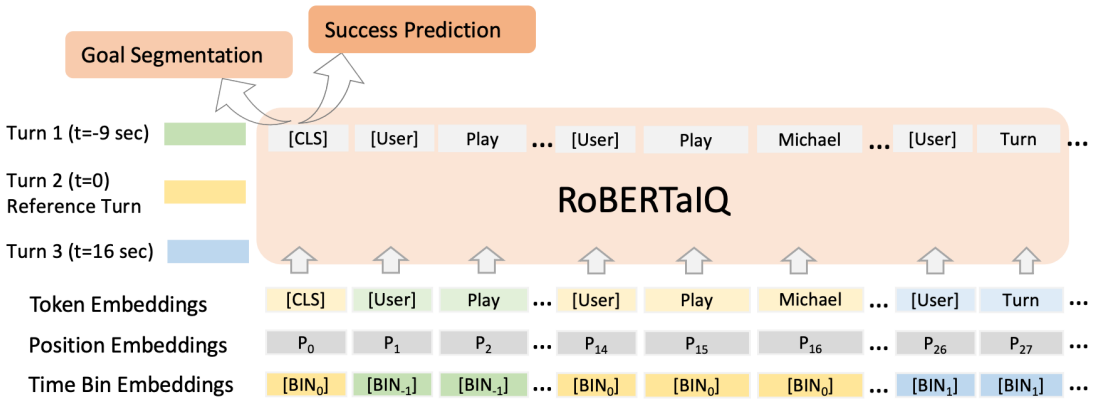

Figure 2: Diagram for RoBERTa-IQ model. The second turn with time bin equals to BIN_0 is the reference turn on which the model will make the prediction.

determined by the prediction on the last turn of the predicted goal.

Metrics We compare the performance of three models under different settings: human model (one-pass human annotation), single task model (two tasks trained separately), multi-task model (two tasks jointly trained). We measure model performance with the following metrics: 1) the accuracy of goal segmentation, which is defined as number of goals with correct boundaries divided by total number of goals; 2) the accuracy of goal evaluation, which is defined as the number of goals with correct boundaries and the right success prediction divided by the total number of goals; 3) weighted F1 score, which is the weighted average of the F1 score of each success prediction class on goals with accurate goal boundary. For machine-learned models, we also report relative metric (accuracy or weighted F1 score) with respect to that of human model in order to illustrate the difference. For example, the relative accuracy of human model is zero. The relative accuracy of single task model is computed as its accuracy minus the accuracy of human model.

4.3 Results

Table 2 summarizes the relative accuracy of segmentation and goal evaluation and relative weighted F1 score on single turn dialogues, multi-turn dialogues and these two combined. From the results, we can see the following points: 1) As shown in both accuracy and weighted F1 score, the performance of the multi-task model (the proposed model) is better than two single task models combined, especially, in multi-turn dialogues. 2) The proposed model has lower accuracy in goal segmentation but higher accuracy in goal evaluation compared to human. 3) The proposed model has higher accuracy but lower weighted F1 compared to human in goal evaluation. The proposed model is optimized for success class since that is the main usecase and has the most data for training while the proposed model has small performance gaps in failure class compared to human due to insufficient training data issue.

Dialogue	Model	Segmentation Accuracy	Goal Accuracy	Weighted F1 Score
Single turn	Human	+0.0%	+0.0%	+0.0%
	Single task	+0.0%	+1.1%	-1.3%
	Multi-task	+0.0%	**+1.4%**	**-1.0%**
Multi-turn	Human	+0.0%	+0.0%	+0.0%
	Single task	-4.7%	+2.2 %	-1.5%
	Multi-task	**-4.4%**	**+2.7%**	**-0.9%**
All	Human	+0.0%	+0.0%	+0.0%
	Single task	-3.8%	+2.0%	-1.4%
	Multi-task	**-3.5%**	**+2.4%**	**-0.9%**

Table 2: Performance of goal segmentation and goal evaluation on golden dataset. The best machine-learned model results are bolded.

5 Conclusion

In this paper, we propose a novel framework to evaluate goal-level performance for multi-domain goal-oriented dialogue systems and a deep learning model that jointly learns goal segmentation and success prediction. Our experiments show that the proposed model reaches an accuracy that is on-par with single-pass human annotation and with multi-task learning, the model performance is better than single task models for both tasks.

References

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Praveen Kumar Bodigutla, Aditya Tiwari, Josep Valls Vargas, Lazaros Polymenakos, and Spyros Matsoukas. 2020. Joint Turn and Dialogue level User Satisfaction Estimation on Multi-Domain Conversations. *arXiv e-prints*, page arXiv:2010.02495.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii. Association for Computational Linguistics.

Saurabh Gupta, Xing Fan, Derek Liu, Benjamin Yao, Yuan Ling, Kun Zhou, Tuan-Hung Pham, and Edward Guo. 2021. Robertaiq: An efficient framework for automatic interaction quality estimation of dialogue systems. In *KDD 2021 Workshop on Data-Efficient Machine Learning*.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pages 33–64.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *EMNLP*.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Yuan Ling, Benjamin Yao, Guneet Kohli, Tuan-Hung Pham, and Chenlei Guo. 2020. IQ-Net: A DNN model for estimating interaction-level dialogue quality with conversational agents. In *Proceedings of the KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption co-located with the 26TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2020), Virtual Workshop, August 24, 2020*, volume 2666 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Dookun Park, Hao Yuan, Dongmin Kim, Yinglei Zhang, Spyros Matsoukas, Young-Bum Kim, Ruhi Sarikaya, Chenlei Guo, Yuan Ling, Kevin Quinn, Tuan-Hung Pham, Benjamin Yao, and Sungjin Lee. 2020. Large-scale hybrid approach for predicting user satisfaction with conversational agents. In *NeurIPS 2020: Human in the Loop Dialogue Systems Workshop*.

Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts - and how it relates to user satisfaction. *Speech Commun.*, 74:12–36.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan J. Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. *ArXiv*, abs/2005.00583.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and M. de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.