Automatic Conversation Turn-Taking Segmentation in Semantic Facet

Dongin Jung
Department of Artificial Intelligence
Chung-Ang University
Seoul, South Korea
dongin1009@cau.ac.kr

Yoon-Sik Cho
Department of Artificial Intelligence
Chung-Ang University
Seoul, South Korea
yoonsik@cau.ac.kr

***Abstract*—Turn-taking is a significant aspect of a smooth conversation system. Detecting end-of-turn can be difficult for automatic conversation systems, and this can cause misleading conversation systems. To make a conversational system recognizing turn transition points, we propose a token-level turn-taking segmentation using linguistic features. This task imitates the automatic speech recognition environment by organizing several settings. Moreover, we utilize GPT-2, which is well known as a pretrained generative language model, to be able to predict in token-level live text stream. We evaluate our model compared to RNN series models in general conversation datasets and explore model prediction with test sample scenarios.**

***Index Terms*—Turn-taking Segmentation, Live Text Stream, Token Classification**

I. Introduction

A conversation aims to interact between humans or even human-computer through dialogue. Conversation participants exchange speaking and listening turns to communicate with each other. Typically, humans catch this turn transition and keep converse with an interlocutor(s). The end-of-turn occurs when the speaker finishes the intent of speaking and is ready to listen interlocutor speaking. However, it is hard to recognize the end-of-turn for conversational assistant systems such as chat assistants, voice assistants, and social robots. Addressing this limitation, turn-taking segmentation is a notable task in conversation systems and related approaches were presented. Previous turn-taking segmentation and end-of-turn prediction approaches used acoustic features for spoken dialogue systems [1]. Also, [2], [3] used linguistic features, however, they predicted sentence-level turn-taking without considering live conversation situations. Unlike these approaches, we focus on linguistic features rather than acoustic features to apply in the textual dialogue systems. Moreover, we utilize a unidirectional contextual language model using attention masking to suppose live conversation.

Bidirectional language models [4] are effective in representing words or subwords by understanding the context of sequences. Since our task supposes a live conversation, we utilize GPT-2 [5] which has a left-to-right manner. GPT models [5], [6] are specialized in text generation by using the decoder part of transformer [7] architecture. We modify the goal of GPT-2 [5] from the text generation task to our token-level turn-taking segmentation task.

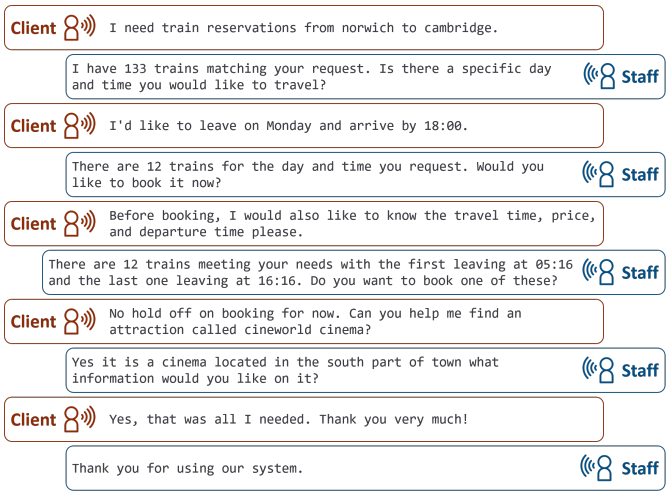

Fig. 1. A sample conversation in MultiWoZ [8] dataset. A conversation conducts giving and taking turns. This situation is turn-taking between client and staff to make a train reservation.

In this paper, we propose a novel token-level turn-taking segmentation task in the linguistic semantic facet designed to target a live conversation. To consider a live conversation situation, input text data are transformed to a speaking-oriented format, and model predictions are made at each token level. Detailed data transformation methods are discussed in Section III. To evaluate our task and model, we use conversation data designed so that speakers exchange their utterance turn-by-turn, named MultiWoZ [8] and DailyDialog [9]. In Figure 1, we display a sample situation of exchange turn to book a train in the MultiWoZ [8] dataset. The token-level turn-taking segmentation task is to predict each token to determine whether it is an end-of-turn or not. Our evaluation results are described in Section V.

II. Related Work

We investigated several pieces of research on turn-taking works, which focused on an acoustic stream segmentation or an utterance-level segmentation. Aldeneh *et al.* [1] made a prediction of a turn-switch statement between two sentences, if the speakers of both sentences are the same, then predict to *hold*; conversely, if the speakers are different, then

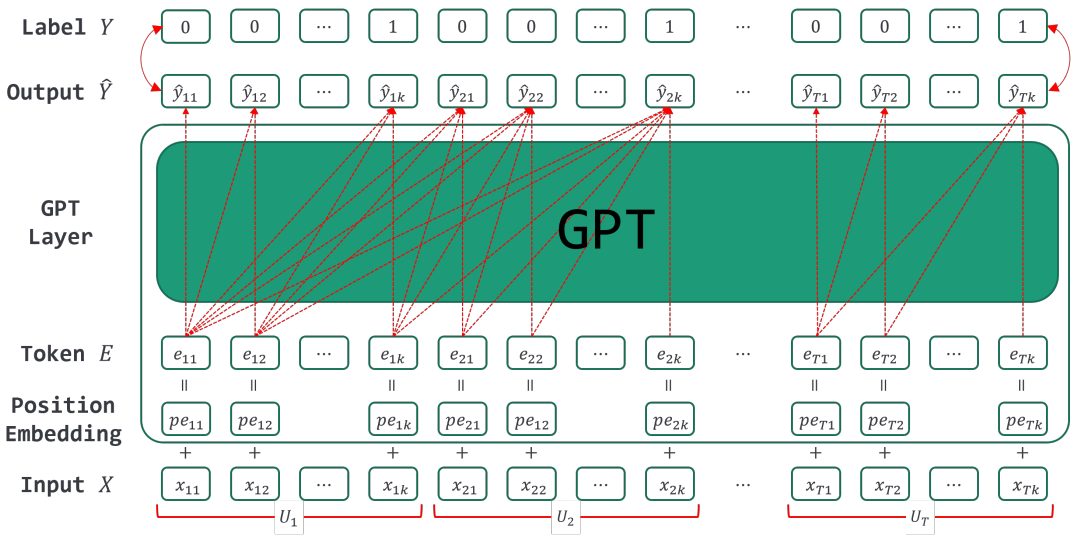

Fig. 2. A pipeline of token-level turn-taking segmentation. Input X is tokens from the tokenizer, and the model predicts in each token level whether the token is end-of-turn or not. In the current time point prediction, the future time point tokens are not used.

predict to *switch*. Acoustic feature sequences were inputted into the unidirectional LSTM and model predicted turn-switch statement. Additionally, they constructed a multitask learning model using speaker intent prediction. However, they used acoustic features from the speakers voice signal rather than the text utterance.

Liang *et al.* [2] used a transformer [7] encoder to predict turn-taking in the subtitle streams. They introduced a semantic recall gate to the transformer encoder for remembering historical information. For utterance embedding, they used BERT [4], which is one of the powerful language models to represent a text by encoding context bidirectionally. Also, they further pretrained BERT due to data discrepancy between the pretrained corpus of BERT and subtitle texts.

Su *et al.* [3] proposed a speaker clustering task with a dialogue act classification task as an auxiliary task. They computed sentence pairwise similarity and assigned it to a specific cluster. The similarity matrix was made in the bilinear form on two sentences and they suppose the number of speakers through the eigenvalue of similarity matrix in multiple-speaker datasets. Each cluster had utterances of the same speaker, and different clusters had utterances with different speakers.

Both [2], [3] used the linguistic features of dialogue, however, they focused on the relation between sentence pairs. They got text representation on the whole sentence level, and they used future tokens even speaker does not speak in the situation of live conversation. We explain our differentiation from these works in the following Section III.

III. Methods

The previous works [2], [3] employed BERT [4] to obtain utterance embedding. BERT has deep bidirectional contextualized representation by stacking the encoder part of transformer [10]. However, it cannot be used in a live speaking situation because of its bidirectional training method. From this restriction, we use GPT-2 [5] model to make the token-level prediction for time step-by-step. Our model is illustrated in Figure 2. For input data to the designed model, we construct the following data preprocessing method and data configuration.

A. Data Processing

To simulate live turn-taking segmentation, we set several settings and environments. We aim to imitate an automatic speech recognition (ASR) system [11].

1) Remove Punctuation: Punctuations are unconsiderable in signal to text writing and they can be a significant clue to segment sentences. Every sentence has a punctuation symbol(s) in the middle and end of the sentence. These symbols can notice the end of the turn rather than the semantic facet. Moreover, ASR system may not represent symbols in certain contexts. So, we remove all punctuations and special characters that contain the apostrophe (') and full stop (.) in text dialogue data.

2) Transform to Lowercase: ASR system [11] converts voice speech to corresponding texts. When a speaker speaks utterances, ASR system recognizes the spoken pronunciation, then converts it to text. However, it cannot recognize the context and letter case of speech, so almost text letters should be lowercase. In addition, since almost all the first letters in a sentence are written in uppercase, we transform all uppercase letters to lowercase to obviate the condition of letter case.

B. Input Data Configuration

A dialogue D consists of multiple utterances U as spoken by two speakers.

$$D = U_1, U_2, ..., U_T, \tag{1}$$

where T is a dynamic number of utterances per conversation. T is same as the frequency of turn-taking within a dialogue.

TABLE I
STATISTICS OF DATASETS USED IN OUR EXPERIMENT. WE CALCULATE THE AVERAGE AND STANDARD DEVIATION NUMBER OF TERMS, AND STANDARD DEVIATION DENOTED IN SUBSCRIPT.

	MultiWoZ	**DailyDialog**
# dialogue	10437	13118
# turns / dialogue	$14.61_{2.38}$	$7.85_{3.99}$
# tokens / dialogue	$202.56_{86.54}$	$103.64_{74.67}$
# tokens / turn	$13.71_{5.25}$	$13.02_{6.93}$

The speaker speaks an utterance that has one or more sentences in each turn. To input the GPT model, we tokenize utterances and concatenate tokens of all utterances within each dialogue. Consequently, end-of-turn labels make follow the corresponding token as the token finished turn or not. The last token label of every utterance is always 1 (end-of-turn), and the remainders are 0 (not end-of-turn).

$$\begin{aligned} X &= \{\psi_{tok}(U_1)||\psi_{tok}(U_2)||...||\psi_{tok}(U_T)\} \\ &= \{x_{11},...,x_{1k}\}||\{x_{21},...,x_{2k}\}||...||\{x_{T1},...,x_{Tk}\}, \end{aligned} \tag{2}$$

where ψ_{tok} indicates the tokenization function that returns output tokens x of the GPT tokenizer, and k is the maximum token number of each utterance. Unlike RNNs feeding tokens sequentially, transformer [7] families such as BERT [4] and GPT [5], [6] architectures process token sequences parallelly. So, BERT and GPT added position embedding to token embedding to obtain positional information on parallel tokens.

C. GPT-2 Model for Token-Level Prediction

GPTs [5], [6] and BERT [4] were trained in the large scale English corpus by self-supervised learning. BERT is a bidirectional encoder that represents the current token by computing attention with past and future tokens, while the live speaking situation does not have the future tokens. Unlike BERT [4], GPT models are the unidirectional manner model to train the context from left to right. To train the input text left-to-right, GPT uses the masked self-attention mechanism that masks behind the current token to avoid the engagement of future tokens. So, GPT-2 [5] is appropriate for our token-level prediction task, and we modified text generation that is GPT models pretraining task to turn-taking prediction task. The output tokens are fed to the two-layer linear classifier, and final binary predictions decide turn-taking segmentation.

IV. DATASET

Our data processing method and task can apply all the other dialogue datasets having turn structure which exists speaker dividing. Among them, we selected MultiWoZ 2.2 [8] and DailyDialog [9] which are the well-known conversation dataset. Statistics of these datasets are shown in Table I.

A. MultiWoZ

The Multi-Domain Wizard-of-Oz dataset (MultiWoZ) [8] is a dyadic conversation dataset of human-to-human interactions in situations of 8 domains. MultiWoZ has intents, consisting of categorical slots and non-categorical slots, such as a type of address, food, name, date, etc. The dataset contains dialogue, categorical and non-categorical types of span, and dialogue acts. Among them, we use only the dialogue part which was divided by utterance.

TABLE II
PERFORMANCE OF THE MODELS IN MULTIWOZ AND DAILYDIALOG.

Model	**MultiWoZ**				**DailyDialog**			
	lr	**R**	**P**	**F1**	lr	**R**	**P**	**F1**
GRU	1e-3	21.5	63.2	32.0	1e-4	28.7	58.1	38.4
LSTM	1e-3	33.9	65.7	44.7	1e-4	32.1	59.9	41.8
GPT-2	1e-5	**68.4**	**70.9**	**69.7**	1e-4	**60.7**	**62.9**	**61.8**

B. DailyDialog

The DailyDialog [9] is a daily life conversation dataset with classified 4 types of dialog acts and 6 types of emotions. The dataset contains dialog text with dialog act and dialog emotion, which are manually annotated. We use only the dialog part which was divided by utterance.

V. EXPERIMENTAL EVALUATION

A. Setup

To evaluate our novel task, we consist of comparable settings. Datasets are split into the train, validation, and test sets following the division of the original datasets provided [8], [9]. We implemented our data preprocess method, model, and task, based on PyTorch framework. The GPT-2 pretrained weights got from HuggingFace[1] platform. Additionally, we use AdamW [12] optimizer to optimize the model with initial learning rate in {1e-3, 1e-4, 1e-5}, and weight decay 1e-5. The model early stops training if validation loss does not increase while 5 epochs.

To compare model performance, we construct baselines using gated recurrent unit (GRU) [13] and long short-term memory (LSTM) [14] which are the RNN family. **GRU** and **LSTM** implement one unidirectional GRU and LSTM. We freeze the GPT-2 layer to gain token embedding values and feed them into GRU or LSTM. Finally, the output of GRU or LSTM is classified through a linear classification layer into binary classes.

B. Result

Table II shows the results of the model evaluation along with the learning rate of the minimum validation loss. Metrics are recall (**R**), precision (**P**), and micro f1-score (**F1**) to evaluate the model performance. Recall means the proportion of real end-of-turn cases that are correctly predicted end-of-turn points. Precision means the proportion of predicted end-of-turn point cases that are correctly real end-of-turns. Micro f1-score is a harmonic mean of recall and precision, which is commonly used in imbalanced class data. **GPT-2** showed the best performance compared to RNN family models.

[1]https://huggingface.co/gpt2

	actual conversation	GRU	LSTM	GPT-2
Multi WoZ	i need train reservations from nor wich to cam bridge // i have 133 trains matching your request is there a specific day and time you would like to travel // id like to leave on m onday and arrive by 1800 // there are 12 trains for the day and time you request would you like to book it now // before booking i would also like to know the travel time price and departure time please // there are 12 trains meeting your needs with the first leaving at 05 16 and the last one leaving at 16 16 do you want to book one of these // no hold off on booking for now can you help me find an attraction called c in eworld cinema // yes it is a cinema located in the south part of town what information would you like on it // yes that was all I needed thank you very much // thank you for using our system	i need train reservations from nor wich to cam bridge i have 133 trains matching your request is there a specific day and time you would like to travel // [70.31%] id like to leave on m onday and arrive by 1800 // [71.35%] there are 12 trains for the day and time you request would you like to book it now // [72.77%] before booking i would also like to know the travel time price and departure time please // [74.53%] there are 12 trains meeting your needs with the first leaving at 05 16 and the last one leaving at 16 16 do you want to book one of these // [69.79%] no hold off on booking for now // [71.62] can you help me find an attraction called c in eworld cinema yes it is a cinema located in the south part of town what information would you like on it // [71.96] yes that was all i needed thank you very much thank you for using our system	i need train reservations from nor wich to cam bridge // [89.71%] i have 133 trains matching your request is there a specific day and time you would like to travel // [93.70%] id like to leave on m onday and arrive by 1800 // [88.07%] there are 12 trains for the day and time you request would you like to book it now // [72.22%] before booking i would also like to know the travel time price and departure time please // [85.21%] there are 12 trains meeting your needs with the first leaving at 05 16 and the last one leaving at 16 16 do you want to book one of these no hold off on booking for now can you help me find an attraction called c in eworld cinema yes it is a cinema located in the south part of town what information would you like on it // [91.95] yes that was all i needed thank you very much thank you for using our system	i need train reservations from nor wich to cam bridge // [54.12%] i have 133 trains matching your request is there a specific day and time you would like to travel // [73.99%] id like to leave on m onday and arrive by 1800 // [70.06%] there are 12 trains for the day and time you request would you like to book it // [57.14%] now // [87.52%] before booking i would also like to know the travel time price and departure time please // [88.65%] there are 12 trains meeting your needs with the first leaving at 05 16 and the last one leaving at 16 16 do you want to book one of these // [78.43%] no hold off on booking for now can you help me find an attraction called c in eworld cinema // [71.26%] yes it is a cinema located in the south part of town what information would you like on it // [87.05%] yes that was all i needed thank you // [60.45%] very much // [65.44%] thank you for using our system
Daily Dialog	do you have maps of downtown area // yes here you are // how much is it // its free of charge // thanks so much //	do you have maps of downtown area yes here you are how much is it // [56.63%] its free // [57.91%] of charge // [58.73%] thanks so much // [52.60%]	do you have maps of downtown area // [72.55%] yes here you are // [68.82%] how much is it // [71.82%] its free of charge // [67.17%] thanks so much // [60.81%]	do you have maps of downtown area // [80.61%] yes here you are // [73.11%] how much is it // [52.79%] its free of charge // [64.57%] thanks // [58.60%] so much // [74.00%]

Fig. 3. Sample turn-taking segmentation output of models. Blue probabilities are denote turn ending probabilities, and reds denote incorrect predictions.

As shown in Figure 3, we extracted output prediction of each model in appropriate test samples. The blue-marked probabilities indicate the end-of-turn predicted probabilities of each model, and the red-marked tokens and probabilities indicate the incorrect prediction. The red-marked tokens with probability are incorrect predictions of end-of-turn, not actual end-of-turns. The red-marked tokens without probability are actual end-of-turns; however, the model did not catch them as end-of-turns. As a result, token-level segmentation performed quite well. Nevertheless, models are confused immediately before the end-of-turn. We think those terms can finish the turn in semantic context. In future work, we toward robust segmentation and multiple aspects based on this work.

VI. Conclusion

In this paper, we proposed a novel turn-taking segmentation task that uses linguistic features based on the GPT-2 model. We configure several environments and data preprocessing to mimic ASR. The token-level prediction has more challenging and unique, regardless, our simple and fundamental task configuration can be utilized in various ways of natural language processing. This task is expected to help conversation systems generate more fluent and smooth utterances.

Acknowledgement

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through the Korea Government (MSIT), Artificial Intelligence Graduate School Program, Chung-Ang University, under Grant 2021-0-01341; and in part by the Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Number: R2020040126).

References

[1] Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6159–6163.

[2] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[3] Z. Su and Q. Zhou, "Speaker clustering in textual dialogue with pairwise utterance relation and cross-corpus dialogue act supervision," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 734–744.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, and J. Chen, "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, ACL 2020*, 2020, pp. 109–117.

[9] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*, 2017.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] J. Levis and R. Suvorov, "Automatic speech recognition," *The encyclopedia of applied linguistics*, 2012.

[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[13] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," pp. 103–111, oct 2014.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.