Neural Dialogue Context Online End-of-Turn Detection

Ryo Masumura, Tomohiro Tanaka, Atsushi Ando
Ryo Ishii, Ryuichiro Higashinaka and Yushi Aono
NTT Media Intelligence Laboratories, NTT Corporation,
1-1, Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847, Japan
ryou.masumura.ba@hco.ntt.co.jp

Abstract

This paper proposes a fully neural network based dialogue-context online end-of-turn detection method that can utilize long-range interactive information extracted from both target speaker's and interlocutor's utterances. In the proposed method, we combine multiple time-asynchronous long short-term memory recurrent neural networks, which can capture target speaker's and interlocutor's multiple sequential features, and their interactions. On the assumption of applying the proposed method to spoken dialogue systems, we introduce target speaker's acoustic sequential features and interlocutor's linguistic sequential features, each of which can be extracted in an online manner. Our evaluation confirms the effectiveness of taking dialogue context formed by the target speaker's utterances and interlocutor's utterances into consideration.

1 Introduction

In human-like spoken dialogue systems, end-of-turn detection that determines whether a target speaker's utterance is ended or not is an essential technology (Sacks et al., 1974; Meena et al., 2014; Ward and Vault, 2015). It is widely known that heuristic end-of-turn detection based on non-speech duration determined by speech activity detection (SAD) is insufficient for smooth turn-taking (Hariharan et al., 2001).

Various methods have been examined for modeling the end-of-turn detection (Koiso et al., 1998; Shriberg et al., 2000; Schlangen, 2006; Gravano and Hirschberg, 2011; Sato et al., 2002; Guntakandla and Nielsen, 2015; Ferrer et al., 2002, 2003; Atterer et al., 2008; Arsikere et al., 2014, 2015). A general approach is discriminative modeling using acoustic or linguistic features extracted from target speaker's current utterance. In addition, recent studies use recurrent neural networks (RNNs) as they are suitable for directly capturing long-range sequential features without manual specification of fixed length features such as maximum, minimum, average values of acoustic features or bag-of-words features (Masumura et al., 2017; Skantze, 2017)

We note, however, that interlocutor's utterances are rarely used for end-of-turn detection. In dialogues, target speaker's utterances are definitely impacted by the interlocutor's utterances (Heeman and Lunsford, 2017). It is expected that we can improve end-of-utterance detection performance by capturing the "interaction" between the target speaker and the interlocutor.

In this paper, we propose a neural dialogue-context online end-of-turn detection method that can flexibly utilize both target speaker's and interlocutor's utterances. To the best of our knowledge, this paper is the first study to utilize dialogue-context information for neural end-of-turn detection. Although some natural language processing tasks recently examine dialogue-context modeling (Liu and Lane, 2017; Tran et al., 2017), they cannot handle multiple acoustic and lexical features individually extracted from both target speaker's and interlocutor's utterances. In the proposed method, target speaker's and interlocutor's multiple sequential features, and their interactions are captured by stacking multiple time-asynchronous long short-term memory RNNs (LSTM-RNNs). In order to achieve low-delayed end-of-turn detection in spoken dialogue systems, acoustic sequential features extracted from target speaker's speech and linguistic sequential features extracted from the interlocutor's (system's) responses are used for capturing interactive information.

Proceedings of the SIGDIAL 2018 Conference, pages 224–228,
Melbourne, Australia, 12-14 July 2018.

In our experiments, human-human contact center dialogue data sets are used with the goal of constructing a human-like interactive voice response system. We show that the proposed method outperforms a variant that uses only target speaker's utterances.

2 Proposed Method

End-of-turn detection is the problem of detecting whether each end-of-utterance point is a turn-taking point or not. The utterance is defined as an internal pause unit (IPU) if it is surrounded by non-speech units (Koiso et al., 1998). The speech/non-speech units are estimated by SAD.

In dialogue-context-based online end-of-turn detection, all past information of both target speaker's and interlocutor's utterances behind the speaker's current end-of-utterance can be utilized for extracting context information. The estimated label is either end-of-turn or not. The label of the t-th target speaker's end-of-utterance in a conversation can be decided by:

$$\hat{l}^{(t)} = \underset{l^{(t)} \in \{0,1\}}{\operatorname{argmax}} P(l^{(t)} | \boldsymbol{S}^{(1:t)}, \boldsymbol{C}^{(1:t)}, \boldsymbol{\Theta}), \quad (1)$$

where $\boldsymbol{\Theta}$ denotes a model parameter. $\hat{l}^{(t)}$ is the estimated label of the t-th speaker's end-of-utterance. $\boldsymbol{S}^{(1:t)}$ represents speaker's utterances $\{\boldsymbol{S}^{(1)}, \cdots, \boldsymbol{S}^{(t)}\}$ where $\boldsymbol{S}^{(t)}$ is the t-th utterance. $\boldsymbol{C}^{(1:t)}$ represents interlocutor's utterances $\{\boldsymbol{C}^{(1)}, \cdots, \boldsymbol{C}^{(t)}\}$ where $\boldsymbol{C}^{(t)}$ is the t-th utterance that occurred just before $\boldsymbol{S}^{(t)}$. Undoubtedly, there are some exceptional cases wherein the t-th interlocutor's utterance is none.

The t-th speaker's utterance involves N kinds of sequential features:

$$\boldsymbol{S}^{(t)} = \{\boldsymbol{s}_1^{(t)}, \cdots, \boldsymbol{s}_N^{(t)}\}, \quad (2)$$

$$\boldsymbol{s}_n^{(t)} = \{\boldsymbol{a}_{n,1}^{(t)}, \cdots, \boldsymbol{a}_{n,I_n^t}^{(t)}\}, \quad (3)$$

where $\boldsymbol{s}_n^{(t)}$ represents the n-th sequential feature in $\boldsymbol{S}^{(t)}$, and $\boldsymbol{a}_{n,i}^t$ is the i-th frame's feature in $\boldsymbol{s}_n^{(t)}$. I_n^t is the length of $\boldsymbol{s}_n^{(t)}$. In the same way, the t-th interlocutor's utterance involves M kinds of sequential features:

$$\boldsymbol{C}^{(t)} = \{\boldsymbol{c}_1^{(t)}, \cdots, \boldsymbol{c}_M^{(t)}\}, \quad (4)$$

$$\boldsymbol{c}_m^{(t)} = \{\boldsymbol{b}_{m,1}^{(t)}, \cdots, \boldsymbol{b}_{m,J_m^t}^{(t)}\}, \quad (5)$$

where $\boldsymbol{c}_m^t$ represents the m-th sequential feature in $\boldsymbol{C}^{(t)}$, and $\boldsymbol{b}_{m,j}^{(t)}$ is the j-th frame's feature in $\boldsymbol{c}_m^{(t)}$. J_m^t is a length of $\boldsymbol{c}_m^{(t)}$.

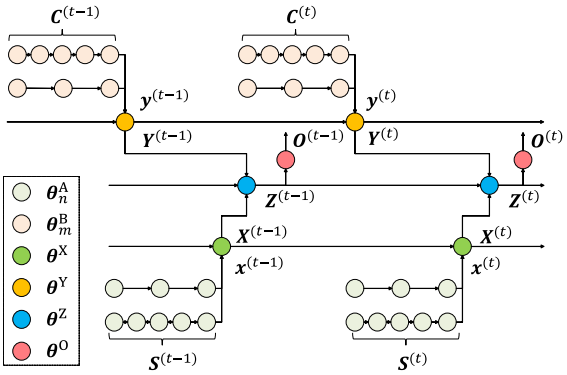

Figure 1: Model structure of neural dialogue-context online end-of-turn detection.

2.1 Fully Neural Network based Modeling

This paper proposes a neural dialogue context online end-of-turn detection method that is modeled using fully neural networks. In order to model $(l^{(t)}|\boldsymbol{S}^{(1:t)}, \boldsymbol{C}^{(1:t)}, \boldsymbol{\Theta})$, we extend stacked time asynchronous sequential networks that include multiple time-asynchronous LSTM-RNNs for embedding complete sequential information into a continuous representation (Masumura et al., 2017). In order to capture long-range dialogue context information, the proposed method employs two stacked time asynchronous sequential networks for both target speaker's and interlocutor's utterances. In addition, the proposed method introduces another sequential network to capture interactions of both side's utterances.

Figure 1 details the structure of the proposed method. In the proposed method, each feature within an utterance is individually embedded into a continuous representation in an asynchronous manner. To this end, LSTM-RNNs are prepared for individual sequential features in both target speaker's and interlocutor's utterances. Each sequential information is embedded as:

$$\boldsymbol{A}_n^{(t)} = \mathtt{LSTM}(\boldsymbol{a}_{n,1}^{(t)}, \cdots, \boldsymbol{a}_{n,I_n^t}^{(t)}; \boldsymbol{\theta}_n^{\mathtt{A}}), \quad (6)$$

$$\boldsymbol{B}_m^{(t)} = \mathtt{LSTM}(\boldsymbol{b}_{m,1}^{(t)}, \cdots, \boldsymbol{b}_{m,J_m^t}^{(t)}; \boldsymbol{\theta}_m^{\mathtt{B}}), \quad (7)$$

where $\boldsymbol{A}_n^{(t)}$ denotes a continuous representation that embeds the n-th sequential feature within the t-th target speaker's utterance. $\boldsymbol{B}_m^{(t)}$ denotes a continuous representation that embeds the n-th sequential feature within the t-th interlocutor's utterance. $\mathtt{LSTM}()$ represents a function of the unidirectional LSTM-RNN layer. $\boldsymbol{\theta}_n^{\mathtt{A}}$ and $\boldsymbol{\theta}_m^{\mathtt{B}}$ are model parameters for the n-th sequence in the target speaker's utterance and the m-th sequence in

the interlocutor's utterance, respectively.

The continuous representations individually formed from each sequential feature are merged to yield an utterance-level continuous representation as follows:

$$\boldsymbol{x}^{(t)} = [\boldsymbol{A}_1^{(t)\top}, \cdots, \boldsymbol{A}_N^{(t)\top}]^\top, \quad (8)$$

$$\boldsymbol{y}^{(t)} = [\boldsymbol{B}_1^{(t)\top}, \cdots, \boldsymbol{B}_M^{(t)\top}]^\top, \quad (9)$$

where $\boldsymbol{x}^{(t)}$ and $\boldsymbol{y}^{(t)}$ represent utterance-level continuous representations for the t-th target speaker's utterance and the t-th interlocutor's utterance, respectively.

In order to capture long-range contexts, target speaker's utterance-level continuous representations and interlocutor's utterance-level continuous representations are individually embedded into a continuous representation. The t-th continuous representation that embeds a start-of-dialogue and the current end-of-utterance is defined as:

$$\boldsymbol{X}^{(t)} = \texttt{LSTM}(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(t)}; \boldsymbol{\theta}^{\texttt{X}}), \quad (10)$$

$$\boldsymbol{Y}^{(t)} = \texttt{LSTM}(\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(t)}; \boldsymbol{\theta}^{\texttt{Y}}), \quad (11)$$

where $\boldsymbol{X}^{(t)}$ denotes a continuous representation that embeds speaker's utterances behind the t-th speaker's end-of-utterance, and $\boldsymbol{Y}^{(t)}$ denotes a continuous representation that embeds interlocutor's utterances behind the t-th interlocutor's end-of-utterance. $\boldsymbol{\theta}^{\texttt{X}}$ and $\boldsymbol{\theta}^{\texttt{Y}}$ are model parameters for the target speaker's utterance-level LSTM-RNN and the interlocutor's utterance-level LSTM-RNN, respectively.

In addition, to consider the interaction between the target speaker and the interlocutor, both utterance-level continuous representations are additionally summarized as:

$$\boldsymbol{z}^{(t)} = [\boldsymbol{X}^{(t)\top}, \boldsymbol{Y}^{(t)\top}]^\top, \quad (12)$$

$$\boldsymbol{Z}^{(t)} = \texttt{LSTM}(\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(t)}; \boldsymbol{\theta}^{\texttt{Z}}), \quad (13)$$

where $\boldsymbol{Z}^{(t)}$ denotes a continuous representation that embeds all dialogue context sequential information behind the t-th target speaker's end-of-utterance. $\boldsymbol{\theta}^{\texttt{Z}}$ represents the model parameter.

In an output layer, posterior probability of end-of-turn detection in the t-th target speaker's end-of-utterance is defined as:

$$\boldsymbol{O}^{(t)} = \texttt{SOFTMAX}(\boldsymbol{Z}^{(t)}; \boldsymbol{\theta}^{\texttt{O}}), \quad (14)$$

where $\texttt{SOFTMAX}()$ is a softmax function, and $\boldsymbol{\theta}^{\texttt{O}}$ is a model parameter for the softmax function. $\boldsymbol{O}^{(t)}$ corresponds to $P(l^{(t)}|\boldsymbol{S}^{(1:t)}, \boldsymbol{C}^{(1:t)}, \boldsymbol{\Theta})$. Summarizing the above, $\boldsymbol{\Theta}$ is represented as $\{\boldsymbol{\theta}_1^{\texttt{A}}, \cdots, \boldsymbol{\theta}_N^{\texttt{A}}, \boldsymbol{\theta}_1^{\texttt{B}}, \cdots, \boldsymbol{\theta}_M^{\texttt{B}}, \boldsymbol{\theta}^{\texttt{X}}, \boldsymbol{\theta}^{\texttt{Y}}, \boldsymbol{\theta}^{\texttt{Z}}, \boldsymbol{\theta}^{\texttt{O}}\}$. In training, the parameter can be optimized by minimizing the cross entropy between a reference probability and an estimated probability:

$$\hat{\boldsymbol{\Theta}} = \operatorname*{argmin}_{\boldsymbol{\Theta}} - \sum_{d \in \mathcal{D}} \sum_{t=1}^{T_d} \sum_{l \in \{0,1\}} \hat{\boldsymbol{O}}_{l,d}^{(t)} \log \boldsymbol{O}_{l,d}^{(t)}, \quad (15)$$

where $\hat{\boldsymbol{O}}_{l,d}^{(t)}$ and $\boldsymbol{O}_{l,d}^{(t)}$ are a reference probability and an estimated probability of label l for the t-th end-of-utterance in the d-th conversation, respectively. $\mathcal{D}$ represents a training data set.

2.2 Features for Spoken Dialogue Systems

In neural dialogue-context-based online end-of-turn detection, various sequential features can be leveraged for capturing both target speaker's and interlocutor's utterances. In spoken dialogue systems, the interlocutor is the system. Therefore, lexical information generated by the system's response generation module can be utilized. This paper uses pronunciation sequences and word sequences as the interlocutor's sequential features. In the proposed modeling, we use both symbol sequences by converting them into continuous vectors. On the other hand, the target speaker's utterances are speech. This paper introduces fundamental frequencies (F0s), and senone bottleneck features inspired by Masumura et al. (2017). The senone bottleneck features, which extract phonetic information as continuous vector representations, offer strong performance without recourse to lexical features.

3 Experiments

This paper employed Japanese simulated contact center dialogue data sets instead of human-computer dialogue data sets. The data sets include 330 dialogues and 6 topics. One dialogue means one telephone call between one operator and one customer, in which each speaker's speech was separately recorded. In order to simulated interactive voice response applications, we regard the operator as the interlocutor, and the customer as the target speaker. We divided each data set into speech units and non-speech units using an LSTM-RNN based SAD (Eyben et al., 2013) trained using various Japanese speech data. An utterance is defined as a unit surrounded by non-speech units whose

	Speaker's features	Interlocutor's features	Dialogue context	Recall	Precision	F-value	Accuracy
(1).	F0	-	-	80.4	69.9	74.8	73.4
(2).	SENONE	-	-	82.7	78.3	80.4	80.3
(3).	F0+SENONE	-	-	84.5	77.4	80.8	80.6
(4)	-	PRON	-	46.2	64.9	54.0	61.3
(5).	-	WORD	-	66.1	64.6	65.4	65.3
(6).	-	PRON+WORD	-	68.3	64.1	66.2	65.9
(7).	SENONE	WORD	√	82.0	80.5	81.2	81.4
(8).	F0+SENONE	PRON+WORD	√	**82.7**	**81.4**	**82.1**	**82.0**

Table 2: *Experimental results: Recall (%), Precision (%), F-value (%), and Accuracy (%).*

Topics	#calls	#utterances	#turns
Finance	50	3,991	2,166
Internet provider	64	3,860	1,799
Local government unit	58	3,741	1,598
Mail-order	52	3,752	1,828
PC repair	45	2,838	1,934
Mobile phone	61	4,453	2.016
Total	330	22,635	11,341

Table 1: *Experimental data sets.*

duration is more than 100 ms. Turn-taking points and backchannel points were manually annotated for all dialogues. The evaluation used 6-fold cross validation in which training and validation data were 5 topics and test data were 1 topic. Detailed setups are shown in Table 1 where #calls, #utterances, and #turns represent number of calls, utterances and end-of-turn points, respectively.

To realize a comprehensive evaluation, we examined various conditions. In the proposed modeling, unit size of LSTM-RNNs was unified to 256. For training, the mini-batch size was set to 2 calls. The optimizer was Adam with the default setting. Note that a part of the training sets were used as the data sets employed for early stopping. We constructed five models by varying an initial parameter for individual conditions and evaluated the average performance. When using either target speaker's utterances or interlocutor's utterances, required components were only used for building the proposed modeling. We used following sequential features. F0 represents 2 dimensional sequential features of F0 and ΔF0; frame shift was set to 5 ms. SENONE represents 256-dimensional senone bottleneck features extracted from 3-layer senone LSTM-RNN with 256 units trained from a corpus of spontaneous Japanese speech (Maekawa et al., 2000). Its frame shift was set to 10 ms, and the bottleneck layer was set to the third LSTM-RNN layer. PRON represents pronunciation sequences, and WORD represents word sequences of interlocutor's utterances. The lexical features were introduced by converting them into 128 dimensional vectors through linear transformation that was also optimized in training.

3.1 Results

Table 2 shows the experimental results. We used the evaluation metrics of recall, precision, macro F-value, and accuracy. The results gained when using only target speaker's utterances are shown in (1)-(3). In terms of F-value and accuracy, (3) outperformed (1) and (2). This confirms that stacked time-asynchronous sequential network based modeling is effective for combining multiple sequential features. The results gained when using only interlocutor's utterances are shown in (4)-(6). Among them, (6) attained the best performance although its performance was inferior to (1)-(3). In fact, (4)-(6) outperformed random end-of-turn decision making. This indicates interlocutor's utterances are effective in improving online end-of-turn detection performance. The proposed method, which takes both target speaker's and interlocutor's utterances into consideration, is shown in (7) and (8). In terms of F-value and accuracy, (7) outperformed (2) and (5). These results indicate that interaction information is effective for detecting end-of-turn points. The best results were attained by (8), which utilized both multiple target speaker's features and multiple interlocutor's features. The sign test results verified that (8) achieved statistically significant performance improvement ($p < 0.05$) over (3).

4 Conclusions

In this paper, we proposed a neural dialogue context online end-of-turn detection method. Main advance of the proposed method is taking long-range interaction information between target speaker's and interlocutor's utterances into consideration. In experiments using contact center dialogue data sets, the proposed method, which leveraged both target speaker's multiple acoustic features and interlocutor's multiple lexical features, achieved significant performance improvement compared to a method that only utilized target speaker's utterances.

References

Harish Arsikere, Elizabeth Shriberg, and Umut Ozertem. 2014. Computationally-efficient endpointing features for natural spoken interaction with personal-assistant systems. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3241–3245.

Harish Arsikere, Elizabeth Shriberg, and Umut Ozertem. 2015. Enhanced end-of-turn detection for speech to a personal assistant. *In Proc. AAAI Spring Symposium, Turn-Taking and Coordination in Human-Machine Interaction*, pages 75–78.

Michaela Atterer, Timo Baumann, and David Schlangen. 2008. Towards incremental end-of-utterance detection in dialogue systems. *In Proc. International Conference on Computational Linguistics (COLING)*, pages 11–14.

Florian Eyben, Felix Weninger, Stefano Squartini, and Bjorn Schuller. 2013. Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 483–487.

Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. 2002. In the speaker done yet? faster and more accurate end-of-utterance detection using prosody in human-computer dialog. *In Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 2061–2064.

Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 608–611.

Agustin Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25:601–634.

Nishitha Guntakandla and Rodney D. Nielsen. 2015. Modelling turn-taking in human conversations. *AAAI Spring Symposium, Turn-Taking and Coordination in Human-Machine Interaction*, pages 17–22.

Ramalingam Hariharan, Juha Hakkinen, and Kari Laurila. 2001. Robust end-of-utterance detection for real-time speech recognition applications. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 249–252.

Peter A Heeman and Rebecca Lunsford. 2017. Turn-taking offsets and dialogue context. *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1671–1675.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, 41:295–321.

Bing Liu and Ian Lane. 2017. Dialogue contect language modeling with recurrent neural networks. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5715–5719.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. *In proc. International Conference on Language Resources and Evaluation (LREC)*, pages 947–952.

Ryo Masumura, Taichi Asami, Hirokazu Masataki, Ryo Ishii, and Ryuichiro Higashinaka. 2017. Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1661–1665.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2014. Data-driven models for timing feedback responses in a map task dialogue system. *Computer Speech and Language*, 28:903–922.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplet systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa. 2002. Learning decison trees to determine turn-taking by spoken dialogue systems. *In Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 861–864.

David Schlangen. 2006. From reaction to prediction: Experiments with computational models of turn taking. *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 17–21.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tur, and Gukhan Tur. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32:127–154.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. *In Proc. Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 220–230.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. *In Proc. Conference of the European Chapter of the Association for Computational Linguistics*, 1:428–437.

Nigel G. Ward and David De Vault. 2015. Ten challenges in highly-interactive dialog systems. *AAAI Spring Symposium, Turn-Taking and Coordination in Human-Machine Interaction*, pages 104–107.