Online Call Scene Segmentation of Contact Center Dialogues based on Role Aware Hierarchical LSTM-RNNs

Ryo Masumura*, Setsuo Yamada*, Tomohiro Tanaka*,
Atsushi Ando*, Hosana Kamiyama* and Yushi Aono*
* NTT Media Intelligence Laboratories, NTT Corporation, Japan
E-mail: ryou.masumura.ba@hco.ntt.co.jp

***Abstract*—This paper proposes novel online call scene segmentation methods for contact center dialogues between an operator and a customer. Call scene segmentation is useful for many useful tasks such as summarization and information retrieval. In addition, its online implementation is beneficial in constructing operator assist systems. We define call scene segmentation as utterance-level sequential labeling with five tag types: opening, requirement confirmation, response, customer confirmation, and closing. In order to perform online call scene segmentation precisely, it is essential to capture interactions between the operator and the customer. In fact, dialogues usually include several sets of interactions; for example, the customer raises a question and the operator answers it, or the customer reveals own personal information and the operator repeats it. In addition, each scene shows some tendencies of interactions; for example, customer confirmation includes operator repetitions more frequently than other scenes. To capture the interactions, we present novel fully neural network based methods called role-aware hierarchical long short-term memory recurrent neural networks. The proposed methods can capture long-range interactive information by simultaneously dealing with, in an online manner, both the sequence of sentences and the sequence of speaker role labels. An experiment on Japanese simulated contact center dialogue data sets demonstrates the effectiveness of the proposed methods.**

I. Introduction

With the progress of automatic speech recognition technologies, utilization of the speech segments being captured by contact centers is increasing. Contact center dialogues include customer needs or typical annoyances, and their information is very effective in improving business performance. In this paper, we focus on call scene segmentation which splits contact center dialogues into scene categories. Of course, extracting the customer's annoyances or customer's personal information is highly useful if not essential. In addition, segmentation is beneficial for subsequent tasks such as summarization, and information retrieval.

Call scene segmentation is closely related to topic segmentation, a target of many studies. The two basic approaches to segmentations are unsupervised methods and supervised methods. Text-tiling which uses passage coherence to detect topic change boundaries is a typical unsupervised method [1]–[3]. More recent unsupervised methods use latent variable models such as hidden Markov models or topic models [4]–[7]. On the other hand, recent supervised methods construct sequential labeling models [8]–[10]. Deep learning based supervised methods have been shown to offer superior performance to unsupervised methods but topic annotated data sets are necessary.

This paper aims to develop supervised methods for call scene segmentation since fixed call scenes can be defined for call center dialogues. It is reasonable to split call scenes into the following five main parts; opening, requirement confirmation, response, customer confirmation, and closing. In addition, we aim to support online segmentation so as to determine scene categories in real-time; the chief goal is to enhance real time operator assist systems.

In order to precisely perform online call scene segmentation, it is essential to capture the interactions between the operator and the customer. In fact, previous supervised topic segmentation methods focused on just single speaker tasks [8]–[10]. This is despite the fact that contact center dialogues include several sets of interactions; for example, the customer raises a question and the operator answers it, or the customer reveals own personal information and the operator repeats it. In addition, each scene shows some tendencies of interactions; for example, customer confirmation includes operator repetitions more frequently than other scenes. Capturing these interactions is expected to improve segmentation performance.

In this paper, we propose fully neural network based online call scene segmentation methods that can take the interactions between speakers into consideration. Our idea is to simultaneously process sentence sequences and speaker role sequences in the dialogues. We implement the idea as role aware hierarchical long short-term memory recurrent neural networks that can consider long-range interaction information between the speakers by exploiting speaker role awareness. In order to extract the speaker role awareness, this paper examines two representations; one-hot vector representations and continuous vector representations. Previous studies examined neural network based methods that can consider interactions in conversations or dialogues for document classification and language modeling [11], [12]. To the best of our knowledge, this paper is the first to detail segmentation methods that can consider the interactions. An experiment on simulated

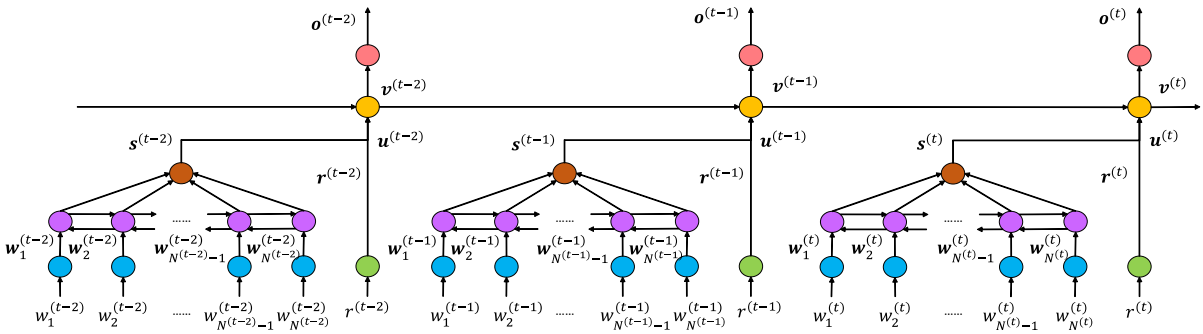

Fig. 1. Network structure of role-aware LSTM-RNNs.

Japanese contact center dialogue data sets demonstrates the effectiveness of the proposed methods.

This paper is organized as follows. Section 2 describes work related to call scene segmentation. We define call scenes in Section 3. Section 4 details our proposed method. In Section 5, we describe the evaluation conducted and the results gained. Section 6 concludes this paper.

II. Related Work

A lot of studies have examined the processing of contact center dialogues. The most common research topic is summarization of dialogues [13]–[15]. In addition, information retrieval technologies have been investigated [16]. Anger detection and customer satisfaction estimation were also examined in the field of contact center enhancement [17], [18]. No study has considered supervised call scene segmentation although summarization and information retrieval are similar to call scene segmentation. We can expect that call scene segmentation will be highly useful in advancing other contact center technologies.

Call scene segmentation is related to online sentence boundary detection and end-of-turn detection. Both sentence boundary detection and end-of-turn detection estimate whether a speaker's utterance is ended or not [19]–[22], so they address the two-class sequential labeling problem, whereas call scene segmentation involves multi-class sequential labeling. While some sequential labeling methods employed hierarchal LSTM-RNNs [9], [10], [22], role-aware methods that can take interactions between speakers had been not examined.

III. Call Scenes in Contact Center Dialogues

This section defines call scenes in contact center dialogues as the utterance sets created during the dialogue between a contact center operator and a customer. In this work, we split call scenes as follows.

1) **Opening**: Utterance sets from start-of-dialogue to end-of-introduction.
2) **Requirement confirmation**: Utterance sets wherein the operator fully understands the customer's requirements.
3) **Response**: Utterance sets where operator responds to the requirements.
4) **Customer confirmation**: Utterance sets where operator confirms customer's personal contact information such as name, address, and telephone number.
5) **Closing**: Utterance sets from closing introduction to end-of-dialogue.

IV. Proposed Methods

This section details our online call scene segmentation methods. The T utterance units of a dialogue are automatically determined by speech activity detection (SAD). The dialogue consists of sentence sequence $\boldsymbol{S} = \{s^{(1)}, \cdots, s^{(T)}\}$ and speaker role label sequence $\boldsymbol{R} = \{r^{(1)}, \cdots, r^{(T)}\}$ where $s^{(t)}$ represents word sequence $\{w_1^{(t)}, \cdots, w_{N^{(t)}}^{(t)}\}$ and $N^{(t)}$ is number of words in the t-th utterance. In processing contact center dialogues, speaker role r^t is either operator or customer. Call scene segmentation is to assign call scene labels $\boldsymbol{L} = \{l^{(1)}, \cdots, l^{(T)}\}$ to the utterance sequence where $l^{(t)} \in \mathcal{L}$ and $\mathcal{L}$ is sets of scene labels. Thus, call scene segmentation is regarded as utterance-level sequential labeling problem.

In online call scene segmentation, the t-th call scene label $l^{(t)}$ is estimated from $\boldsymbol{S}^{(1:t)} = \{s^{(1)}, \cdots, s^{(t)}\}$ and $\boldsymbol{R}^{(1:t)} = \{r^{(1)}, \cdots, r^{(t)}\}$. For this, we model conditional probability $P(l^{(t)}|\boldsymbol{S}^{(1:t)}, \boldsymbol{R}^{(1:t)}, \boldsymbol{\Theta})$ where $\boldsymbol{\Theta}$ represents a model parameter. The t-th call scene can be categorized by:

$$\hat{l}^{(t)} = \arg\max_{l^{(t)} \in \mathcal{L}} P(l^{(t)}|\boldsymbol{S}^{(1:t)}, \boldsymbol{R}^{(1:t)}, \boldsymbol{\Theta}). \quad (1)$$

In this paper, $P(l^{(t)}|\boldsymbol{S}^{(1:t)}, \boldsymbol{R}^{(1:t)}, \boldsymbol{\Theta})$ is modeled by role-aware hierarchical long short-term memory recurrent neural networks (LSTM-RNNs) which are based on fully neural networks. Online call scene segmentation can be performed in an incremental manner.

A. Role-Aware Hierarchical LSTN-RNNs

Role-aware hierarchical LSTM-RNNs must simultaneously handle the sentence sequence and speaker role sequence. To this end, we introduce word-level LSTM-RNN and utterance-level LSTM-RNN. Word-level LSTM-RNN converts a word sequence into a continuous vector representation. We introduce self-attention bidirectional LSTM-RNN as the word-level LSTM-RNN [23]. Utterance-level LSTM-RNN handles

TABLE I
Experimental data sets.

Topics	#calls	#words	#utterances				
			Opening	Requrement confirmation	Response	Customer confirmation	Closing
Finance	59	55,933	140	542	3,027	2,023	349
Internet provider	57	47,668	112	346	1,901	1,221	235
Government unit	73	48,998	191	794	4,118	181	333
Mail-order	56	46,574	116	509	2,152	1,815	346
PC repair	55	55,101	126	664	3,824	1,319	330
Mobile phone	61	51,061	126	451	4,090	748	323
Total	361	305,351	811	3,306	19,112	7,307	1,916

a concatenated vector of speaker role representation and the sentence representation. For speaker role representation, this paper examines both one-hot vector representation and continuous vector representation. Fig. 1 shows the network structure of the role-aware hierarchical LSTM-RNNs.

In order to compose a sentence representation from words, each word in the sentence is first converted into a continuous representation. The continuous vector representation of the n-th word in the t-th sentence is given by:

$$\boldsymbol{w}_n^{(t)} = \texttt{EMBED}(w_n^{(t)}; \boldsymbol{\theta}^{\mathtt{w}}), \tag{2}$$

where $\texttt{EMBED}()$ is a linear transformational function to embed a symbol into a continuous vector and $\boldsymbol{\theta}^{\mathtt{w}}$ is the trainable parameter. In self-attention bidirectional LSTM-RNN, each word vector representation is also converted into a hidden representation that takes neighboring context information into consideration. The hidden representation for the n-th word in the t-th utterance is calculated as:

$$\boldsymbol{h}_n^{(t)} = \texttt{BLSTM}(\boldsymbol{w}_1^{(t)}, \cdots, \boldsymbol{w}_{N^{(t)}}^{(t)}, n; \boldsymbol{\theta}^{\mathtt{h}}), \tag{3}$$

where $\texttt{BLSTM}()$ is a function of the bidirectional LSTM-RNN layer, and $\boldsymbol{\theta}^{\mathtt{h}}$ is the trainable parameter. In addition, the hidden representations are summarized as a sentence representation using a self-attention mechanism that can consider the importance of individual hidden representations. The t-th sentence continuous representation $\boldsymbol{s}^{(t)}$ is calculated as:

$$\boldsymbol{z}_n^{(t)} = \texttt{tanh}(\boldsymbol{h}_n^{(t)}; \boldsymbol{\theta}^{\mathtt{z}}), \tag{4}$$

$$\boldsymbol{s}^{(t)} = \sum_{n=1}^{N^{(t)}} \frac{\exp({\boldsymbol{z}_n^{(t)}}^\top \bar{\boldsymbol{z}})}{\Sigma_{j=1}^{N^{(t)}} \exp({\boldsymbol{z}_j^{(t)}}^\top \bar{\boldsymbol{z}})} \boldsymbol{h}_n^{(t)}, \tag{5}$$

where $\texttt{tanh}()$ is a non-linear transformational function with tanh activation and $\boldsymbol{\theta}^{\mathtt{z}}$ is the trainable parameter. $\bar{\boldsymbol{z}}$ is the trainable context vector used to measure the importance of individual hidden representations.

Speaker role label is also converted into vector representation. This paper examines one-hot vector representation and continuous vector representation. The one-hot vector representation of the t-th speaker role is defined as:

$$\boldsymbol{r}^{(t)} = \texttt{ONEHOT}(r^{(t)}), \tag{6}$$

where $\texttt{ONEHOT}()$ is a function to covert a symbol into a one-hot vector. On the other hand, the continuous vector representation of the t-th speaker role is defined as:

$$\boldsymbol{r}^{(t)} = \texttt{EMBED}(r^{(t)}; \boldsymbol{\theta}^{\mathtt{r}}), \tag{7}$$

where $\boldsymbol{\theta}^{\mathtt{r}}$ is the trainable parameter.

In the utterance-level LSTM-RNN, interaction information from start-of-dialogue to the t-th utterance is incrementally embedded into a continuous vector representation. To this end, the t-th sentence vector representation and the t-th speaker role vector representation is merged as follows:

$$\boldsymbol{u}^{(t)} = [{\boldsymbol{s}^{(t)}}^\top, {\boldsymbol{r}^{(t)}}^\top]^\top. \tag{8}$$

The t-th continuous vector representation that embeds all dialogue context sequential information behind the t-th utterance is given as:

$$\boldsymbol{v}^{(t)} = \texttt{LSTM}(\boldsymbol{u}^{(1)}, \cdots, \boldsymbol{u}^{(t)}; \boldsymbol{\theta}^{\mathtt{v}}), \tag{9}$$

where $\texttt{LSTM}()$ is a function of the unidirectional LSTM-RNN layer, and $\boldsymbol{\theta}^{\mathtt{v}}$ represents the trainable parameter.

In the output layer, predictive probabilities of the scene labels for the t-th utterance are defined as:

$$\boldsymbol{o}^{(t)} = \texttt{SOFTMAX}(\boldsymbol{v}^{(t)}; \boldsymbol{\theta}^{\mathtt{o}}), \tag{10}$$

where $\texttt{SOFTMAX}()$ is a softmax function, and $\boldsymbol{\theta}^{\mathtt{o}}$ is a model parameter for the softmax function. $\boldsymbol{o}^{(t)}$ corresponds to $P(l^{(t)}|\boldsymbol{S}^{(1:t)}, \boldsymbol{R}^{(1:t)}, \boldsymbol{\Theta})$.

Summarizing the above, when the speaker role continuous representation is utilized, $\boldsymbol{\Theta}$ is written as $\{\boldsymbol{\theta}^{\mathtt{w}}, \boldsymbol{\theta}^{\mathtt{h}}, \boldsymbol{\theta}^{\mathtt{z}}, \bar{\boldsymbol{z}}, \boldsymbol{\theta}^{\mathtt{r}}, \boldsymbol{\theta}^{\mathtt{v}}, \boldsymbol{\theta}^{\mathtt{o}}\}$. In training, the parameter can be optimized by minimizing the cross entropy between a reference probability and the corresponding estimated probability:

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} - \sum_{d \in \mathcal{D}} \sum_{t=1}^{T_d} \sum_{l \in \mathcal{L}} \hat{\boldsymbol{o}}_{l,d}^{(t)} \log \boldsymbol{o}_{l,d}^{(t)}, \tag{11}$$

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

V. Experiment

A. Setups

The experiment used a simulated Japanese contact center dialogue data set consisting of 361 dialogues in 6 business fields. One dialogue means one telephone call between one operator

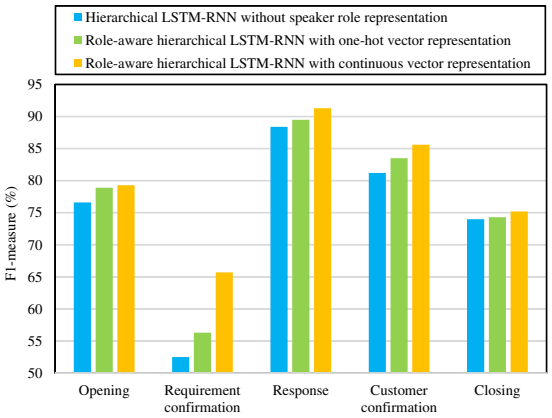

Fig. 2. F1-measure values (%) with respect to call scene.

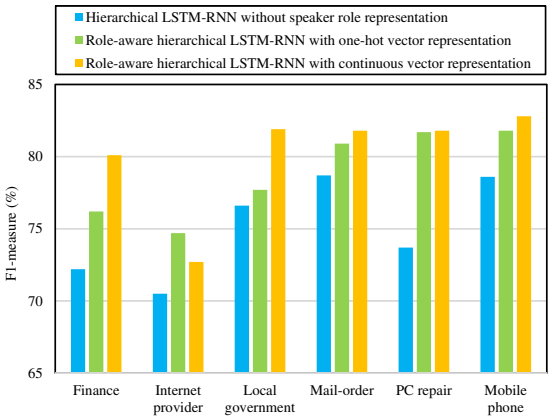

Fig. 3. F1-measure values (%) with respect to business type.

and one customer; all utterances were manually transcribed. Each dialogue was divided into speech units using LSTM-RNN based SAD [24] trained from various Japanese speech samples. We manually annotated the call scenes. Detailed setups are shown in Table 1 where #calls, #utterances, and #words represent the number of calls, words, utterances with respect to the call scenes, respectively. We can see that each business type has unique characteristics in terms of the about frequencies call scenes.

The evaluation involved 6-fold cross validation open to business type, in which 5 business types were used for training and the remaining business type was used for testing. For the evaluation, we constructed the following hierarchical LSTM-RNNs.

- Hierarchical LSTM-RNN without speaker role representation. This is our baseline method.
- Role-aware hierarchical LSTM-RNN with one-hot vector representation.
- Role-aware hierarchical LSTM-RNN with continuous vector representation.

In these models, we unified the network configurations as follows. We defined the word vector representation as a 128-dimensional vector. Words that appeared only once in the training data sets were treated as unknown words. For speaker role representation, one-hot vector representation was defined as a 2-dimensional vector, and continuous vector representation as a 32-dimensional vector. LSTM-RNN unit size was set to 400, and sentence representation was set to 400. Dropout was used for `BLSTM()` and `LSTM()`, and the dropout rate was set to 0.2. For training, the mini-batch size was set to 5 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 initial parameters and evaluated using the model that had the lowest validation loss for individual setups.

B. Results

The resulting F1-measure values are shown in Figures 2 and 3. Blue plots show hierarchical LSTM-RNN without speaker role representations, green ones show role-aware hierarchical LSTM-RNN with one-hot vector representations, and yellow show Role-aware hierarchical LSTM-RNN with continuous vector representations.

Fig. 2 shows F1-measure with respect to the call scene type. Response and customer confirmation were frequent scenes, so they were comparatively easy to discern. Opening and closing were more difficult than response or customer confirmation although it is clear that they occur at the start-of-dialogues or end-of-dialogues. The most difficult scene type was requirement confirmation. First, the proposed role-aware hierarchical LSTM-RNNs outperformed the alternatives without speaker role representations in all scenes. It indicates that interactions can be captured by introducing speaker role representations. In particular, continuous vector representation of the speaker role was effective in improving call scene segmentation. This is because the continuous vector representation can embed richer information than the one-hot vector representations. In fact, it substantially improved segmentation performance for the most difficult requirement confirmation.

Next, Fig. 3 shows F1-measure with respect to the business type. The role-aware hierarchical LSTM-RNN achieved performance superior to hierarchical LSTM-RNN without speaker role representation in all business types. This indicates that the proposed methods are robust to differences in business types as they well handle the interactions between speakers.

VI. Conclusions

This paper proposed role-aware hierarchical LSTM-RNNs for robust online call scene segmentation in contact center dialogues. We defined call scene segmentation as utterance-level sequential labeling using five scene types. The strength of the proposal is that it well utilizes long-range interaction information by simultaneously handling both sequences of sentences and sequences of speaker role labels. An experiment on simulated contact center dialogues demonstrated that the proposal could yield improved performance for all five scenes and all business types compared with a method without speaker role awareness.

REFERENCES

[1] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, pp. 33–64, 1997.

[2] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," *In Proc. International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pp. 25–32, 2006.

[3] Y. Song, L. Mou, R. Yan, L. Yi, Z. Zhu, X. Hu, and M. Zhang, "Dialogue session segmentation by embedding-enhanced texttilling," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2706–2710, 2016.

[4] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulvregt, "A hidden Markov model approach to text segmentation and event tracking," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 333–336, 1998.

[5] H. Misra, F. Hopfgartner, A. Goyal, P. Punitha, and J. M. Jose, "TV news story segmentation based on semantic coherence and content similarity," *In Proc. International Conference on Multimedia Modeling (MMM)*, pp. 347–357, 2010.

[6] M. Lu, L. Z. C. C. Leung, L. Xie, B. Ma, and H. Li, "Broadcast news segmentation using probabilistic latent semantic analysis and laplacian eigenmaps," *In Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, pp. 356–360, 2011.

[7] X. L. C. C. Leung, L. Xie, B. Ma, and H. Li, "Broadcast news story segmentation using latent topics on data mainfold," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8465–8469, 2013.

[8] J. Yu, X. Xiao, L. Xie, E. S. Chng, and H. Li, "A DNN-HMM approach to story segmentation," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1527–1531, 2016.

[9] E. Tsunoo, P. Bell, and S. Renals, "Hierarchical recurrent neural network for story segmentation," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2919–2923, 2017.

[10] E. Tsunoo, O. Klejch, P. Bell, and S. Renals, "Hierarchical recurrent neural network for story segmentation using fusion of lexical and acoustic features," *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 525–532, 2017.

[11] N. Sawada, R. Masumura, and H. Nishizaki, "Parallel hierarchical attention networks with shared memory reader for multi-stream conversational document classification," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3311–3315, 2017.

[12] B. Liu and I. Lane, "Dialogue context language modeling with recurrent neural networks," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5715–5719, 2017.

[13] R. J. Byrd, M. S. Neff, W. Teiken, Y. Park, K.-S. F. Cheng, S. C. Gates, and K. Visweswariah, "Logging of contact center telephone calls," *In Proc. International Conference on Information and Knowledge Management (CIKM)*, pp. 133–142, 2008.

[14] R. Higashinaka, Y. Minami, H. Nishikawa, K. Dohsaka, T. Meguro, S. Takahashi, and G. Kikui, "Learning to model domain-specific utterance sequences for extractive summarization of contact center dialogues," *In Proc. International Conference on Computational Linguistics (COLING)*, pp. 400–408, 2010.

[15] A. Tamura, K. Ishikawa, M. Saikou, and M. Tsuchida, "Extractive summarization method for contact center dialogues based on call logs," *In Proc. International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 500–508, 2011.

[16] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," *In Proc. Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 51–58, 2006.

[17] C. Chastagnol and L. Devillers, "Analysis of anger across several agent-customer interactions in french call centers," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4963, 2011.

[18] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Hierarchical LSTMs with joint learning for estimating customer satisfaction from contact center callis," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1716–1720, 2017.

[19] E. Shirberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1, pp. 127–154, 2000.

[20] Y. Liu, A. Stolcke, E. Shirberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech," *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 451–458, 2005.

[21] C. Xu, L. Xie, G. Huang, X. Xiao, E. S. Chng, and H. Li, "A deep neural network approach for sentence boundary detection in broadcast news," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2887–2891, 2014.

[22] R. Masumura, T. Asami, H. Masataki, R. Ishii, and R. Higashinaka, "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)*, pp. 1661–1665, 2017.

[23] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1480–1489, 2016.

[24] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "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)*, pp. 483–487, 2013.