INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Parallel Hierarchical Attention Networks with Shared Memory Reader for Multi-Stream Conversational Document Classification

Naoki Sawada[1,2], *Ryo Masumura*[1], *Hiromitsu Nishizaki*[3]

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2]The Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Japan
[3]Graduate School of Interdisciplinary Research, Faculty of Engineering, University of Yamanashi, Japan

sawada@alps-lab.org, masumura.ryo@lab.ntt.co.jp, hnishi@yamanashi.ac.jp

Abstract

This paper describes a novel classification method for multi-stream conversational documents. Documents of contact center dialogues or meetings are often composed of multiple source documents that are transcriptions of the recordings of each speaker's channel. To enhance the classification performance of such multi-stream conversational documents, three main advances over the previous method are introduced. The first is a parallel hierarchical attention network (PHAN) for multi-stream conversational document modeling. PHAN can precisely capture word and sentence structures of individual source documents and efficiently integrate them. The second is a shared memory reader that can yield a shared attention mechanism. The shared memory reader highlights common important information in a conversation. Our experiments on a call category classification in contact center dialogues show that PHAN together with the shared memory reader outperforms the single document modeling method and previous multi-stream document modeling method.

Index Terms: Multi-stream conversational documents, hierarchical attention networks, memory reader, call category classification

1. Introduction

Conversational documents that transcribe conversational speech such as contact center dialogue or speech during meetings have been attracting much attention [1, 2, 3]. The conversational documents are often composed of multiple source documents that are transcriptions recorded by each speaker's channel. This paper aims to enhance the performance of multi-stream conversational documents classification such as theme-identification tasks [4, 5, 6].

For single stream classification tasks, i.e., sentence or document classification, modern technologies are deep learning. Several deep learning technologies such as convolution neural networks or recurrent neural networks including gated recurrent units (GRUs) and long short-term memories (LSTMs) were applied for modeling the sentence classification, and they displayed superior performance to conventional discriminative modeling [7, 8, 9, 10]. It is reported that the networks can precisely capture sequential semantics by combining attention mechanism that can focus on a key part of sequence [11, 12]. In addition, hierarchical networks that can take into account not only word structure but also sentence structure were examined [13, 14]. Furthermore, hierarchical attention networks (HANs) that support both the hierarchical networks and the attention mechanism were recently proposed [15].

There are few studies for multi-stream classification [16, 17]. A representative method is parallel LSTM that integrates outputs of multiple LSTMs [16]. This modeling can simultaneously manage multiple streams; however, there are two limitations for classifying multi-stream conversational documents. First, simple LSTM was introduced for modeling individual streams although conversational documents include multiple utterances. It can be thought that attention mechanism or hierarchical networks are more suitable for the conversational documents because each source document involves a lot of utterances in conversations. In addition, the second limitation is that each LSTM independently manages individual streams, although speakers talk about a common theme in conversations. It is likely that theme sharing mechanism between speaker-dependent networks yields further performance improvements.

In this paper, we introduce two main advances over the previous method. First, this paper proposes parallel hierarchical attention networks (PHANs) that use multiple HANs to model individual source documents. PHANs can precisely capture word and sentence structures of source documents and efficiently integrate them. Second, this paper proposes a shared memory reader that can yield a shared attention mechanism. The shared memory reader highlights common important information in a conversation. This idea is inspired by dialogue topic modeling in which common important information is shared among each speaker [18]. Moreover, we introduce an additional mechanism that repeatedly updates the shared memory reader. The mechanism can reflect the entire information of a target conversation to the shared attention mechanism. This idea is inspired by end-to-end memory networks where multiple computational steps called multi-hops were performed [12].

In our experiments for a call theme classification task, we present the effectiveness of PHAN and the shared memory reader. In addition, we verify that PHAN with the shared memory reader outperforms a method that ignores the difference between source documents and a method that ignores common information between source documents.

2. Multi-Stream Conversational Document Classification

Conversational documents are composed by multiple speaker's utterances. Multi-stream conversational documents, one form of conversational documents, are constituted by multiple source documents that are recorded by each speaker's channel. For ex-

http://dx.doi.org/10.21437/Interspeech.2017-269

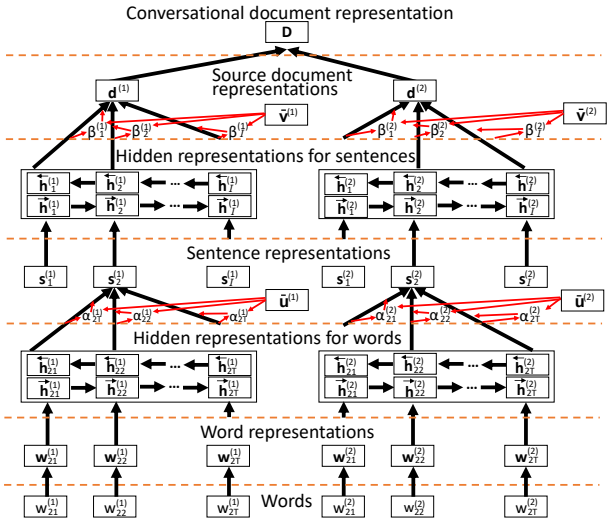

Figure 1: *PHAN using two source documents.*

ample, multi-stream conversational documents in contact center dialogues are constituted by two source documents: an operator's document and an customer's document. In the multi-stream conversational document classification, a label of a multi-stream conversational document D is determined as:

$$\hat{l} = \underset{l}{\operatorname{argmax}}\, P(l|D, \boldsymbol{\Theta}), \quad (1)$$

$$D = \{d^{(1)}, \cdots, d^{(M)}\}, \quad (2)$$

where $\boldsymbol{\Theta}$ represents a model parameter. $\hat{l}$ is an estimated label. M is a number of source documents, and $d^{(m)}$ is the m-th source document in a conversational document. Each source document includes multiple sentences that correspond to utterances in a conversation. $d^{(m)}$ is represented as:

$$d^{(m)} = \{s_1^{(m)}, \cdots, s_{I_m}^{(m)}\}, \quad (3)$$

where $s_i^{(m)}$ means the i-th sentence in the m-th source document. I_m is number of sentences in the m-th source document. In addition, each sentence includes multiple words. $s_i^{(m)}$ is represented as:

$$s_i^{(m)} = \{w_{i1}^{(m)}, \cdots, w_{iT_i}^{(m)}\}, \quad (4)$$

where $w_{it}^{(m)}$ denotes the t-th word in the i-th sentence, for the m-th source document. T_i is number of words in the i-th sentence, for the m-th source document.

3. Proposed Method

3.1. Parallel Hierarchical Attention Networks

This paper proposes parallel hierarchical attention networks (PHANs) that introduce multiple HANs for individual streams [15]. Figure 1 shows the detailed structure of PHAN using two source documents. In PHAN, various continuous representations, i.e., word representations, sentence representations, source document representations, and a conversational document representation are hierarchically composed. The conversational document representation is directly used for classification. In addition, an attention mechanism is introduced when summarizing word and sentence information. To this end, a word memory reader and a sentence memory reader are introduced for each source document.

3.1.1. Definitions

In PHAN, each word in individual sentences is first converted into a continuous representation [19]. A word representation of t-th word in the i-th sentence is defined as:

$$\boldsymbol{w}_{it}^{(m)} = \texttt{EMBEDDING}(w_{it}^{(m)}; \boldsymbol{\theta}_{\texttt{e}}^{(m)}), \quad (5)$$

where $\texttt{EMBEDDING}()$ is a linear transformational function to embed a word to a continuous vector, and $\boldsymbol{\theta}_{\texttt{e}}^{(m)}$ is a model parameter of the function for the m-th source document. $\boldsymbol{w}_{it}^{(m)}$ is a word representation of $w_{it}^{(m)}$.

Next, each word representation is converted into a hidden representation that summarizes the neighboring words in a sentence using stream-dependent word encoders. In order to summarize the information from both directions for words, this paper uses bidirectional GRU as the word encoders [20]. The hidden representation for the t-th word in the i-th sentence, in the m-th source document, is calculated as:

$$\overrightarrow{\boldsymbol{h}}_{it}^{(m)} = \overrightarrow{\texttt{GRU}}(\boldsymbol{w}_{it}^{(m)}; \boldsymbol{\theta}_{\texttt{rw}}^{(m)}), \quad (6)$$

$$\overleftarrow{\boldsymbol{h}}_{it}^{(m)} = \overleftarrow{\texttt{GRU}}(\boldsymbol{w}_{it}^{(m)}; \boldsymbol{\theta}_{\texttt{lw}}^{(m)}), \quad (7)$$

$$\boldsymbol{h}_{it}^{(m)} = [\overrightarrow{\boldsymbol{h}}_{it}^{(m)\top}, \overleftarrow{\boldsymbol{h}}_{it}^{(m)\top}]^\top, \quad (8)$$

where $\overrightarrow{\texttt{GRU}}()$ and $\overleftarrow{\texttt{GRU}}()$ are a forward GRU function and a backward GRU function, respectively. $\boldsymbol{\theta}_{\texttt{rw}}^{(m)}$ and $\boldsymbol{\theta}_{\texttt{lw}}^{(m)}$ are model parameters for word-level GRUs. $\boldsymbol{h}_{it}^{(m)}$ is a concatenated hidden representation of $\overrightarrow{\boldsymbol{h}}_{it}^{(m)}$ and $\overleftarrow{\boldsymbol{h}}_{it}^{(m)}$.

In addition, sentence representations are composed by summarizing hidden representations of the word encoder. To this end, word attention mechanism is introduced. The i-th sentence representation in the m-th source document is calculated as:

$$\boldsymbol{u}_{it}^{(m)} = \texttt{tanh}(\boldsymbol{h}_{it}^{(m)}; \boldsymbol{\theta}_{\texttt{w}}^{(m)}), \quad (9)$$

$$\alpha_{it}^{(m)} = \frac{\exp(\boldsymbol{u}_{it}^{(m)\top} \bar{\boldsymbol{u}}^{(m)})}{\Sigma_{n=1}^{T_i} \exp(\boldsymbol{u}_{in}^{(m)\top} \bar{\boldsymbol{u}}^{(m)})}, \quad (10)$$

$$\boldsymbol{s}_i^{(m)} = \sum_{t=1}^{T_i} \alpha_{it}^{(m)} \boldsymbol{h}_{it}^{(m)}, \quad (11)$$

where $\texttt{tanh}()$ is a non-linear transformational function with $\tanh$ activation, and $\boldsymbol{\theta}_{\texttt{w}}^{(m)}$ is a model parameter of the non-linear transformational function for the m-th source document. $\alpha_{it}^{(m)}$ means a normalized importance weight for the t-th word in the i-th sentence for the m-th source document. $\bar{\boldsymbol{u}}^{(m)}$ is a word memory reader for the m-th source document, which is used for the word attention mechanism. The word memory reader is jointly optimized during the training process.

Then, each sentence representation is additionally converted into a hidden representation that summarizes neighbor sentences using stream-dependent sentence encoders. The sentence encoders are also composed by bidirectional GRUs. The hidden representation for i-th sentence in m-th source document is calculated as:

$$\overrightarrow{\boldsymbol{h}}_i^{(m)} = \overrightarrow{\texttt{GRU}}(\boldsymbol{s}_i^{(m)}; \boldsymbol{\theta}_{\texttt{rs}}^{(m)}), \quad (12)$$

$$\overleftarrow{\boldsymbol{h}}_i^{(m)} = \overleftarrow{\texttt{GRU}}(\boldsymbol{s}_i^{(m)}; \boldsymbol{\theta}_{\texttt{ls}}^{(m)}), \quad (13)$$

$$\boldsymbol{h}_i^{(m)} = [\overrightarrow{\boldsymbol{h}}_i^{(m)\top}, \overleftarrow{\boldsymbol{h}}_i^{(m)\top}]^\top, \quad (14)$$

where $\boldsymbol{\theta}_{\mathtt{rs}}^{(m)}$ and $\boldsymbol{\theta}_{\mathtt{ls}}^{(m)}$ are model parameters for sentence-level GRUs. $\boldsymbol{h}_i^{(m)}$ is a concatenated continuous representation of $\overrightarrow{\boldsymbol{h}}_i^{(m)}$ and $\overleftarrow{\boldsymbol{h}}_i^{(m)}$.

Source document representations are composed by summarizing hidden representations of the sentence encoder with a sentence attention mechanism. Moreover, a conversational document representation is composed by adding the source document representations. The source document representations and the conversational document representation are calculated as

$$\boldsymbol{v}_i^{(m)} = \mathtt{tanh}(\boldsymbol{h}_i^{(m)}; \boldsymbol{\theta}_{\mathtt{s}}^{(m)}), \quad (15)$$

$$\beta_i^{(m)} = \frac{\exp(\boldsymbol{v}_i^{(m)\top} \bar{\boldsymbol{v}}^{(m)})}{\Sigma_{j=1}^{I_m} \exp(\boldsymbol{v}_j^{(m)\top} \bar{\boldsymbol{v}}^{(m)})}, \quad (16)$$

$$\boldsymbol{d}^{(m)} = \sum_{i=1}^{I_m} \beta_i^{(m)} \boldsymbol{h}_i^{(m)}, \quad (17)$$

$$\boldsymbol{D} = \sum_{m=1}^{M} \boldsymbol{d}^{(m)}, \quad (18)$$

where $\beta_i^{(m)}$ means a normalized importance weight for the i-th sentence in the m-th stream. $\boldsymbol{\theta}_{\mathtt{s}}^{(m)}$ is a model parameter in the non-linear transformational function for the m-th source document. $\bar{\boldsymbol{v}}^{(m)}$ is a sentence memory reader for the m-th stream, which is jointly optimized during the training process. $\boldsymbol{d}^{(m)}$ and $\boldsymbol{D}$ are m-th source document representation and conversational document representation, respectively.

In an output layer of PHAN, predicted probabilities are produced using the conversational document representation:

$$\boldsymbol{O} = \mathtt{SOFTMAX}(\boldsymbol{D}; \boldsymbol{\theta}_{\mathtt{o}}), \quad (19)$$

where $\mathtt{SOFTMAX}()$ is a softmax function, and $\boldsymbol{\theta}_{\mathtt{o}}$ is its parameter. The l-th dimension in an output $\boldsymbol{O}$ corresponds to $P(l|D, \boldsymbol{\Theta})$.

3.1.2. Optimization

In PHAN, trainable parameter $\boldsymbol{\Theta}$ is represented as:

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{\mathtt{e}}^{(m)}, \boldsymbol{\theta}_{\mathtt{rw}}^{(m)}, \boldsymbol{\theta}_{\mathtt{lw}}^{(m)}, \boldsymbol{\theta}_{\mathtt{w}}^{(m)}, \boldsymbol{\theta}_{\mathtt{rs}}^{(m)}, \boldsymbol{\theta}_{\mathtt{ls}}^{(m)}, \boldsymbol{\theta}_{\mathtt{s}}^{(m)}, \boldsymbol{\theta}_{\mathtt{o}}, \bar{\boldsymbol{u}}^{(m)}, \bar{\boldsymbol{v}}^{(m)}\}, \quad (20)$$

where $m \in [1, \cdots, M]$. The parameters can be optimized by minimizing cross entropy between a reference probability and an estimated probability:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} - \sum_{D \in \mathcal{D}} \sum_{l} \hat{\boldsymbol{O}}_l^D \log \boldsymbol{O}_l^D, \quad (21)$$

where $\hat{\boldsymbol{O}}_l^D$ and $\boldsymbol{O}_l^D$ are a reference probability and an estimated probability of label l for a conversational document D, respectively. $\mathcal{D}$ denotes a training data set.

3.2. Shared Memory Readers

This paper proposes shared memory readers in order to highlight common important information through a conversation. The shared memory readers can yield a shared attention mechanism. In fact, similar idea was introduced in machine translation area where attention mechanism was shared between languages [21]. In standard PHANs, word memory readers and sentence memory readers are provided for each source document. On the other hand, a single word memory reader and a single sentence memory reader are shared between all source documents. Thus, $\alpha_{it}^{(m)}$ and $\beta_i^{(m)}$ are defined as:

$$\alpha_{it}^{(m)} = \frac{\exp(\boldsymbol{u}_{it}^{(m)\top} \bar{\boldsymbol{u}})}{\Sigma_{n=1}^{T_i} \exp(\boldsymbol{u}_{in}^{(m)\top} \bar{\boldsymbol{u}})}, \quad (22)$$

$$\beta_i^{(m)} = \frac{\exp(\boldsymbol{v}_i^{(m)\top} \bar{\boldsymbol{v}})}{\Sigma_{j=1}^{I_m} \exp(\boldsymbol{v}_j^{(m)\top} \bar{\boldsymbol{v}})}, \quad (23)$$

where $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{v}}$ are the shared word memory reader and the shared sentence memory reader, respectively. The shared memory readers are jointly optimized with other trainable parameters.

Table 1: *The number of dialogs of experimental data sets for call theme classification .*

Label	Training	Validation	Test
New contract	54	54	108
Downgrading	30	30	60
Upgrading	29	28	56
Add option	13	12	25
Delete option	13	12	25
Request for ID/PW	14	12	25
Change of name	25	25	50
Cancellation	26	25	50
Total	204	198	399

3.3. Multi-Hops

This paper introduces additional mechanism called multi-hops that repeatedly updates the shared sentence memory reader. This mechanism is inspired by end-to-end memory networks [12]. In PHANs, a conversational document representation is repeatedly used for updating the shared sentence memory reader. Update rules are defined as:

$$\bar{\boldsymbol{v}}_{k-1} = \boldsymbol{D}_{k-1}, \quad (24)$$

$$\beta_{i,k}^{(m)} = \frac{\exp(\boldsymbol{v}_i^{(m)\top} \bar{\boldsymbol{v}}_{k-1})}{\sum_{j=1}^{I_m} \exp(\boldsymbol{v}_j^{(m)\top} \bar{\boldsymbol{v}}_{k-1})}, \quad (25)$$

$$\boldsymbol{d}_k^{(m)} = \sum_{i=1}^{I_m} \beta_{i,k}^{(m)} \boldsymbol{h}_i^{(m)}, \quad (26)$$

$$\boldsymbol{D}_k = \sum_{m=1}^{M} \boldsymbol{d}_k^{(m)} + \mathtt{Linear}(\boldsymbol{D}_{k-1}; \boldsymbol{\theta}_{\mathtt{h}}^{(m)}), \quad (27)$$

where $\boldsymbol{D}_0$ corresponds to $\boldsymbol{D}$ in Eq. (18). $\bar{\boldsymbol{v}}_{k-1}$ is reconstructed shared sentence memory reader in the k-th hop. After one hopping, a conversational document representation is reconstructed using both source document representations in the current hop and previous conversational document representation. $\mathtt{Linear}()$ is a linear transformational function, and $\boldsymbol{\theta}_{\mathtt{h}}^{(m)}$ is a parameter that is jointly optimized with other trainable parameters. After K-hopping, updated conversational document representation $\boldsymbol{D}_K$ is used for classification in Eq. (19). By introducing multi-hops, we can reflect the entire information of a target conversation to the shared attention mechanism.

4. Experiments

4.1. Setups

Our evaluation task is call theme classification. We employed the Japanese contact center dialogue data sets, which include

Table 2: *Call theme classification accuracy [%]*

		Multi-stream	HAN	Shared memory reader	Multi-hop	Validation	Test
Baseline	BGRU					88.8	83.7
	PBGRU	√				85.3	80.2
Proposed	PHAN	√	√			87.5	86.6
	PHAN-SMR	√	√	√		90.7	87.3
	PHAN-SMR-MH	√	√	√	√	**91.4**	**87.9**

several call themes. One dialogue set means one telephone call between one operator and one customer. Each dialogue was separately recorded. Table 1 details themes about the training, validation and test sets. The data sets includes eight themes. This paper used manual transcriptions of the contact center dialogue and divided them into utterances by deep neural network based speech activity detector [22] trained from the Corpus of Spontaneous Japanese [23]. Each dialogue included about 148 utterances per each speaker. Each utterance involved from 1 to 442 words.

For evaluation, we employed five methods.

- **BGRU**: Single document classification method based on bidirectional GRU. This method regarded multi-stream document as a single document. Specifically, each word was converted into word representations using a single linear function. Next, each word representation was fed into both a single forward GRU function and a single backward GRU function. A document representation was obtained by averaging word-level hidden representations in the bidirectional GRU. In an output layer, predicted probabilities were produced using the document representation.
- **PBGRU**: Multi-stream document classification method based on parallel bidirectional GRU. This method introduced different parameters with respect to each stream. Specifically, each word in each source document was converted into word representations using stream-dependent linear functions. Next, each word representation in each source document was fed into both a stream-dependent forward GRU function and a stream-dependent backward GRU function. Source document representations were obtained by averaging word-level hidden representations in the stream-dependent bidirectional GRUs. A multi-stream document representation was obtained by averaging the source document representations. In an output layer, predicted probabilities were produced using the multi-stream document representation.
- **PHAN**: Multi-stream document classification method based on PHAN presented in Section 3.1.
- **PHAN-SMR**: Multi-stream document classification method based on PHAN with shared memory reader presented in Sections 3.1-3.2.
- **PHAN-SMR-MH**: Multi-stream document classification method based on PHAN with multi-hopped shared memory reader presented in Sections 3.1-3.3. A number of hops K was set to 3.

BGRU and **PBGRU** are positioned as baseline methods which are similar setup to previous work [16]. Several parameters between each method were unified. 32-dimensional word representations, 64-dimensional sentence representations and 64-dimensional document representation were used. For training, a mini-batch size was set to 5. Adam was used for the optimizer. The training epoch was stopped when the validation loss was not improved six consecutive times. For each method, we constructed five models by varying an initial parameter randomly selected for individual conditions and evaluated averaged performance.

4.2. Results

Table 2 shows the experimental results, in terms of call theme classification accuracy for both validation and test sets. Table 2 also denotes whether each mechanism can be taken into account or not.

First, among baseline methods, PBGRU was inferior to BGRU. This is because training data was insufficient for PBGRU. In fact, PBGRU had about two times more parameters than BGRU. It can be thought that PBGRU will outperform BGRU if much more training data can be obtained.

Next, among multi-stream document classification methods, PHAN presented superior performance to PBGRU. This indicates that both a hierarchical network and an attention mechanism were well performed in conversational document classification tasks. PHAN could outperform BGRU in the test set although PHAN quite increased a model parameter size. In addition, PHAN-SMR outperformed PHAN. This is because PHAN with shared memory reader could capture important common information from source documents while keeping precise information about individual source documents. PHAN-SMR could surpass simple BGRU in both the validation set and the test set. This result implies that the shared memory reader is also useful to reduce model parameter size and efficiently train the parameters using limited training data. Moreover, the highest performance was attained by PHAN-SMR-MH. This results shows multi-hops could enhance shared attention mechanism by reflecting entire information of a target conversation.

Summarizing the above, in terms of classification accuracy, the proposed method could yield 2.6 point performance improvement in validation set and 4.2 point performance improvement in test set compared to baseline methods.

5. Conclusion

This paper proposed a PHAN with a shared memory reader for multi-stream conversational document classification. PHAN can hierarchically summarize information in a conversational document and precisely capture important information using a attention mechanism. The shared memory reader can yield a shared attention mechanism that highlights common important information between the speakers. We also proposed multi-hops that repeatedly updates the shared memory reader. Our experiments showed that PHAN together with the shared memory reader outperformed single document modeling method and previous multi-stream document modeling method. In future work, we will examine an evaluation using automatic speech recognition transcriptions. Moreover, we will introduce an additional mechanism that takes a detailed conversational structure into consideration.

6. References

[1] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," *In Proc. Human Language Technology Conference (HLT)*, pp. 1–7, 2001.

[2] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the AMI and AMIDA project," *In Proc. biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 238–247, 2007.

[3] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 499–513, 2012.

[4] M. Morchid, G. Linarès, M. El Bèze, and R. De Mori, "Theme identification in telephone service conversations using quaternions of speech features," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1394–1398, 2013.

[5] M. Morchid, R. Dufour, M. Bouallegue, G. Linarès, and R. De Mori, "Theme identification in human-human conversations with features from specific speaker type hidden spaces," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 248–252, 2014.

[6] Y. Estève, M. Bouallegue, C. Lailler, M. Morchid, R. Dufour, G. Linarès, D. Matrouf, and R. D. Mori, "Integration of word and semantic features for theme identification in telephone conversations," *Natural Language Dialog Systems and Intelligent Assistants*, pp. 223–231, 2015.

[7] Y. Kim, "Convolutional neural networks for sentence classification," *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

[8] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *In Proc. International Conference on Neural Information Processing Systems (NIPS)*, pp. 649–657, 2015.

[9] S. Ravuri and A. Stolcke, "Recurrent neural network and LSTM models for lexical utterance classification," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 135–139, 2015.

[10] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2267–2273, 2015.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] S. Sukhbaatar, a. szlam, J. Weston, and R. Fergus, "End-to-end memory networks," *In Proc. International Conference on Neural Information Processing Systems (NIPS)*, pp. 2440–2448, 2015.

[13] M. Denil, A. Demiraj, N. Kalchbrenner, P. Blunsom, and N. de Freitas, "Modelling, visualising and summarising documents with a single convolutional neural network," *arXiv prerint arXiv:1406.3830*, 2014.

[14] P. Bhatia, Y. Ji, and J. Eisenstein, "Better document-level sentiment analysis from rst discourse parsing," *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2212–2218, 2015.

[15] 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.

[16] M. Bouaziz, M. Morchid, R. Dufour, G. Linars, and R. D. Mori, "Parallel long short-term memory for multi-stream classification," *In Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 218–223, 2016.

[17] M. Bouaziz, M. Morchid, R. Dufour, and G. Linars, "Improving multi-stream classification by mapping sequence-embedding in a high dimensional space," *In Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 224–231, 2016.

[18] R. Masumura, T. Oba, H. Masataki, O. Yoshioka, and S. Takahashi, "Role play dialogue topic model for language model adaptation in multi-party conversation speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4873–4877, 2014.

[19] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[20] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

[21] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 866–875, 2016.

[22] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 697–710, 2013.

[23] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," *In Proc. International Conference on Language Resources and Evaluation (LREC)*, pp. 947–952, 2000.