Vol.2017-NL-231 No.1
Vol.2017-SLP-116 No.1
2017/5/15

マルチストリーム音声ドキュメントのための Parallel Hierarchical Attention Networkの検討

澤田 直輝[1,2,a)]　増村 亮[2,b)]　西崎 博光[3,c)]

概要：コンタクトセンタやミーティングでは話者ごとに音声を分割して保存されるため，マルチストリームな音声ドキュメントとして音声データが蓄積されている．本研究ではマルチストリーム音声ドキュメントのための新しいテーマ分類手法を提案する．提案手法の特徴は 2 つある．まず 1 つ目は，Hierarchical Attention Networks を拡張した Parallel Hierarchical Attention Networks（並列階層型注意ネットワーク）(PHAN) を提案し，これを用いることである．それぞれのドキュメントの単語構造と文構造を精緻に捉えることができる．2 つ目は，PHAN の注意機構をドキュメント間で共有化する共有メモリリーダを導入することである．この共有メモリリーダは話者間の重要な共通情報を際立たせることができる．マルチストリーム音声ドキュメントとしてコンタクトセンタのテーマ分類タスクの評価実験を行った結果，従来法よりも提案手法のほうが高い分類性能となった．

Parallel hierarchical attention networks for multi-stream conversation documents

Naoki SAWADA[1,2,a)]　Ryo Masumura[2,b)]　Hiromitsu NISHIZAKI[3,c)]

1. はじめに

コンタクトセンタやミーティングの会話から作成される対話音声ドキュメントのテーマ分類が注目されている [1], [2], [3]．コンタクトセンタやミーティングの会話では複数話者が存在し，話者ごとに音声を保存するため，話者ごとの複数の音声ドキュメントが生成される．これを本稿ではマルチストリーム音声ドキュメントと呼んでいる．

[1] 山梨大学大学院医工農学総合教育部
Department of Education, Interdisciplinary Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi Takeda 4–3–11, Kofu-shi, Yamanashi, 400–8511 Japan
[2] 日本電信電話株式会社 NTT メディアインテリジェンス研究所
(NTT Corporation NTT Media Intelligence Laboratories)
1-1, Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847, Japan
[3] 山梨大学大学院総合研究部工学域
The Graduate School of Interdisciplinary Research, Faculty of Engineering, University of Yamanashi Takeda 4–3–11, Kofu-shi, Yamanashi, 400–8511 Japan
[a)] sawada@alps-lab.org
[b)] masumura.ryo@lab.ntt.co.jp
[c)] hnishi@yamanashi.ac.jp

本研究では，コンタクトセンタ等で収録されるマルチストリーム音声ドキュメントのテーマ分類の性能向上を目的とする [4], [5], [6].

シングルストリームの文（発話）や文書に対するクラス分類では，最新技術として深層学習が用いられている [7], [8], [9], [10]．畳み込みニューラルネットワークやリカレントニューラルネットワークなどのいくつかの深層学習技術が文分類技術に適用されており，従来のモデリング手法よりも高い性能であることが示されている．

さらに，重要な箇所を際立たせる注意機構（attention mechanism）をネットワークに組み込むことで，正確に意味的な特徴をつかむことができる手法が提案されている [11], [12]．これに加えて，単語特徴と文特徴を別々に特徴抽出する階層構造も有効であることが示されている [13], [14]．また，階層構造と注意機構の両方に対応した Hierarchical Attention Networks（HANs）が提案されている [15].

本研究が対象とするマルチストリーム音声ドキュメント

のクラス分類の研究例もいくつか存在している [16], [17]. 例えば，複数の LSTM を用いた LSTM を並列化する手法が提案されている [16]．この手法では，話者ごとに独立した LSTM を用いることで，マルチストリーム音声ドキュメントを同時に扱うことができる．しかし，この手法でマルチストリーム音声ドキュメントをクラス分類するためには 2 つの課題が存在する．まず 1 つ目は，単純な LSTM では複数の発話を含むマルチストリーム音声ドキュメントに対してうまく特徴を捉えることが難しいことである．この課題に対しては，階層構造や注意機構を導入することで複数の発話を含むマルチストリーム音声ドキュメントの単語特徴と文特徴を捉えることができると考えられる．2 つ目の課題は，それぞれの LSTM が各ストリームを独立に扱っていることである．この方法では，話者ごとの音声ドキュメントを独立に扱ってしまい，それぞれの話者が話している共通の話題を正確に扱うことができない．これに対して，話者間の共通情報を際立たせる共有した注意機構を導入することで性能の改善が期待できる．

そこで本稿では，これらの課題を解決するための 2 つの特徴を持ったテーマ分類手法を提案する．まず 1 つ目は，HANs をマルチストリーム音声ドキュメントに対応させた Parallel Hirarchical Attention Networks (PHAN) を提案し，これを用いることである．PHAN はマルチストリーム音声ドキュメントに対して単語構造と文構造の特徴を正確に捉えることができる．2 つ目は，共有注意機構である共有メモリリーダを PHAN に導入することである．これは，対話音声は共通の重要情報を持っているため [18]，各話者の重要情報を共有することで共通の重要情報を際立たせることができると考えた方法である．さらに，共有メモリリーダを繰り返し更新するマルチホップを提案する．この方法は，他の話者の重要情報を考慮した情報の抽出が可能となる．これは，end-to-end memory networks の multi-hops と呼ばれる方法を参考としている [12]．コンタクトセンタのテーマ分類タスクにおける評価実験の結果，PHAN と共有メモリリーダによって分類性能が改善された．さらに，共有メモリリーダを繰り返し更新するマルチホップによりさらなる性能の改善が示された．

2. マルチストリーム音声ドキュメントのクラス分類

マルチストリーム音声ドキュメントは複数話者が存在する．この話者ごとの発話を録音し，複数の音声ドキュメントを生成する．例えば，コンタクトセンタのマルチストリーム音声ドキュメントでは，オペレータと顧客の 2 つドキュメントが存在する．マルチストリーム音声ドキュメントのクラス分類において，M 個のストリームからなるドキュメント集合 D のクラスラベルは以下の式で表される．

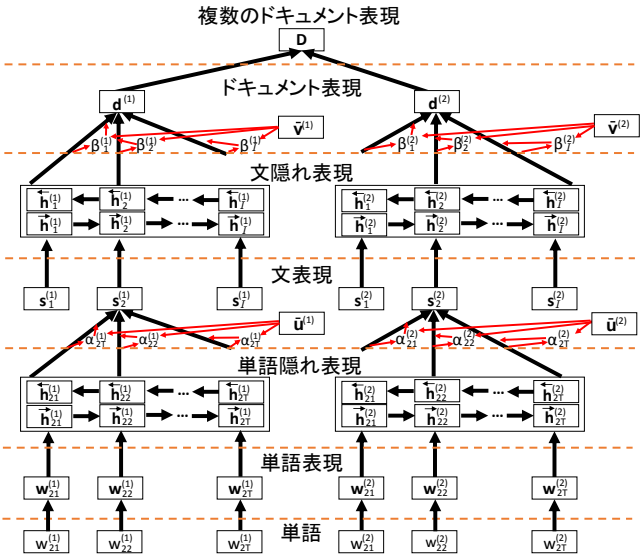

図 1 2 個のストリームを持つマルチストリームドキュメントにおける PHAN 構造

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

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

ここで，$\boldsymbol{\Theta}$ はモデルパラメータである．また $\hat{l}$ は推定ラベルである．$d^{(m)}$ は m 番目のドキュメントを表す．それぞれのドキュメントは発話に対応する複数の文を持っている．$d^{(m)}$ は，以下の式で表される．

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

ここで，$s_i^{(m)}$ は m 番目のドキュメントの i 番目の文を表している．また，I_m は m 番目のドキュメントの文の数を表す．さらに，それぞれの文は複数の単語を持っている．$s_i^{(m)}$ は以下の式で表される．

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

ここで，$w_{it}^{(m)}$ は m 番目のドキュメントの i 番目の文の t 番目の単語を表している．また，T_i は m 番目のドキュメントの i 番目の文の単語数を表す．

3. 提案手法

3.1 Parallel Hierarchical Attention Network

本稿では，各ストリームごとに HAN を導入した PHAN を提案する．図 1 に，ドキュメントが 2 つの場合の概要図を示す．PHAN は，単語表現，文表現，ドキュメント表現，複数のドキュメント表現という 4 つの連続表現を階層的に持っている．複数のドキュメント表現がクラス分類に使用される．加えて，単語表現から文表現，文表現からドキュメント表現に変換する際に注意機構を利用する．この注意機構を実現するため，単語メモリリーダと文メモリリーダが存在する．各メモリリーダはストリームごとに独立して用いられる．

3.1.1 PHAN の定義

PHAN では，最初に各文の全ての単語を連続表現に変換する [19]．そのため i 番目の文の t 番目の単語表現は以下の式で示される．

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

ここで，EMBEDDING() は線形変換関数であり単語をベクトル表現に変換する．そして，$\boldsymbol{\theta}_{\texttt{e}}^{(m)}$ は m 番目のドキュメントにおける線形変換関数のモデルパラメータである．さらに $\boldsymbol{w}_{it}^{(m)}$ は $w_{it}^{(m)}$ の単語ベクトル表現である．

次に，各単語ベクトル表現から単語エンコーダを用いて前後の単語を考慮した単語隠れ表現に変換する．前後の単語情報を考慮するために，本稿では双方向 GRU を単語エンコーダとして使用した [20]．m 番目のドキュメントにおける i 番目の文の t 番目の単語隠れ表現は以下の式で求められる．

$$\overrightarrow{\boldsymbol{h}}_{it}^{(m)} = \overrightarrow{\texttt{GRU}}(\boldsymbol{w}_{it}^{(m)}; \boldsymbol{\theta}_{\texttt{rw}}^{(m)}), \tag{6}$$
$$\overleftarrow{\boldsymbol{h}}_{it}^{(m)} = \overleftarrow{\texttt{GRU}}(\boldsymbol{w}_{it}^{(m)}; \boldsymbol{\theta}_{\texttt{lw}}^{(m)}), \tag{7}$$
$$\boldsymbol{h}_{it}^{(m)} = [\overrightarrow{\boldsymbol{h}}_{it}^{(m)\top}, \overleftarrow{\boldsymbol{h}}_{it}^{(m)\top}]^{\top}, \tag{8}$$

ここで，$\overrightarrow{\texttt{GRU}}()$ と $\overleftarrow{\texttt{GRU}}()$ は，前方向の GRU 関数と後ろ方向の GRU 関数である．また，$\boldsymbol{\theta}_{\texttt{rw}}^{(m)}$ と $\boldsymbol{\theta}_{\texttt{lw}}^{(m)}$ は単語レベルの GRU のモデルパラメータである．さらに，$\boldsymbol{h}_{it}^{(m)}$ は $\overrightarrow{\boldsymbol{h}}_{it}^{(m)}$ と $\overleftarrow{\boldsymbol{h}}_{it}^{(m)}$ を結合した結果である．

次に，単語隠れ表現から文表現を求める．ここに，単語注意機構を導入する．m 番目のドキュメントにおける i 番目の文表現は以下の式で表される．

$$\boldsymbol{u}_{it}^{(m)} = \texttt{tanh}(\boldsymbol{h}_{it}^{(m)}; \boldsymbol{\theta}_{\texttt{w}}^{(m)}), \tag{9}$$
$$\alpha_{it}^{(m)} = \frac{\exp(\boldsymbol{u}_{it}^{(m)\top} \bar{\boldsymbol{u}}^{(m)})}{\Sigma_{n=1}^{T_i} \exp(\boldsymbol{u}_{in}^{(m)\top} \bar{\boldsymbol{u}}^{(m)})}, \tag{10}$$
$$\boldsymbol{s}_i^{(m)} = \sum_{t=1}^{T_i} \alpha_{it}^{(m)} \boldsymbol{h}_{it}^{(m)}, \tag{11}$$

ここで，tanh() は活性化関数として tanh を用いた非線形変換関数である．また，$\boldsymbol{\theta}_{\texttt{w}}^{(m)}$ は m 番目のドキュメントにおける非線形変換関数のモデルパラメータである．$\alpha_{it}^{(m)}$ は m 番目のドキュメントにおける i 番目の文の t 番目の単語の重要度の重みを表す．$\bar{\boldsymbol{u}}^{(m)}$ は m 番目のドキュメントの単語メモリリーダであり，これは単語注意機構に用いられる．単語メモリリーダは他のモデルパラメータと同様に学習で最適化される．

それぞれの文表現は，文エンコーダを用いて前後の文を考慮した文隠れ表現に変換される．文エンコーダは，単語エンコーダと同様に双方向 GRU を使用している．m 番目のドキュメントにおける i 番目の文隠れ表現は以下の式で表される．

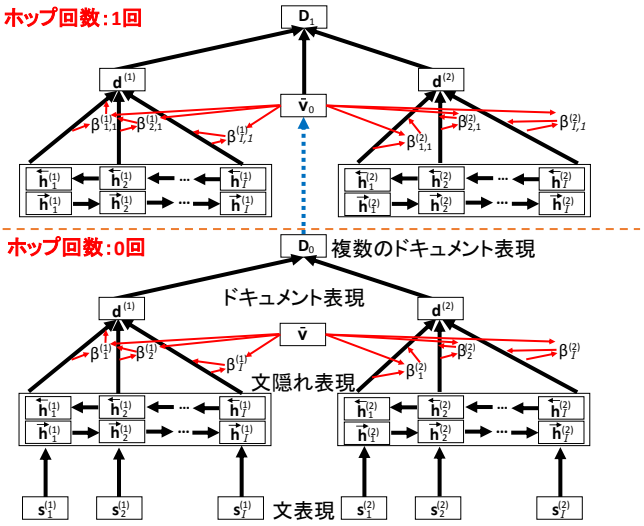

図 2 1 回ホップ時におけるマルチホップの概要

$$\overrightarrow{\boldsymbol{h}}_i^{(m)} = \overrightarrow{\texttt{GRU}}(\boldsymbol{s}_i^{(m)}; \boldsymbol{\theta}_{\texttt{rs}}^{(m)}), \tag{12}$$
$$\overleftarrow{\boldsymbol{h}}_i^{(m)} = \overleftarrow{\texttt{GRU}}(\boldsymbol{s}_i^{(m)}; \boldsymbol{\theta}_{\texttt{ls}}^{(m)}), \tag{13}$$
$$\boldsymbol{h}_i^{(m)} = [\overrightarrow{\boldsymbol{h}}_i^{(m)\top}, \overleftarrow{\boldsymbol{h}}_i^{(m)\top}]^{\top}, \tag{14}$$

ここで，$\boldsymbol{\theta}_{\texttt{rs}}^{(m)}$ と $\boldsymbol{\theta}_{\texttt{ls}}^{(m)}$ は文レベルの GRU のモデルパラメータである．$\boldsymbol{h}_i^{(m)}$ は，$\overrightarrow{\boldsymbol{h}}_i^{(m)}$ と $\overleftarrow{\boldsymbol{h}}_i^{(m)}$ を結合した結果である．

ドキュメント表現は文注意機構を使用し文隠れ表現をまとめることで求める．さらに，複数のドキュメント表現は全てのドキュメント表現を足し込むことで求める．ドキュメント表現と複数のドキュメント表現は以下のように求められる．

$$\boldsymbol{v}_i^{(m)} = \texttt{tanh}(\boldsymbol{h}_i^{(m)}; \boldsymbol{\theta}_{\texttt{s}}^{(m)}), \tag{15}$$
$$\beta_i^{(m)} = \frac{\exp(\boldsymbol{v}_i^{(m)\top} \bar{\boldsymbol{v}}^{(m)})}{\Sigma_{j=1}^{I_m} \exp(\boldsymbol{v}_j^{(m)\top} \bar{\boldsymbol{v}}^{(m)})}, \tag{16}$$
$$\boldsymbol{d}^{(m)} = \sum_{i=1}^{I_m} \beta_i^{(m)} \boldsymbol{h}_i^{(m)}, \tag{17}$$
$$\boldsymbol{D} = \sum_{m=1}^{M} \boldsymbol{d}^{(m)}, \tag{18}$$

ここで，$\beta_i^{(m)}$ は m 番目のドキュメントにおける i 番目の文の重要度の重みを表す．$\boldsymbol{\theta}_{\texttt{s}}^{(m)}$ は m 番目のドキュメントにおける非線形変換関数のモデルパラメータである．また，$\bar{\boldsymbol{v}}^{(m)}$ は m 番目のドキュメントの文メモリリーダであり，学習により最適化される．$\boldsymbol{d}^{(m)}$ と $\boldsymbol{D}$ は m 番目のドキュメント表現と複数ドキュメント表現である．

最後に，複数のドキュメント表現を使ってラベルの確率を推定する．

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

ここで，SOFTMAX() はソフトマック関数で，$\boldsymbol{\theta}_{\texttt{o}}$ はソフトマック関数のパラメータである．出力である l 次元の $\boldsymbol{O}$ は $P(l|D, \boldsymbol{\Theta})$ に対応する．

表 1 評価実験のためのデータセットの内訳

ラベル	学習	評価	テスト
新規契約	54	54	108
ダウングレード	30	30	60
アップグレード	29	28	56
オプション追加	13	12	25
オプション削除	13	12	25
ID・パスワード問い合わせ	14	12	25
名義変更	25	25	50
解約	26	25	50
合計	204	198	399

3.1.2 最適化

PHAN における，学習パラメータ $\boldsymbol{\Theta}$ を以下に示す．

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

ここで，m は $m \in [1, \cdots, M]$ である．パラメータは，正解ラベルと推定ラベルにおける最小クロスエントロピーを用いて最適化する．

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

ここで，$\hat{\boldsymbol{O}}_l^D$ と $\boldsymbol{O}_l^D$ は，複数ドキュメント D のラベル l の正解確率と推定確率である．$\mathcal{D}$ は学習データを示す．

3.2 共有メモリリーダ

本稿では，共有メモリリーダを提案する．共有メモリリーダとは，ドキュメントごとの共通な重要情報を際立たせる方法である．実際に，言語間で共有される注意機構は機械翻訳においても導入されている [21]．標準の PHAN では，単語メモリリーダと文メモリリーダはドキュメントごとに持っている．そこで，単一の単語メモリリーダと単一の文メモリリーダを全てのドキュメントで共有する．このとき，$\alpha_{it}^{(m)}$ と $\beta_i^{(m)}$ は以下の式で決定される．

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

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

ここで，$\bar{\boldsymbol{u}}$ と $\bar{\boldsymbol{v}}$ は共有単語メモリリーダと共有文メモリリーダである．共有メモリリーダは他の学習パラメータと同時に学習される．

3.3 マルチホップ

本稿では，マルチホップと呼ばれる追加機構を提案する．このマルチホップは共有文メモリリーダを繰り返し更新する機構である．この機構は，end-to-end memory networks[12] から着想している．マルチホップの概要を図

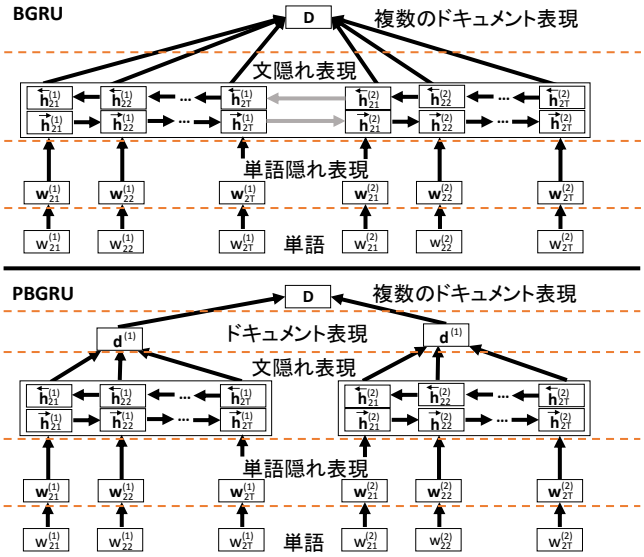

図 3 ベースライン手法の概要

2 に示す．PHAN では，共有文メモリリーダの更新のために複数ドキュメント表現を繰り返し使用する．更新方法を以下に示す．

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

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

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

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

ここで，$\boldsymbol{D}_0$ は式 (18) の $\boldsymbol{D}$ に対応される．$\bar{\boldsymbol{v}}_{k-1}$ は，k 回繰り返し更新した共有メモリリーダである．複数のドキュメント表現を求めた後，複数のドキュメント表現を次のホップの共有メモリリーダとして用いる．$\mathtt{Linear}()$ は線形変換関数であり，そして $\boldsymbol{\theta}_{\mathtt{h}}^{(m)}$ は他の学習パラメータと同時に最適化する．K 回ホップした後，更新した複数のドキュメント表現 $\boldsymbol{D}_K$ は式 (19) でのクラス分類に用いられる．マルチホップを導入することで，それぞれの対話の情報を全体に共有する注意機構に反映することができる．

4. 評価実験

4.1 実験条件

評価実験のタスクは，コンタクトセンタのテーマ分類である．コンタクセンタ対話データセットは，日本語で話された 8 つのテーマが存在するデータセットを用いた．表 1 に学習と評価，テストデータのそれぞれの対話数を示している．1 つの対話に 1 人のオペレータと 1 人の顧客が電話で会話しており，それぞれ話者ごとに別々に記録されている．本稿では，このデータセットの書き起こしを用いて，日本語話し言葉コーパス [23] から訓練されたディープニューラルネットワークに基づく発話検出器 [22] を用いて発話を

表 2 コンタクトセンタのテーマ分類精度 [%]

		マルチストリーム	HAN	共有メモリリーダ	マルチホップ	評価セット	テストセット
ベースライン	BGRU					88.8	83.7
	PBGRU	√				85.3	80.2
提案手法	PHAN	√	√			87.5	86.6
	PHAN-SMR	√	√	√		90.7	87.3
	PHAN-SMR-MH	√	√	√	√	**91.4**	**87.9**

分割した．各対話には，話者ごとに平均して約 148 発話が含まれている．また，各発話には最大 442 語存在する．

提案手法を評価するために 5 つの手法を用いた．

- **BGRU**: 双方向 GRU に基づくシングルドキュメントクラス分類手法である．この手法は，マルチストリーム文書を単一の文書とみなした手法である．具体的には，各単語を単一の線形関数を使用して単語表現で変換する．次に，各単語表現は，単一の双方向 GRU に渡される．ドキュメント表現は，双方向 GRU の出力を平均化することで獲得する．出力レイヤーは，ドキュメント表現を使用して予測確率を生成する．
- **PBGRU**: 複数の双方向 GRU に基づくマルチストリームドキュメントクラス分類手法である．この方法は，各ストリームに対して異なるパラメータを導入している．具体的には，各ドキュメントの各単語は，各ストリームごとの線形関数を使用して単語表現に変換する．次に，各ドキュメントの各単語表現は，ドキュメントごとの双方向 GRU 関数に渡される．ドキュメント表現は，各ドキュメントの双方向 GRU の出力を平均化することで獲得する．複数のドキュメント表現は，ドキュメント表現を平均化することで獲得する．この，複数のドキュメント表現を用いて予測確率を計算する．
- **PHAN**: 3.1 節で述べた PHAN に基づくマルチストリームドキュメントクラス分類手法である．
- **PHAN-SMR**: 3.2 節で述べた共有メモリリーダを導入した PHAN に基づくマルチストリームドキュメントクラス分類手法である．
- **PHAN-SMR-MH**: 3.3 節で述べたマルチホップと共有メモリリーダを導入した PHAN に基づくマルチストリームドキュメントクラス分類手法である．ホップ回数 K は 3 回とした．

BGRU と PBGRU は，従来手法と同様の構造であり，ベースライン手法となる．2 つのベースライン手法の概要を図 3 に示す．全ての手法のハイパーパラメータは統一した．まず，単語表現は 32 次元とし，文表現とドキュメント表現は 64 次元とした．学習時のミニバッチサイズを 5 に設定した．最適化関数を Adam とし，学習エポック数は，評価データの損失の最適値が 6 回連続で更新されなかった場合に学習を終了させた．それぞれの手法において，5 回の学習を行い，全てのモデルでの平均精度で評価を行った．

4.2 実験結果

表 2 に，評価セットとテストセットにおけるコンタクトセンタのテーマ分類精度の結果を示す．また，表 2 は各方法を考慮しているかどうかを示している．

まず，ベースライン手法の 2 つを比較すると，PBGRU は BGRU より劣っていた．これは，今回の学習データ量が PBGRU を学習するには不十分であることが考えられる．実際，PBGRU は BGRU よりパラメータ数が 2 倍多くなっている．そのため，学習データが多い場合は PBGRU のほうが高い性能が得られると考えられる．

次に，マルチストリーム音声ドキュメントのクラス分類手法で比較すると，PHAN が PBGRU より高い性能が示された．このことから，マルチストリーム音声ドキュメントのクラス分類において，階層構造と注意機構の両方がうまく性能改善に寄与したことが分かる．PHAN はベースラインよりモデルパラメータ数が増加しているが，テストセットにおいて BGRU 手法より高い性能が得られた．

さらに，PHAN-SMR は PHAN よりも高い性能が得られた．これは，共有メモリリーダを導入した PHAN は，別のストリームのドキュメントから重要な情報を抽出し，複数のドキュメントにまたぐ情報を正確に持てるからだと考えられる．加えて PHAN-SMR は評価セットとテストセットの両方で BGRU より高い性能が得られている．この結果から，共有メモリリーダの導入によってモデルパラメータ数が削減でき，学習データ量が限られている中で効率的な学習が実現していると言える．

最後に，PHAN-SMR-MH が最も高い性能が得られた．これは，マルチホップが全ドキュメントの情報をそれぞれのドキュメントに反映することで，他のドキュメントの内容を考慮した重要情報の抽出を可能にできたためだと考えられる．

以上をまとめると，提案手法の分類精度が，ベースラインと比較して，評価セットでは 2.6 ポイント，テストセットでは 4.2 ポイント改善した．

4.3 考察

評価実験の結果から，コンタクトセンタのテーマ分類タスクにおいて提案手法が高い性能であることが分かった．これは，それぞれの注意機構がクラス分類における複数話者の重要な情報を正しく抽出できたからだと考えられる．そこで，今回の実験においてどのような発話が抽出されて

いるかの調査を行った．ここでは，一例として「ダウングレード」をテーマとした対話について述べる．これは以下に示すような内容の発話を含んだ対話である．

(1) 顧客：値段が高いため昔のプランに変更したい
(2) オペレータ：従量制のプランがおすすめですよ
(3) 顧客：じゃあそのプランにしようかな

ベースライン手法である BGRU では，この対話をアップグレードと誤分類してしまった．しかし，提案手法は正しくダウングレードと分類した．この対話において提案手法が抽出した顧客の文は，「昔に値段が安いプランから高いプランに変更した」という文であった．この文からは顧客が契約しておりアップグレードしたことがあることが分かる．このことから，顧客が値段が高いプラン変更に変更したことについて意見があることが分かる．また，オペレータ側では，「昔使っていたプランにするなら従量制のプランに変更することもできます」という文が特に抽出されていた．このオペレータの文からも，顧客が値段の安いプランに変更したいということが分かる．提案手法では，このような文を抽出できたことにより正しく分類できたと考えれる．

5. まとめ

本稿では，マルチストリーム音声ドキュメントのクラス分類のための共有メモリリーダを導入した PHAN を提案した．PHAN は，ドキュメント内の情報を階層的にまとめ，注意機構により重要な情報を正確に抽出することができる．そして，共有メモリリーダは者間で共通情報を際立たせる共通注意機構を実現することができる．さらに，共有メモリリーダを繰り返し更新するマルチホップを提案した．コンタクトセンタのテーマ分類タスクを用いた評価実験の結果，PHAN と共有メモリリーダは従来手法よりも高い性能が示された．

今後は，音声認識結果を用いた評価実験を行う．また，マルチストリーム音声ドキュメントの構造を考慮したさらなる追加機構を導入することを考えている．

参考文献

[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.