会話ドキュメントに対する発話単位系列ラベリングのための自己教師あり事前学習

増村 亮　庵 愛　高島 瑛彦
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

ryou.masumura.ba@hco.ntt.co.jp

1 はじめに

音声認識の進展に伴い，人対人の会話の言語情報を理解して活用するニーズが高まってきている．例えばコンタクトセンタにおける電話の会話ドキュメントを理解することにより，顧客のニーズや自社の課題を発見するサービスが展開されており，さらなる高度化が望まれる．

本稿では，会話ドキュメントの理解に重要となる**発話単位系列ラベリング**に焦点を当てる．発話単位系列ラベリングは，ドキュメントが与えられた場合に発話ごとにラベルを推定する教師あり学習のタスクであり，トピックセグメンテーションや発話行為推定などに用いられる [1–5]．複数人が登場する会話ドキュメントを扱う場合では，誰が何をどんな順番で話したかを精緻に捉えることが重要となり，話者ラベル系列と発話文系列から階層的なネットワーク構造でモデル化する方法が高い性能を実現している [3]．

会話ドキュメントに対する発話単位系列ラベリングの学習には，ラベル付き会話ドキュメントが必要となる．しかしながら，このようなデータセットを大量に集めることは困難であり，限られたデータ量のラベル付き会話ドキュメントのみでは会話コンテキストを精緻に捉えたモデル化は難しい．そこで，本稿ではラベルなしデータセットを活用した**自己教師あり事前学習** [6–10] に着目する．自己教師あり事前学習は，大量のラベルなしデータからネットワーク構造をあらかじめ事前学習しておき，ラベル付きデータを用いてファインチューニングする方法であり，様々な自然言語処理タスクで成功を収めている．ラベルなし会話ドキュメントは，音声認識を活用することにより比較的大量に得られるため，自己教師あり事前学習は適したアプローチであると考えられる．一方で，会話ドキュメントに対する発話単位系列ラベリングに適した自己教師あり事前学習は，これまで検討されていない．

そこで本稿では，**会話ドキュメントに特化した自己教師あり事前学習 (Large-Context Conversational Representation Learning; LCCRL)** を提案する．提案手法では，会話中のある発話を，過去と未来の全ての発話系列情報をコンテキストとして予測する問題を自己教師あり事前学習のタスクとして設定する．具体的には，階層リカレントニューラルネットワークに基づく双方向長期コンテキスト言語モデル [11] に対して，話者ラベルと発話文を同時に予測するような拡張を行うことで自己教師あり学習を実現する．ファインチューニング時は，事前学習によって得られたネットワーク構造を発話系列全体を理解する構造に組み替えて，条件付き確率場 (CRF) を出力層とすることで教師あり学習を行う [2]．なお，提案手法のネットワークの一部で従来のトークン表現の自己教師あり事前学習を用いることが可能であり，テキストのみのデータとラベルなし会話データの両者を同時に活用することが期待できる．

コンタクトセンタ会話に対する発話単位系列ラベリングの評価実験において，従来の自己教師あり事前学習を用いる方法や全く用いない方法と比較して，提案手法の自己教師あり事前学習が有効であることを示す．

2 関連文献

発話単位系列ラベリング:　発話単位系列ラベリングは，トピックセグメンテーションや対話行為推定など用いられてる [1–5]．トークン間，および文間の長距離のコンテキストを捉えるために，トークン単位と文単位を組み合わせた階層リカレントニューラルネットワークがよく用いられている．本稿の発話単位系列ラベリングでは，会話ドキュメントに特化した階層リカレントニューラルネットワークを採用する．さらに，会話ドキュメント向けの階層構造に特化した自己教師あり事前学習を導入する．

自己教師あり事前学習:　初期の自己教師あり事前学習では，コンテキスト独立なトークン表現 [6] が検討されてきたが，近年は ELMo や BERT 等の前後コンテキストに依存したトークン表現 [8–10] が大きく注目されている．一方，文表現に対しては，Skip-Thought Vector [7] 等の自己教師あり事前学習が検討されているが，コンテキスト独立な文表現であることが課題である．本稿では，会話ドキュメントを対象として，誰が何をどんな順番で話してきたかの前後コンテキストを考慮した発話表現を自己教師あり事前学習することを目的としており，我々の知る限り，会話ドキュメントに対する自己教師あり事前学習に関する初めての検討である．

長期コンテキスト言語モデル:　会話や談話などを扱うための長期コンテキスト言語モデルが近年提案されている [11–15]．最も代表的なアプローチは，階層リカレントニューラルネットワークを用いて長期コンテキストを捉えるアプローチである．長期コンテキスト言語モデルは，これまで主に生成やポストエディッティングのために用いられてきた．本稿では，長期コンテキスト言語モデルを会話ドキュメントに特化した形に拡張し，自己教師あり事前学習のために利用する．

3 問題設定

会話ドキュメントの発話単位系列ラベリングの問題設定について述べる．本稿では，会話ドキュメントを

発話系列 $\boldsymbol{U} = \{U^1, \cdots, U^T\}$ として表す．ここで t 番目の発話 U^t は，発話文 W^t と話者ラベル q^t の組から表され，W^t はトークン系列 $W^t = \{w_1^t, \cdots, w_{N^t}^t\}$ として表される．q^t は t 番目の発話に対応した話者ラベルであり，例えばコンタクトセンタ通話の会話ドキュメントにおいてはオペレータかカスタマーの 2 値ラベルである．このとき，発話単位系列ラベリングでは，$\boldsymbol{U}$ から発話単位のラベル系列 $\boldsymbol{O} = \{o^1, \cdots, o^T\}$ を予測する問題を扱う．ラベルはタスクに依存し，例えばトピックラベルや発話意図ラベルなどである．

4 提案手法

自己教師あり事前学習を利用した会話ドキュメントの発話単位系列ラベリングのモデル化方法について述べる．以降では，ラベル付き会話ドキュメント集合 $\mathcal{D}_{\texttt{pair}} = \{(\boldsymbol{U}_1, \boldsymbol{O}_1), \cdots, (\boldsymbol{U}_K, \boldsymbol{O}_K)\}$，ラベルなし会話ドキュメント集合 $\mathcal{D}_{\texttt{unpair}} = \{\boldsymbol{U}_1, \cdots, \boldsymbol{U}_M\}$ を用いて発話単位系列ラベリングをモデル化する流れについて述べる．なお，会話ドキュメント以外のテキスト集合も活用することを我々は想定しているが，以降の説明では省略する．

4.1 自己教師あり事前学習

会話ドキュメントに特化した自己教師あり事前学習 (Large-Context Conversational Representation Learning; LCCRL) について述べる．

モデル化　LCCRL では，会話ドキュメント内のある発話を全ての過去と未来の発話系列をコンテキストとして予測するモデル化を行う．すなわち，t 番目の発話 U^t の予測確率を (1) 式の通りモデル化する．

$$P(U^t|U^{1:t-1}, U^{t+1:T}; \boldsymbol{\Theta}_{\texttt{pre}}) = P(q^t|U^{1:t-1}, U^{t+1:T}; \boldsymbol{\Theta}_{\texttt{pre}}) \prod_{n=1}^{N_t} P(w_n^t|w_1^t, \cdots, w_{n-1}^t, q^t, U^{1:t-1}, U^{t+1:T}; \boldsymbol{\Theta}_{\texttt{pre}}) \quad (1)$$

ここで，$\boldsymbol{\Theta}_{\texttt{pre}}$ はモデルパラメータを表す．本モデル化は，発話エンコーダ，過去コンテキストエンコーダ，未来コンテキストエンコーダ，そして発話デコーダの 4 つの部位によって構成される．図 1 に，LCCRL に基づく自己教師あり事前学習のネットワーク構造を示す．

発話エンコーダ:　発話エンコーダでは，双方向 LSTM (BLSTM) と Self-Attetion 機構を用いることで，発話中の重要な言語コンテキストを話者ラベルを考慮して固定長ベクトルに埋め込む．t 番目の発話に対して，話者ラベルの埋め込みベクトル $\boldsymbol{q}^t$ とトークンの埋め込みベクトル系列 $\{\boldsymbol{w}_1^t, \cdots, \boldsymbol{w}_{N^t}^t\}$ から (2) 式の通りコンテキストを考慮したベクトル変換を行う．

$$\boldsymbol{C}^t = \texttt{BLSTM}([\boldsymbol{q}^{t\top}, \boldsymbol{w}_1^{t\top}]^\top, \cdots, [\boldsymbol{q}^{t\top}, \boldsymbol{w}_{N^t}^{t\top}]^\top; \boldsymbol{\theta}_{\texttt{c}}) \quad (2)$$

ここで、$\texttt{BLSTM}()$ は，BLSTM の機能を持つ関数であり，$\boldsymbol{\theta}_{\texttt{c}}$ は学習可能なモデルパラメータを表す．なお，

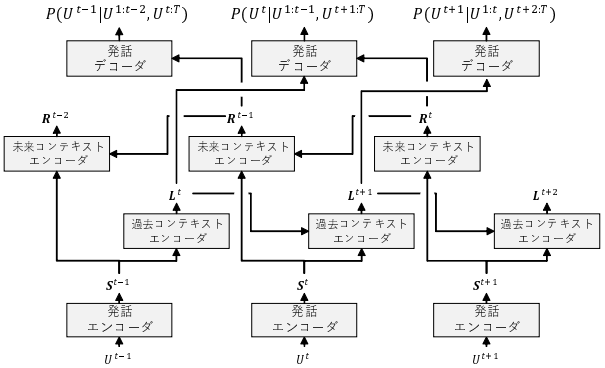

図 1: 自己教師あり事前学習時のネットワーク構造．

ELMo や BERT 等のトークン単位の自己教師あり事前学習は，トークンの埋め込みベクトルを構成するために利用できる．

次にコンテキストを考慮したベクトル系列を用いて，t 番目の発話文に対応する固定長ベクトルを (3) 式の通り構成する．

$$\boldsymbol{S}^t = \texttt{SelfAttention}(\boldsymbol{C}^t; \boldsymbol{\theta}_{\texttt{s}}) \quad (3)$$

ここで，$\texttt{SelfAttention}()$ は，入力されたベクトル系列に対して，各要素の重要度を考慮して重み付けて足し合わせることで固定長ベクトルに変換する関数であり，$\boldsymbol{\theta}_{\texttt{s}}$ は学習可能なモデルパラメータを表す．

過去コンテキストエンコーダ:　過去コンテキストエンコーダでは，発話単位の順方向 LSTM を用いることで，過去の発話系列全体を固定長ベクトルに変換する．1 番目の発話から $t-1$ 番目の発話までを埋め込んだ固定長ベクトルは (4) 式で表される．

$$\boldsymbol{L}^t = \overrightarrow{\texttt{LSTM}}(\boldsymbol{S}^1, \cdots, \boldsymbol{S}^{t-1}; \boldsymbol{\theta}_{\texttt{l}}) \quad (4)$$

ここで，$\overrightarrow{\texttt{LSTM}}()$ は順方向 LSTM の機能を持つ関数であり，$\boldsymbol{\theta}_{\texttt{l}}$ は学習可能なモデルパラメータを表す．

未来コンテキストエンコーダ:　未来コンテキストエンコーダでは，発話単位の逆方向 LSTM を用いることで，未来の発話系列全体を固定長ベクトルに変換する．T 番目の発話から $t+1$ 番目の発話までを埋め込んだ固定長ベクトルは (5) 式で表される．

$$\boldsymbol{R}^t = \overleftarrow{\texttt{LSTM}}(\boldsymbol{S}^{t+1}, \cdots, \boldsymbol{S}^T; \boldsymbol{\theta}_{\texttt{r}}) \quad (5)$$

ここで，$\overleftarrow{\texttt{LSTM}}()$ は逆方向 LSTM の機能を持つ関数であり，$\boldsymbol{\theta}_{\texttt{r}}$ は学習可能なモデルパラメータを表す．

発話デコーダ:　発話デコーダでは，過去と未来の全ての発話コンテキストから発話文と話者ラベルの予測を行うタスクを構成する．最初に話者ラベルの予測は (6) 式に従う．

$$P(q^t|U^{1:t-1}, U^{t+1:T}; \boldsymbol{\Theta}_{\texttt{pre}}) = \texttt{SOFTMAX}([\boldsymbol{L}^{t\top}, \boldsymbol{R}^{t\top}]^\top; \boldsymbol{\theta}_{\texttt{q}}) \quad (6)$$

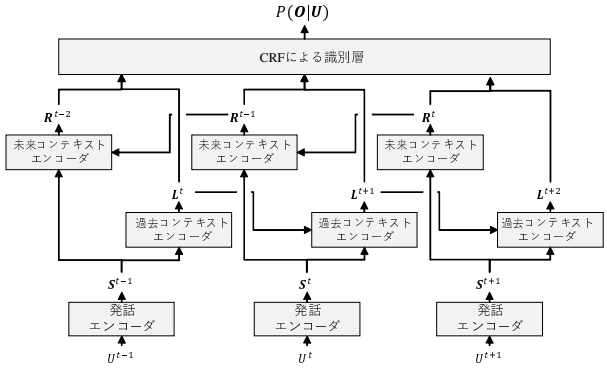

図 2: ファインチューニング時のネットワーク構造.

ここで SOFTMAX() は Softmax アクティベーションを含む線形変換関数であり，$\boldsymbol{\theta}_{\mathrm{q}}$ は学習可能なモデルパラメータを表す．次に発話文の予測は，自己回帰生成モデルとしてモデル化を行う．t 番目の発話の n 番目のトークンの生成確率は (7)-(8) 式に従い計算できる．

$$P(w_n^t|w_1^t,\cdots,w_{n-1}^t,q^t,U^{1:t-1},U^{t+1:T};\boldsymbol{\Theta}_{\mathtt{pre}}) = \mathtt{SOFTMAX}(\boldsymbol{v}_n^t;\boldsymbol{\theta}_{\mathrm{d}}) \quad (7)$$

$$\boldsymbol{v}_n^t = \overrightarrow{\mathtt{LSTM}}([\boldsymbol{w}_{n-1}^{t\top},\boldsymbol{q}^{t\top},\boldsymbol{L}^{t\top},\boldsymbol{R}^{t\top}]^\top,\boldsymbol{v}_{n-1}^t;\boldsymbol{\theta}_{\mathrm{v}}) \quad (8)$$

ここで $\boldsymbol{\theta}_{\mathrm{d}}$，$\boldsymbol{\theta}_{\mathrm{v}}$ は学習可能なモデルパラメータを表す．

最適化 LCCRL によるモデルパラメータの最適化について述べる．ここで，モデルパラメータは $\boldsymbol{\Theta}_{\mathtt{pre}} = \{\boldsymbol{\theta}_{\mathrm{c}},\boldsymbol{\theta}_{\mathrm{s}},\boldsymbol{\theta}_{\mathrm{r}},\boldsymbol{\theta}_{\mathrm{q}},\boldsymbol{\theta}_{\mathrm{d}},\boldsymbol{\theta}_{\mathrm{v}}\}$ として表されることになる．このとき，モデルパラメータの最適化は (9) 式に従う．

$$\hat{\boldsymbol{\Theta}}_{\mathtt{pre}} = \underset{\boldsymbol{\Theta}_{\mathtt{pre}}}{\mathrm{argmin}} - \sum_{\boldsymbol{U}\in\mathcal{D}_{\mathtt{unpair}}}\sum_{t=1}^{T} \log P(U^t|U^{1:t-1},U^{t+1:T};\boldsymbol{\Theta}_{\mathtt{pre}}) \quad (9)$$

この最適化は，会話単位のデータでミニバッチを構成することにより，ミニバッチ確率的勾配降下法により実施できる．

4.2 ファインチューニング

事前学習したネットワーク構造を用いて会話ドキュメントに対する発話単位系列ラベリングのタスクにファインチューニングする場合について述べる．

モデル化 会話ドキュメントに対する発話単位系列ラベリングでは，$P(\boldsymbol{O}|\boldsymbol{U};\boldsymbol{\Theta}_{\mathtt{fine}})$ をモデル化する．ファインチューニング時には，事前学習したネットワークを組み替えて CRF の出力層を設けた階層 BLSTM-CRF [2] を会話ドキュメント向けに拡張したネットワーク構造により，全体のラベルの整合性まで考慮した発話単位系列ラベリングをモデル化する．すなわち，$P(\boldsymbol{O}|\boldsymbol{U};\boldsymbol{\Theta}_{\mathtt{fine}})$ を (10) 式に従い算出する．

$$P(\boldsymbol{O}|\boldsymbol{U};\boldsymbol{\Theta}_{\mathtt{fine}}) = \frac{\prod_{t=1}^{T}\exp\phi(o^{t-1},o^t,\boldsymbol{y}^t;\boldsymbol{\theta}^\circ)}{\sum_{\bar{\boldsymbol{O}}}\prod_{t=1}^{T}\exp\phi(\bar{o}^{t-1},\bar{o}^t,\boldsymbol{y}^t;\boldsymbol{\theta}^\circ)} \quad (10)$$

表 1: ラベル付き会話ドキュメントの詳細.

業種	通話数	発話数	単語数
金融	59	6,081	55,933
プロバイダ販売	57	3,815	47,668
地方自治体	73	5,617	48,998
通信販売	56	4,938	46,574
パソコン修理	55	6,263	55,101
携帯電話販売	61	5,738	51,061
6 業種の合計	361	32,452	305,351

$$\boldsymbol{y}^t = [\boldsymbol{L}^{t+1\top},\boldsymbol{R}^{t-1\top}]^\top \quad (11)$$

ここで，$\bar{\boldsymbol{O}} = \{\bar{o}^1,\cdots,\bar{o}^T\}$ である．$\phi()$ は CRF において入力の素性ベクトルに重みをかけるための線形変換関数であり，$\boldsymbol{\theta}^\circ$ はそのモデルパラメータを表す．

ファインチューニング時の重要な点は，過去コンテキストエンコーダと未来コンテキストエンコーダの出力を，事前学習時とは異なる使い方をしている点であり，ファインチューニング時には入力の全ての情報を考慮できるようにネットワーク構造を構成する．図 2 にネットワーク構造を示す．

最適化 ファインチューニング時の最適化について述べる．ここでモデルパラメータは $\boldsymbol{\Theta}_{\mathtt{fine}} = \{\boldsymbol{\theta}_{\mathrm{c}},\boldsymbol{\theta}_{\mathrm{s}},\boldsymbol{\theta}_{\mathrm{r}},\boldsymbol{\theta}_{\mathrm{q}},\boldsymbol{\theta}_{\mathrm{o}}\}$ として表され，$\{\boldsymbol{\theta}_{\mathrm{c}},\boldsymbol{\theta}_{\mathrm{s}},\boldsymbol{\theta}_{\mathrm{r}},\boldsymbol{\theta}_{\mathrm{q}}\}$ は自己教師あり事前学習済みのものとする．ファインチューニング時のモデルパラメータの最適化は (12) 式に従う．

$$\hat{\boldsymbol{\Theta}}_{\mathtt{fine}} = \underset{\boldsymbol{\Theta}_{\mathtt{fine}}}{\mathrm{argmin}} - \sum_{(\boldsymbol{U},\boldsymbol{O})\in\mathcal{D}_{\mathtt{pair}}} \log P(\boldsymbol{O}|\boldsymbol{U};\boldsymbol{\Theta}_{\mathtt{fine}}) \quad (12)$$

この最適化は，会話単位のデータでミニバッチを構成することにより，ミニバッチ確率的勾配降下法により実施できる．なお，本稿では事前学習済みのパラメータを初期値として最適化時に更新することとする．

5 評価実験

データ 本稿では，コンタクトセンタにおけるオペレータとカスタマーの模擬対話データを用いてトピックセグメンテーションのタスクに対する評価を行った．我々はラベル付き会話ドキュメントとして 6 業種 361 通話を準備した．1 つの通話は，オペレータとカスタマーの対話データであり，本稿では人手による書き起こしテキストを発話文として用い，話者ごとに話者ラベルを付与した．我々はトピックラベルとして，オープニング，要件把握，要件対応，カスタマー情報把握，クロージングのシーンに対応した 5 ラベルを設定し，各発話に対して人手によりラベルを一意に割り当てた．表 1 に通話数，発話数，単語数等のラベル付き会話ドキュメントデータセットの詳細を示す．学習と評価は，6 分割交差検定とし，6 業種中 5 業種で学習し 1 業種で評価することにより，業種についてオープンな評価とした．また，自己教師あり事前学習のために，様々なドメインのコンタクトセンタデータのラベルなし会話ドキュメントを 4,000 通話準備した．さらに，Web 上のテキストを約 5 億文準備した．

実験条件 本評価では，様々なセットアップで学習した発話単位系列ラベリング手法を比較した．ベースライン

表 2: 発話単位系列ラベリングの評価結果．

手法	正解率 (%)
HBLSTM-CRF	83.8
HBLSTM-CRF-SPK	85.0
HBLSTM-CRF-SPK+ELMo	88.2
HBLSTM-CRF-SPK+SkipThought	87.2
HBLSTM-CRF-SPK+ELMo+SkipThought	88.7
HBLSTM-CRF-SPK+LCCRL	89.6
HBLSTM-CRF-SPK+ELMo+LCCRL	**90.4**

として，階層 BLSTM-CRF (HBLSTM-CRF)，および会話ドキュメントの話者ラベルまで考慮した HBLSTM-CRF-SPK を準備した．単語の埋め込みベクトルの次元数は 512，話者ラベルの埋め込みベクトルの次元数は 32，LSTM および BLSTM は，出力連続ベクトルの次元数が 512 次元となるようにそれぞれ 2 層用いてユニットを形成した．自己教師あり事前学習を用いる方法として，ELMo，Skip-Thought Vector，LCCRL を用いた．ELMo の学習には Web 上のテキストを用い，Skip-Thought Vector と LCCRL の学習にはラベルなし会話ドキュメントを使用した．また，2 段階の自己教師あり事前学習についても検討した．この場合，事前学習時のみに必要なネットワーク構造も前述と同様に設定した．ネットワークの教師あり事前学習，ファインチューニングでは 5 通話で 1 ミニバッチを形成し，Adam により最適化を行った．その際，学習データの一部を開発データに用いてアーリーストッピングを行った．なお，初期値を変化させて 1 条件につき 5 回ネットワークを構築し，開発データについて最もロスが減少したモデルを用いて評価を行った．

実験結果　表 2 に各セットアップについての正解率に関する評価結果を示す．最初にベースラインの 2 手法を比較すると，話者ラベルまで考慮した HBLSTM-CRF-SPK の方が高い性能を示すことが見て取れ，会話ドキュメントの理解には「誰が」までを捉えることが有用であることが示された．次に，HBLSTM-CRF-SPK に自己教師あり事前学習を含めた結果を比較すると，LCCRL は ELMo や Skip-Thought Vector よりも有効であることが見て取れる．これは，LCCRL による事前学習により，発話間の長期コンテキストにわたる関係性を学習できたためであると考えられる．さらに，LCCRL と ELMo の二段階の事前学習を導入することにより最高性能が得られ，テキストのみのデータとラベルなし会話データの両者を同時に活用して自己教師あり事前学習することが有用であることが確認できた．図 3 に学習データ量を変化させた場合の正解率に関する評価結果を示す．ELMo のみを用いる場合では発話間の関係性を事前学習時に捉えられないため，ラベル付き会話ドキュメントの量が少ないと性能が低いことが見て取れる．一方，LCCRL を用いる場合は，学習に用いるラベル付き会話ドキュメントの量が少ない際も高い性能が得られることが見て取れ，事前学習の時点で有用なパラメータが学習できていることが確認できる．以上の結果から，会話の発話単位系列ラベリングにおける提案手法の有効性が示された．

6　おわりに

本稿では，会話ドキュメントに対する発話単位系列ラベリングのための自己教師あり事前学習 (Large-Context Conversational Representation Learning; LCCRL) を提案した．提案手法の強みは，発話間の長期コンテキストにわたる関係性をラベルなし会話ドキュメント群から学習できる点であり，従来のテキスト集合からの自己教師あり事前学習と組み合わせることで強力な事前学習を実現できる．コンタクトセンタ会話に対する発話系列ラベリングの評価実験において，提案手法の自己教師あり事前学習を用いることにより，ラベル付きドキュメントが少ない場合，およびある程度収集できる場合の両者において，従来の自己教師あり事前学習を用いる方法や全く用いない方法と比較して高い性能を達成できることを示した．

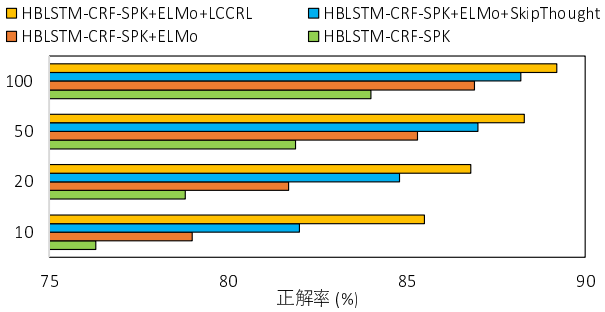

図 3: ラベル付き会話ドキュメントの量を変化させた場合の評価結果．

参考文献

[1] Q. H. Tran *et al.*, “A hierarchical neural model for learning sequences of dialogue acts,” *In Proc. EACL*, vol. 1, pp. 428–437, 2017.

[2] H. Kumar *et al.*, “Dialogue act sequence labeling using hierarchical encoder with CRF,” *In Proc. AAAI*, pp. 3440–3447, 2018.

[3] R. Masumura *et al.*, “Online call scene segmentation of contact center dialogues based on role aware hierarchical LSTM-RNNs,” *In Proc. APSIPA ASC*, pp. 811–815, 2018.

[4] W. Jiao *et al.*, “HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition,” *In Proc. NAACL-HLT*, pp. 397–406, 2019.

[5] Y. Yu *et al.*, “Modeling long-range context for concurrent dialogue acts recognition,” *In Proc. CIKM*, p. 22772280, 2019.

[6] T. Mikolov *et al.*, “Distributed representations of words and phrases and their compositionality,” *In Proc. NIPS*, pp. 3111–3119, 2013.

[7] R. Kiros *et al.*, “Skip-thought vectors,” *In Proc. NIPS*, pp. 3294–3302, 2015.

[8] M. E. Peters *et al.*, “Semi-supervised sequence tagging with bidirectional language models,” *In Proc. ACL*, pp. 1756–1765, 2017.

[9] M. E. Peters *et al.*, “Deep contextualized word representations,” *In Proc. NAACL-HLT*, pp. 2227–2237, 2018.

[10] J. Devlin *et al.*, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *In Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[11] R. Masumura *et al.*, “Generalized large-context language models based on forward-backward hierarchical recurrent encoder-decoder models,” *In Proc. ASRU*, pp. 554–551, 2019.

[12] R. Lin *et al.*, “Hierarchical recurrent neural network for document modeling,” *In Proc. EMNLP*, pp. 899–907, 2015.

[13] A. Sordoni *et al.*, “A hierarchical recurrent encoder-decoder for generative context-aware query suggestion,” *In Proc. CIKM*, pp. 553–562, 2015.

[14] I. V. Serban *et al.*, “Building end-to-end dialogue systems using generative hierarchical neural network models,” *In Proc. AAAI*, pp. 3776–3783, 2016.

[15] T. Wang and K. Cho, “Larger-context language modelling with recurrent neural network,” *In Proc. ACL*, pp. 1319–1329, 2016.