システム発話の文脈を考慮した発話意図理解

高津 弘明[1]　横山 勝矢[1]　本田 裕[2]　藤江 真也[1,3]　小林 哲則[1]

早稲田大学[1]　本田技術研究所[2]　千葉工業大学[3]

{takatsu,katsuya}@pcl.cs.waseda.ac.jp, Hiroshi_01_Honda@n.t.rd.honda.co.jp,
shinya.fujie@p.chibakoudai.jp, koba@waseda.jp

1　はじめに

システム発話の文脈を考慮した発話意図識別モデルを提案する．

我々はニュース記事のようなまとまった量の情報を効率的に伝達する会話システムの開発を行っている [高津 18a]．ここで「効率的」とは，伝達対象となる記事の中から，ユーザーにとって不要な情報を除き，必要な情報だけを伝えることを意味する．我々のシステムの特徴は，あらかじめ主計画，副計画と呼ぶ複数のシナリオを用意しておき，このシナリオに沿って会話を進めることで，リズムの良い会話を実現するうえで必須となる迅速な応答を可能としたところにある．主計画に沿って記事の要点となる情報を提示する傍らで随時ユーザーからのフィードバックを理解し，必要に応じて副計画に遷移して補足情報を提示する．このようにユーザーの興味や理解状態に応じて提示する情報を柔軟に切り替えながら会話を進めていく仕組みを持つ．一方で，高い情報伝達効率（EoIT; Efficiency of Information Transfer）[高津 18a] を実現するには，ユーザーからのフィードバックを正しく理解することが重要となる．

ユーザーのフィードバックは必ずしも言語的に明示された形で表れるとは限らない．場合によっては，抑揚で表現されるニュアンスなどにユーザーの意図が表れることもある．そこで，フィードバックの種類（以下，発話意図）を分類し，スペクトログラムから抽出した音響特徴量とユーザー発話から抽出した言語特徴量から発話意図を認識するモデルを提案した [高津 18b, Yokoyama 18]．しかしながら，フィードバックに込められる意図をユーザーの発話情報のみから判断するのは難しい．例えば，「え？」という発話が驚きなのか，質問なのか，聞き返しなのかは文脈に依存する．

そこで，本研究では，システム発話の文脈を考慮した発話意図認識モデルを提案する．さらに，言語情報のエンコーダーに BERT（Bidirectional Encoder Representations from Transformers）[Devlin 18] を採用したうえで，Masked Language Model と Next Sentence Prediction による BERT の事前学習の発話意図認識における効果についても確認する．

本稿の構成は次の通りである．2 章で関連研究について述べる．3 章で発話意図データセットについて説明し，4 章で提案する発話意図認識モデルについて説明する．5 章で発話意図データセットを用いてモデルを評価した結果を報告する．6 章でまとめと今後の課題について述べる．

2　関連研究

従来，発話意図認識や感情認識のタスクでは，人手で設計した特徴量が利用されてきた [Ando 15, 藤江 03, 林 14, Nisimura 06, 田中 98]．例えば，藤江らは，システムに対する利用者の発話態度（肯定的か否定的か）を推定するために，第 1 モーラの基本周波数（F0）の傾き，発話全体の F0 レンジ，最終モーラの継続長からなる 3 次元の特徴量を使用した [藤江 03]．Ando らは，対話データにおいてその発話が肯定的であるか否定的であるかを推定するために，音響特徴量として単語ごとに算出した F0，パワーの最大・最小，継続長，間などの情報を，言語特徴量としてバイグラム言語モデルのパープレキシティを使用した [Ando 15]．林らは，発話タグ（疑問文，平叙文，相槌，同意，笑い）の推定において，F0 値やパワー，話速などの 23 次元の音響特徴量を使用した [林 14]．しかしながら，F0 などを特徴量として利用する場合，これらの推定誤りが認識精度の低下につながることが考えられる．

近年では，スペクトログラムを直接入力して感情認識などを行う研究が増えてきている [Guo 18, Luo 18, Satt 17, Tang 18, Yenigalla 18]．例えば，Guo らは，スペクトログラム，位相情報，MGDCC [Hegde 07] を CNN に入力して得られた特徴量を双方向 LSTM に与え感情（怒りや喜びなど）を識別するモデルを提案した [Guo 18]．Luo らは，スペクトログラムを CRNN に入力して得られた特徴量と人手で設計された特徴量（F0 や MFCC など）を組み合わせて感情を識別するモデルを提案した [Luo 18]．Yenigalla らは，スペクトログラムを CNN に入力して得られた特徴量と word2vec [Mikolov 13] で得られた音素の埋め込み表現を組み合わせて感情を識別するモデルを提案した [Yenigalla 18]．

我々は，スペクトログラムを CNN を含む AutoEncoder に入力し，その中間層出力を音響特徴量として LSTM に与え，その LSTM の最終状態の出力と音声認識結果から得られるユーザー発話の言語情報を組み合わせて発話意図を識別するモデルを提案した [高津 18b, Yokoyama 18]．本研究では，言語情報のエンコーダーに BERT [Devlin 18] を採用し，ユーザー発話だけでなくシステム発話の文脈も考慮した発話意図認識モデルを提案する．

3　発話意図データセット

発話意図の情報が付与されたコーパスとして横山らが構築したデータセットを使用した [Yokoyama 18]．このデータセットは，我々が情報伝達のために開発した即応性に富む会話システム [高津 18a] と 24 名の大学生が会話して得られた約 2,000 対話分の音声対話データに基づいて作られたデータセットである．収集したユーザー発話のうち，VAD で切り出した 1.5 秒以下の音声に対して 10 人のアノテーターが発話意図に関するアノテーションを行った．

発話意図の分類と対応するシステム動作を表 1 に示す．伝達情報の増加を求める発話意図として「質問」「補足要求」「反

表 1: 発話意図の分類と対応するシステム動作

効果	発話意図	システムアクション
伝達情報を増やす	質問	質問応答
	補足要求	補足説明
	反復要求	繰り返し
伝達情報を減らす	無関心	他トピックへの移行
	既知	詳細説明の省略
発話の衝突を避ける	待機要求	傾聴

表 2: 実験で使用した発話意図データセットの統計

	訓練セット		開発セット		テストセット	
	正例	負例	正例	負例	正例	負例
質問	2042	2042	292	276	584	584
補足要求	3228	3197	461	457	924	915
反復要求	360	358	52	52	102	102
無関心	1776	1763	254	252	508	504
既知	355	350	51	51	101	102
待機要求	361	359	52	52	104	102

復要求」を，伝達情報の減少を求める発話意図として「無関心」「既知」を，発話衝突の回避を求める発話意図として「待機要求」を定めた．本研究の実験で使用した発話意図データセットの統計を表 2 に示す．

4 発話意図認識モデル

提案する発話意図認識モデルの全体像を図 1 に示す．まず，短い時間幅で切り出した音声の断片からスペクトログラムを生成する．次に，得られたスペクトログラムを CNN を含む AutoEncoder（CNN-AutoEncder）に入力し，その中間層に圧縮された韻律特徴量を時系列に沿って LSTM に入力する．LSTM は逐次発話意図ラベルの確率を出力するが，音声認識結果が得られた段階で，LSTM の隠れ層の情報とユーザー発話の言語情報と直前のシステム発話の言語情報を統合して最終的な発話意図の推定結果を得る．以下，特徴抽出部と識別部について説明する．

4.1 特徴抽出部の設計

一般に発話意図の推定では，特徴量として基本周波数（F0）が用いられる．しかしながら，音声波形の準周期性や周辺雑音，有声音中の基本周波数の変化が広域に渡るなどの理由により，基本周波数を正確に抽出するのは難しい．そこで，F0 の推定を介さずに音声の時間・周波数スペクトルから直接特徴量を抽出する方法を提案した [高津 18b, Yokoyama 18]．

音韻や声の高さに関する特徴はスペクトログラムの模様として表れる．そこで，このスペクトログラムの模様を二次元の画像と見なして，5 層の畳み込み層と 3 層の全結合層からなるネットワークを折り返した全 16 層で構成される CNN-AutoEncoder（図 1.b）を学習する．そして，その中間層に圧縮された韻律特徴量を発話意図の識別で利用する．

表 3: 文脈として用いるシステム発話の範囲（太字部分）（S がシステム発話，U がユーザー発話を表す．）

S：	ツイッターのアカウントがハイジャックされる事件が起きたらしいよ
S：	**攻撃は XSS の脆弱性を突いたものだって**
S：	**ツイートにスクリプトを**
U：	何それ

4.2 識別部の設計

識別部は，韻律情報のみから発話意図を識別する識別部（P）と，識別部（P）で得られる韻律情報に加え，ユーザー発話および直前のシステム発話の言語情報を考慮して発話意図を識別する識別部（L）から構成される．

4.2.1 識別部（P）

識別部（P）では，CNN-AutoEncoder から取り出された韻律特徴量を LSTM に逐次入力して発話意図を識別する（図 1.c）．

音声は時間方向に可変長であり，時間方向の長さに頑健なモデルであることが望まれる．同じ文字列の音声でも人の違いや発話する条件や状態によってその継続長は異なる．また，発話末のピッチが上昇すると「質問」と捉えやすくなるなど，発話意図認識において韻律の時間方向の変化は有用な情報である．本研究では，RNN の中でも長期の依存性に長けた LSTM を用いた．

4.2.2 識別部（L）

識別部（P）の LSTM の状態とユーザー発話および直前のシステム発話の言語情報を用いて発話意図を識別する（図 1.d）．ここで，直前のシステム発話は，ユーザー発話を受け付けた時点での現在の発話内容とその一つ前の発話内容を表す．例えば，表 3 のような会話において，ユーザー発話「何それ」の発話意図を識別する場合，文脈として用いるシステム発話は現在の発話内容「ツイートにスクリプトを」とその一つ前の発話内容「攻撃は XSS の脆弱性を突いたものだって」となる．

ユーザー発話および直前のシステム発話の言語情報のエンコードには，BERT [Devlin 18] を用いた．BERT は，Transformer [Vaswani 17] の Encoder 部分をユニットとする双方向 Transformer モデルである．文の単語をランダムにマスクし，そのマスクされた単語を予測する Masked Language Model と二つ文が隣接しているかどうかを予測する Next Sentence Prediction の 2 つのタスクで事前学習したモデルを転移学習させることで，自然言語処理の様々なタスクで SOTA を達成し，汎用的な言語表現を獲得できるモデルとして注目されている．本研究では，BERT を発話意図認識タスクに適用し，事前学習の効果を検証する．

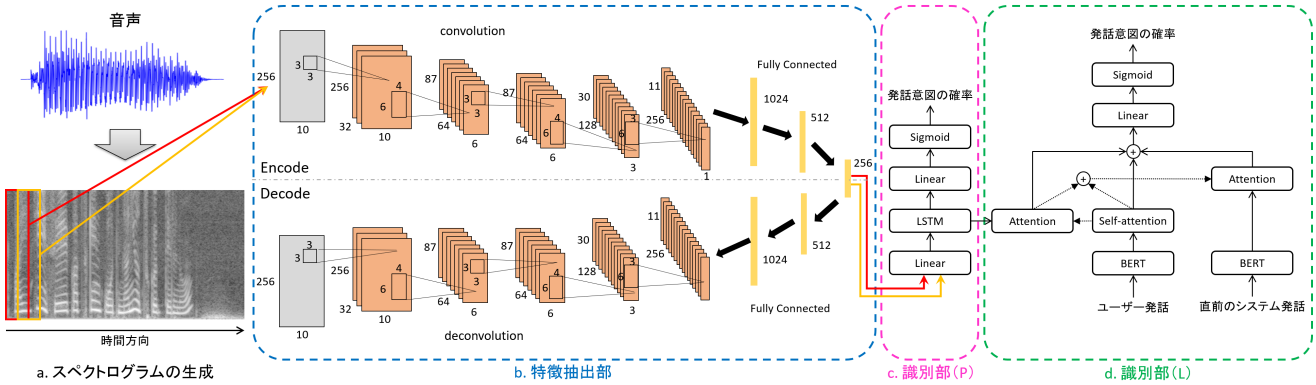

図 1: 発話意図認識モデル

ユーザー発話を BERT でエンコードした結果に対して self-attention [Lin 17] を適用し，得られたベクトル $\boldsymbol{v}_u$ をクエリとして韻律情報を保持する LSTM の各状態に対して attention [Bahdanau 15] を計算する．その結果得られる文脈ベクトル $\boldsymbol{v}_p$ とユーザー発話のベクトル表現 $\boldsymbol{v}_u$ を結合したベクトルをクエリとして，システム発話のエンコーダー BERT に対して attention [Bahdanau 15] を計算する．その結果得られる文脈ベクトル $\boldsymbol{v}_s$ を $\boldsymbol{v}_u$ および $\boldsymbol{v}_p$ と結合したベクトルを出力層に与え，最終的な発話意図の確率を求める．

5 発話意図識別実験

5.1 特徴抽出部の学習

特徴抽出部（図 1.b）の CNN-AutoEncoder の学習には，大規模な『日本語話し言葉コーパス（CSJ）』[1] を用いた．

CNN-AutoEncoder の入力は，フレームサイズ 800(50ms)，フレームシフト 160（10ms），チャンクサイズ 1024 で切り出した音声から生成したスペクトログラムを時系列に並べたものとし，そのサイズは 10×256 とした．このデータをもとにネットワークの学習を行い，特徴抽出器を構築した．

5.2 識別部の実験設定

Linear 層および LSTM の隠れ層の次元は 64 に設定した．BERT の事前学習の詳細および補助情報として用いた言語特徴量の説明を以下に示す．

5.2.1 BERT の事前学習

日経新聞の 200476 個のニュース記事から段落を超えないように隣接文ペアを重複なく抽出した．この内，700000 文ペアを訓練セット，37094 文ペアを開発セットとして，Masked Language Model と Next Sentence Prediction の 2 タスクで BERT の事前学習を行った．語彙には訓練セットにおいて頻度が 7 以上であった 63272 語を用いた．モデルのパラメータは，Transformer のブロック数を $L = 8$，隠れ層の次元を $H = 256$，self-attention のヘッド数を $A = 8$ に設定した．

[1] http://pj.ninjal.ac.jp/corpus_center/csj/

5.2.2 補助情報

システム発話の補助情報には，JUMAN++[2]（Ver.1.02）の形態素情報（品詞大分類，品詞細分類，活用形，活用型，カテゴリ，ドメイン），単語の TF，IDF，TF-IDF，「」『』内の単語かどうか，KNP[3]（Ver.4.19）を適用して得られる IREX の 8 種類の固有表現クラス，係り受けの種類，係り受け木の深さ，係り元の文節数，文頭からの文節位置を用いた．

シナリオにあるシステム発話は事前に解析しておくことができるのに対し，ユーザー発話はリアルタイムに解析する必要がある．そのため，ユーザー発話の補助情報には，JUMAN[4]（Ver.7.01）を適用して得られる形態素情報のみを使用した．

5.3 実験設定

発話意図データセット（3 章）を用いて，識別部（P）を学習し，LSTM の最終状態の出力結果をもとに，韻律情報のみを用いたときのテストセットに対する Accuracy を計算した．また，学習済みの識別部（P）の LSTM の隠れ層の値とユーザー発話（書き起こし）およびシステム発話の言語情報を用いて，識別部（L）を学習し，テストセットに対する Accuracy を計算した．

韻律情報のみを用いたとき（P）と，これにユーザー発話の言語情報を加えたとき（P+U），さらにシステム発話の言語情報を加えたとき（P+U+S）の比較を行った．また，BERT を事前学習したときとしなかったとき，補助情報を加えたときと加えなかったときについても比較を行った．

5.4 実験結果

実験結果を図 4 に示す．全体の傾向として，韻律情報のみよりもユーザー発話の言語情報を加えたときの方が良く，さらに直前のシステム発話の言語情報を加えることで Accuracy が向上することが分かった．また，BERT を事前学習することで，補助情報を加えることで Accuracy が向上することが分かった．

しかしながら，「無関心」の識別に関してはシステム発話の情報を加えることで性能が悪化した．これは，コンテンツ

[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++
[3] http://nlp.ist.i.kyoto-u.ac.jp/?KNP
[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

表 4: 発話意図認識の実験結果

		事前学習あり+補助情報あり		事前学習あり+補助情報なし		事前学習なし+補助情報なし	
	P	P+U	P+U+S	P+U	P+U+S	P+U	P+U+S
質問	0.902	0.955	**0.960**	0.932	0.944	0.929	0.935
補足要求	0.687	0.753	**0.759**	0.699	0.703	0.697	0.697
反復要求	0.593	0.775	**0.789**	0.618	0.637	0.598	0.613
無関心	0.760	**0.779**	0.772	0.770	0.768	0.768	0.760
既知	0.552	0.749	**0.759**	0.709	0.734	0.571	0.606
待機要求	0.709	0.738	**0.752**	0.728	0.733	0.714	0.728

に対する興味の傾向がユーザーごとに様々であることに起因すると考えられる．そのため，今後は，ユーザーを区別する識別子を補助情報に加えるなどして改善を図る．

6 おわりに

システム発話の文脈を考慮した発話意図識別モデルを提案し，直前のシステム発話の言語情報を文脈に用いることで発話意図認識の精度が向上することを確認した．

また，ユーザー発話およびシステム発話の言語情報のエンコーダーに BERT を採用し，Masked Language Model と Next Sentence Prediction の 2 タスクで事前学習することで，精度が向上することを確認した．

今後は，ユーザーの違いを考慮できるように特徴量を工夫するなどして性能の改善を目指すとともに，表情などの視覚情報も考慮したマルチモーダルな発話意図認識手法についても検討したい．

参 考 文 献

[Ando 15] Ando, A., Asami, T., Okamoto, M., Masataki, H., and Sakauchi, S.: Agreement and disagreement utterance detection in conversational speech by extracting and integrating local features, in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, pp. 2494-2498 (2015)

[Bahdanau 15] Bahdanau, D., and Cho, K., and Bengio, Y.: Neural machine translation by jointly learning to align and translate, in *Proceedings of the 3th International Conference on Learning Representations*, pp. 1-15 (2015)

[Devlin 18] Devlin, J., Chang, M.W., Lee, K., and Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding, *arXiv:1810.04805*, pp. 1-14 (2018)

[藤江 03] 藤江真也, 江尻康, 菊池英明, 小林哲則: パラ言語の理解能力を有する対話ロボット, 情報処理学会研究報告音声言語情報処理（SLP）, Vol. 2003, No. 104(2003-SLP-048), pp. 13-20 (2003)

[Guo 18] Guo, L., Wang, L., Dang, J., Zhang, L., Guan, H., and Li, X.: Speech emotion recognition by combining amplitude and phase information using convolutional neural network, in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pp. 1611-1615 (2018)

[林 14] 林佑樹, 大佛駿介, 中野有紀子: 協調学習における韻律特徴を用いた発話タグ推定モデル, 教育システム情報学会第 39 回全国大会, pp. 441-442 (2014)

[Hegde 07] Hegde, R.M., Murthy, H.A., and Gadde, V.R.R.: Significance of the modified group delay feature in speech recognition, *IEEE Transactions on Audio Speech and Language Processing*, Vol. 15, No. 1, pp. 190-202 (2006)

[Lin 17] Lin, Z., Feng, M., Santos, C.N., Yu, M., Xiang, B., Zhou, B., and Bengio, Y.: A structured self-attentive sentence embedding, in *Proceedings of the 5th International Conference on Learning Representations*, pp. 1-15 (2017)

[Luo 18] Luo, D., Zou, Y., and Huang, D.: Investigation on joint representation learning for robust feature extraction in speech emotion recognition, in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pp. 152-156 (2018)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, in *Proceedings of the 1st International Conference on Learning Representations*, pp. 1-12 (2013)

[Nisimura 06] Nisimura, R., Omae, S., Kawahawa, H., and Irino, T.: Analyzing dialogue data for real-world emotional speech classification, in *Proceedings of the 7th Annual Conference of the International Speech Communication Association*, pp. 1822-1825 (2006)

[Satt 17] Satt, A., Rozenberg, S., and Hoory, R.: Efficient emotion recognition from speech using deep learning on spectrograms, in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, pp. 1089-1093 (2017)

[高津 18a] 高津弘明, 福岡維新, 藤江真也, 林良彦, 小林哲則: 意図性の異なる多様な情報行動を可能とする音声対話システム, 人工知能学会論文誌, Vol. 33, No. 1, pp. 1-24 (2018)

[高津 18b] 高津弘明, 横山勝矢, 本田裕, 藤江真也, 林良彦, 小林哲則: 会話によるニュース記事伝達のための発話意図理解, 人工知能学会第 32 回全国大会論文集, 4Pin1-29, pp. 1-4, (2018)

[田中 98] 田中真詞 , 川端豪: 文型と音調によるユーザの発話意図の推定, 情報処理学会研究報告音声言語情報処理（SLP）, Vol. 1998, No. 68(1998-SLP-022), pp. 55-60 (1998)

[Tang 18] Tang, D., Zeng, J., and Li, M.: An end-to-end deep learning framework for speech emotion recognition of atypical individuals, in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pp. 162-166 (2018)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., and Kaiser, L.: Attention iss all you need, in *Proceedings of the 31st Conference on Neural Information Processing System*, pp. 6000-6010 (2017)

[Yenigalla 18] Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., and Vepa, J.: Speech emotion recognition using spectrogram & phoneme embedding, in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pp. 3688-3692 (2018)

[Yokoyama 18] Yokoyama, K., Takatsu, H., Honda, H., Fujie, S., and Kobayashi, T.: Investigation of users' short responses in actual conversation system and automatic recognition of their intentions, in *Proceedings of the the 2018 IEEE Workshop on Spoken Language Technology* (2018)