The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

会話によるニュース記事伝達のための発話意図理解

Utterance Intention Understanding for News Articles Transfer by Conversation

高津 弘明 *1 Hiroaki Takatsu　横山 勝矢 *1 Katsuya Yokoyama　本田 裕 *2 Hiroshi Honda　藤江 真也 *1*3 Shinya Fujie　林 良彦 *1 Yoshihiko Hayashi　小林 哲則 *1 Tetsunori Kobayashi

*1早稲田大学 Waseda University　*2本田技術研究所 Honda R&D Co.,Ltd.　*3千葉工業大学 Chiba Institute of Technology

We are developing a conversation system which efficiently transfers a massive amount of information like news articles by spoken dialogue. Here, "efficient" means that only the necessary information is transferred except unnecessary information for the user from target articles. In our system, feedbacks from the user are indispensable in order to realize high EoIT (Efficiency of Information Transfer). Therefore, we propose a utterance intention recognition method combining language information and prosodic information for the purpose of understanding diverse feedbacks from users. The feature of the proposed method is that it automatically extracts prosodic features with a high contribution ratio by using deep learning. We confirmed the effectiveness of the proposed method using a corpus with utterance intention tags designed based on dialogue data collected using our conversation system.

1. はじめに

発話意図理解のための深層学習に基づく韻律特徴量の自動抽出手法を提案する．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 18] を用いて収集した対話データを分析し設計した発話意図タグ付きコーパス [横山 18] を用いて提案手法の有効性を示す．

我々はニュース記事のようなまとまった量の情報を音声対話によって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 18]．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上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）を実現するにはユーザーからのフィードバックが必要不可欠である．フィードバックに関する技術課題として，フィードバックの誘発とフィードバックの理解という観点がある．本研究は後者に関するもので，ユーザーからの多様なフィードバックを理解することが目的である．

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，本研究では言語情報と韻律情報を組み合わせた発話意図認識手法を提案する．提案手法の特徴は，従来意図の種類ごとに観察に基づいて行われてきた特徴パラメータの設計を深層学習を用いることで，寄与率の高い韻律特徴量を自動で抽出しているところにある．実験ではデータセットとして我々の会話システムを用いて収集した対話データを分析し設計した発話意図タグ付きコーパス [横山 18] を使用し，提案手法の有効性を示す．

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. 章で提案する発話意図認識モデルについて説明し，4. 章で発話意図タグ付きコーパスを用いた実験結果を報告する．

連絡先: 早稲田大学 理工学術院 知覚情報システム研究室
〒 162-0042 東京都新宿区早稲田町 27
E-mail: takatsu@pcl.cs.waseda.ac.jp

2. 関連研究

従来，発話意図の認識・理解は韻律情報を用いて行なわれてきた．例えば，[藤江 03] 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ている．[Ando 15] 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ている．[Nisimura 06] らは，収集した子供の音声に対して喜んでいるか嫌悪を示しているかの推定を行う際，F0 値，パワーから求められる 16 次元の特徴量の中から，因子分析により寄与の高い因子を選択することで，より識別率の高い特徴量を使用している．[林 14] らは，発話タグ（疑問文，平叙文，相槌，同意，笑い）の推定を行う際，F0 値やパワー，話速などの 23 次元の音響特徴量を使用している．

これらの研究は，韻律情報を詳細に観察・分析することで意図理解のための特徴パラメータを設計している．しかしながら，人手で設計可能な特徴パラメータは，推定対象の発話意図の違いが文章や音声を観察することで見いだせる場合に限られる．つまり，観察・分析だけでは容易に特徴が見つけられないような微妙なニュアンスの意図については特徴量の設計が困難になる．加えて，発話意図ごとに特徴パラメータの設計が必要となる．

本研究では，発話意図理解の韻律特徴量をニューラルネットワークを用いて自動で抽出する枠組みを提案する．具体的には，音響情報としてスペクトログラムを CNN を含む AutoEncoder に入力し，その中間層出力を韻律特徴量として用いる．これにより，発話意図の種類に依らず同一の仕組みで韻律特徴量の獲得が可能となる．

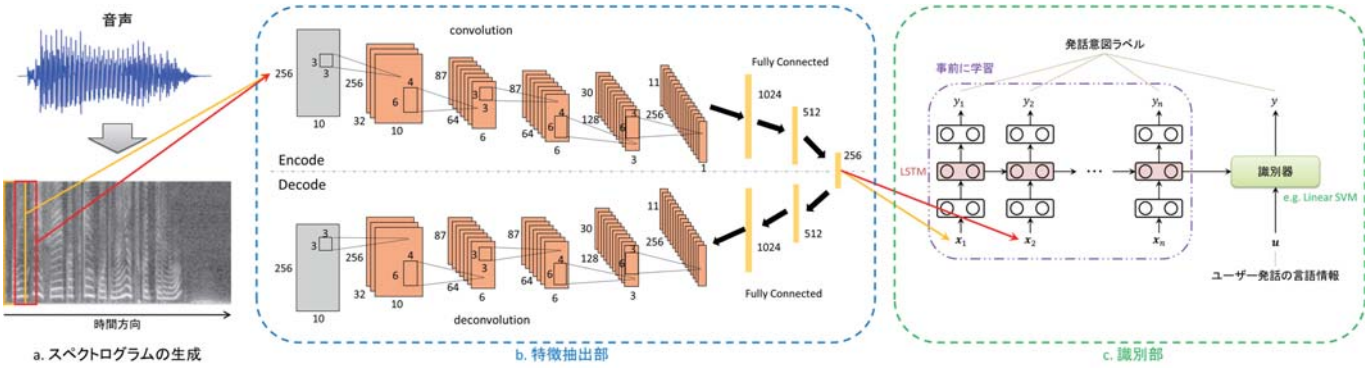

図 1: 発話意図認識モデル

3. 発話意図認識モデル

時間方向に可変な音声の入力を短い時間幅で逐次処理し，その都度過去の履歴を考慮して現時点で予測される発話意図ラベル（e.g. 質問，待機要求，反復要求）を出力するモデルを構築した．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を出力するが，音声認識結果が得られた段階で，LSTM の最終層と音声認識結果から得られる発話の言語特徴量を識別器に与え，最終的な発話意図ラベルの推定結果を得る．モデルの全体像を図 1 に示す．以下，特徴抽出部と識別部について説明する．

3.1 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する．

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用いて発話意図の識別を行った．

3.2 識別部の設計

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．さらに，音声認識結果が得られた段階で，LSTM の最終層と音声認識結果から得られる発話の言語特徴量を用いて再度識別を行うことで，発話内容に対しても頑健なモデルを構築した（図 1.c）．

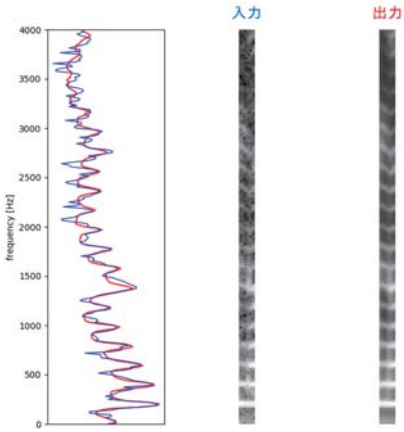

図 2: CNN-AutoEncoder の出力例（左:各スペクトログラムの 5 番目のスペクトルの周波数強度（青：入力，赤：出力），中：入力したスペクトログラム，右：出力したスペクトログラム）

4. 発話意図識別実験

4.1 特徴抽出部の学習と結果

特徴抽出部ではスペクトルの周期性を表すスペクトルパターンの情報を圧縮する．図 1.b の CNN-AutoEncder でスペクトログラムを復元するように学習を行うことで，その中間層にスペクトログラムの情報を圧縮する．この学習過程で期待することは，CNN におけるフィルターが音声情報，つまりスペクトログラムを扱うのに長けたものに学習されることである．そのため，学習で使うデータはドメインが異なる音声データであってもよい．そこで，本研究では公開されている日本語音声コーパスの中でも大規模な『日本語話し言葉コーパス（CSJ）』[*1] を用いて特徴抽出部の学習を行った．

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．

学習した CNN-AutoEncoder の出力例を図 2 に示す．ここで，中央のスペクトログラムが入力として与えたスペクトログラムで，右側のスペクトログラムが CNN-AutoEncoder が出力したスペクトログラムである．左側のグラフは各スペクトロ

*1 http://pj.ninjal.ac.jp/corpus_center/csj/

表 1: 「質問」「非質問」データセットの統計

	総数	「質問」の数	「非質問」の数
訓練セット	2,000	925	1,075
テストセット	1,000	467	533
開発セット	257	138	119

グラムの 5 番目のスペクトルの周波数強度を表している．これらの結果からモデルが入力スペクトログラムの模様パターンを復元できていることが分かった．

4.2 識別部の学習と結果:発話意図「質問」の識別

横山らが構築したデータセット [横山 18] のうち，ラベラー間の一致率が高くデータ数が多い発話意図「質問」に関して識別実験を行った．

4.2.1 発話意図タグ付きコーパス

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用いた [横山 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 18] 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 人のラベラーが発話意図ラベルを付与した．ここでは，その中でもラベラー間の一致率が高くデータ数が多い発話意図「質問」に関して識別実験を行った．本実験で用いた「質問」データセットの統計を表 1 に示す．ここで，「質問」以外の発話意図に分類された発話を「非質問」とする．

4.2.2 LSTM の学習と識別結果

まず，4.1 節で学習した CNN-AutoEncoder の中間層の値（256 次元）を入力として LSTM の学習を行った．ここで，CNN-AutoEncoder への入力には，100ms のスペクトログラムを 50ms ずつシフトさせながら与えた．つまり，現時刻の入力と次の時刻の入力で 50ms のオーバーラップが存在する．これは，LSTM に時間変化の情報を陽に学習させるためである．モデルのパラメータは，入力層のユニット数を 100，中間層のユニット数を 50, 出力層のユニット数を 50 に設定した．

LSTM の最終層の出力ラベルと正解ラベルの一致率を求めたところ，Accuracy は 0.896,「質問」を「質問」と識別する精度は 0.878,「非質問」を「非質問」と識別する精度は 0.912 であった．エラー分析を行ったところ，「なにそれ」や「どうして」「誰が」など，言語表記を見れば「質問」と判別できるような発話を「非質問」と判定してしまう事例が多々見られた．このような誤りを解消すべく，次節では発話の言語情報も組み合わせた発話意図の識別について検討した．

4.2.3 韻律情報と言語情報を用いた識別器の学習と識別結果

本実験では以下の観点について検証を行った．

- 有効な識別器
  scikit-learn[*2] の識別器を一通り比較し，有効な識別器について調査する．各識別器のパラメータには初期値を用いた．

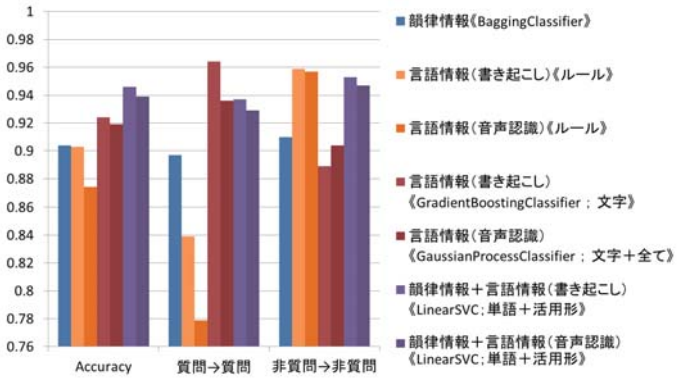

図 3: 発話意図「質問」の識別性能

表 2: 識別誤りの例

正解	出力	書き起こし	音声認識結果
質問	非質問	えっ	
		てなに	
		え何年	奈良根
		何ドルに	マンドリン
非質問	質問	低調	
		おー	
		んなんじゅ	なん j
		の	の

- 言語情報として有効な素性の組み合わせ
  言語情報には，基本素性としてユーザー発話の文字または単語の Bag-of-Words（BoW）を使用し，追加素性として JUMAN（Ver.7.01）[*3] を適用して得られる形態素情報 { 品詞（大分類と細分類）, カテゴリ, ドメイン, 活用形, 活用型 } を使用した．これらの素性に関して有効な組み合わせについて調査する．
- 言語情報のみ，韻律情報のみ，韻律情報と言語情報の 3 通りで識別器を学習したときの比較
  識別器を言語情報のみを用いて学習した場合と韻律情報のみを用いて学習した場合，韻律情報と言語情報を組み合わせて学習した場合で性能を比較し，各特徴量の有効性を確認する．韻律情報には 4.2.2 節で学習した LSTM の最後の隠れ層の値を用いた．
- 発話の表記として人手で書き起こした結果を用いたときと音声認識結果を用いたときの比較
  人手で書き起こした結果から抽出した言語特徴量を用いた場合と音声認識結果から抽出した言語特徴量を用いた場合でどの程度性能に差が見られるか確認する．音声認識には Google Cloud Speech API[*4] を使用した．
- ユーザー発話の言語表記に関してルールで識別したときと機械学習で識別したときの比較
  ベースラインとして基本的なルールでどの程度の識別性能を実現できるか確認する．実際に用いたルールは付録 A に示す．これらのルールは訓練セットの書き起こしデータに基づいて設計した．

*2 http://scikit-learn.org/stable/

*3 http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

*4 https://cloud.google.com/speech/?hl=ja

実験結果を図 3 に示す．まず，言語情報のみを用いてルールで識別した場合と機械学習で識別した場合について比較すると，ルールによる識別では「質問」を「質問」と識別する性能が低いことが分かった．次に韻律情報のみを用いたときと言語情報のみを用いたときの結果を比較すると，言語情報のみ（機械学習）の方が高い性能を示した．

最も良い結果を示したのは，韻律情報と言語情報（単語の表層形と活用形）を特徴量に用いて線形カーネルの SVM を学習し識別を行ったときであった．ここで，書き起こしではなく音声認識結果を用いたときの Accuracy の劣化は 0.7%にとどまっていた．誤り例を表 2 に示す．書き起こしを用いたときは正しかったが，音声認識結果を用いたときに誤った事例について確認したところ，音声認識できなかったもの（e.g. 「えっ」→「」）や音声認識誤りで疑問詞が消えてしまったもの（e.g. 「え何年」→「宇奈根」,「え何」→「えなり」）が誤りとして多く見られた．また，言語的に誤りを含むものだけでなく，言い淀みなどで音声的に不完全な発話も誤りやすいことが分かった．

5. おわりに

情報伝達効率の高い会話インタラクションを実現するためには，ユーザーからの多様なフィードバックを理解する必要がある．しかしながら，全てのフィードバックが言語的に明示された形で行われるとは限らない．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，本研究では言語情報と韻律情報を組み合わせた発話意図認識手法を提案した．提案手法の特徴は，従来意図の種類ごとに観察に基づいて行われてきた特徴パラメータの設計を深層学習を用いることで，寄与率の高い韻律特徴量を自動で抽出しているところにある．

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 18] を用いて収集した対話データを分析し設計した発話意図タグ付きコーパス [横山 18] を用いて評価実験を行ったところ，発話意図「質問」の識別に関して 94%程度の識別率を達成した．今後は，他の発話意図に関してもデータを増やし，識別実験を行う．

参考文献

[高津 18] 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 (2018)

[横山 18] 横山勝矢, 高津弘明, 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分類とデータベースの構築, 第 32 回人工知能学会全国大会論文集 (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)

[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)

[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)

[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)

[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)

[岩田 12] 岩田和彦, 小林哲則: 終助詞とその音調とによって聞き手に伝わる発話意図の分析, 電子情報通信学会技術研究報告, Vol. 112, No. 281, pp. 31-36 (2012)

[Schuller 04] Schuller, B., Rigoll, G., and Lang, M.: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 577-580 (2004)

[Han 14] Han, K., Yu, C., and Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine, in *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, pp. 223-227 (2014)

[Ghosh 16] Ghosh, S., Laksana, E., Morency, L.P., Scherer, S.: Representation learning for speech emotion recognition, in *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, pp. 3603-3607 (2016)

[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)

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks, in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097-1105 (2012)

A 4.2.3 節の実験で用いた識別ルール

- 以下を含めば「質問」
  「何」「なに」「なーに」「どこ」「だれ」「誰」「なんで」「なんの」「なん日」「どう」「いつ」「いくら」「なぜ」「どんな」「どの」「どれ」「例えば」「たとえば」「なんて」
- 以下で終われば「質問」
  「か」「かね」「かな」「の」「って」「っけ」「は」「で」「が」「に」「ってこと」
- 以下に一致していれば「質問」
  「え」「ん」「えっ」
- 以下を含めば「非質問」
  「そっか」「そうか」「なんでもない」「何でもない」