2H3-OS-35a-3

インタビュー対話における 重要シーン推定のための言語・非言語特徴量の分析

Analysis of Verbal and Non-verbal Features for Estimating Important Scenes in Interview Dialogue

石原 卓弥 *1 Takuya Ishihara　長澤 史記 *1 Fuminori Nagasawa　岡田 将吾 *1 Shogo Okada　新田 克己 *1 Katsumi Nitta

*1東京工業大学 情報理工学院 情報工学系
Department of Computer Science, School of Computing, Tokyo Institute of Technology

Abstract: The goal of this research is to summarize the dialog contents which are collected through interaction with a robot interviewer by combining speech recognition techniques and multimodal recognition techniques. The key technique is to identify important statements in the dialog contents. We focus on using nonverbal behaviors including prosody, gesture and posture for the prediction of the important statements that the user would like to emphasize. As a first step, this study investigates the relationship between multimodal features and important statements (or scene) using statistical tests. The result shows that features : "Number of Nouns", "Number of Verbs", "Voice Pitch", "Voice Energy", "MFCC", "Head Movement" and "Shoulder Movement" have significant difference between important one and non-important one.

1. はじめに

近年までに，対話システムに関する研究 [中野 15] は盛んに行われており，電話応答システム，カーナビゲーションといった多くの事例で応用されている．一方で，対話において交わされる，会話・対話内容（コンテンツ）には様々な知識が含まれており，この知識を管理する技術は会話情報学 [Nishida 08] の分野で研究されている．本研究では，対話システムを対話コンテンツ・知識を収集するために用いることで，個人の発信したい内容，聴衆が聞きたい内容を，対話を通じて獲得し，対話内容を要約・提示する技術に焦点を当てる．対話内容の収集・活用が行われる事例の一つとして，インタビューが挙げられる．本研究ではヒューマノイドロボットを用いたインタビューシステムを構築し，インタビュー対話内容を音声認識・自然言語処理により獲得し，重要なインタビュー内容を推定し要約・活用することを目標とする．提案するロボットシステムの要素技術は,（1）インタビュー相手（ユーザ）の発話態度に応じて，インタビュー方法を変えることで，ユーザに多くの事を語らせるインタビュー戦略の実装と（2）インタビューで得られた対話コンテンツにおける重要箇所の推定・要約である．本論文では（2）の対話コンテンツの要約のための検討を行い，その結果を報告する．対話コンテンツ要約のために，インタビュー内容の内，要約に残すべき重要箇所（重要発言の）にアノテーションを付与し，重要箇所の推定に有用な特徴量を分析する．音声テキスト要約は主に，音声認識の分野で盛んに行われてきた [Tur 11]．本研究では，ロボットとユーザの対面対話を対象としているため，音声認識結果より得られたテキストデータだけでなく，対話時のユーザの韻律情報や上半身の動作情報を含め，重要箇所の推定に有効な特徴量の分析を行う．

2. 関連研究

二瓶ら [二瓶 17][Nihei 14] は，議論参加者の注視行動，頭部動作，韻律情報といった非言語情報に着目している．二瓶らは [二瓶 17] で，発言の重要度を F 値 0.7, 再現率約 0.7 の性能で検出できることを報告している．また，議論データについて，時間を 45 %に短縮した要約の生成が可能になったことを報告している．しかしながら，これらの研究に用いられている情報は非言語情報のみであり，言語情報は用いられていない．

岡田ら [岡田 16] は，グループディスカッションにおけるコミュニケーション能力の推定のモデルの構築・評価に取り組んだ．人事経験者によるアノテーションを行い，そこからグループディスカッションの参加者の能力を推定するモデルの構築を行っている．この研究は，コミュニケーション能力の高低を 90 %以上の精度で分類出来ることを示した．

これらの研究が対象とするデータは，本研究で行うロボットと人とのコミュニケーションではなく，人と人とがコミュニケーションするシチュエーションであるため，本研究との差異が存在する．

3. インタビュー概要

3.1 インタビュー環境と質問内容

インタビュー環境を図 1 に示す．インタビュー対象者 (以下，対象者と呼ぶ) は Pepper との一対一の環境でインタビューに答える．対象者は博士前後期課程を含む研究者である．研究内容や，私生活について問う質問セットをあらかじめ用意し，それを Pepper が発することによりインタビューが進行する．この際，指向性マイクを対象者の頭部に取り付ける．また，kinect を設置し，Pepper と対象者の動作を検出し，PC 側でその動作を保存する．加えて，対象者の正面に web カメラを設置し，対象者の表情や身体の動きと共に，俯瞰音声を取得する．

3.2 インタビューの流れと取得データの処理

インタビューとその処理の流れを図 2 に示す．インタビューは Pepper が質問を発し，対象者がそれに答える一方向型の対話で進められる．この際，質問のタイミングは実験管理者 (以下，管理者と呼ぶ) によって制御される．管理者は適切なタイミングで Pepper に対して発話の命令を送信する．

インタビューにより，kinect から対象者の行動データ，指向性マイクからの音声データ，ウェブカメラからの正面映像・俯瞰音声データが取得される．インタビュー後に，アノテータによって，重要シーンの開始時刻と終了時刻までの区間を指定し

連絡先: 石原卓弥, 東京工業大学情報理工学院情報工学系, 神奈川県横浜市緑区長津田町 4259 J2-53, TEL: 045-924-5214

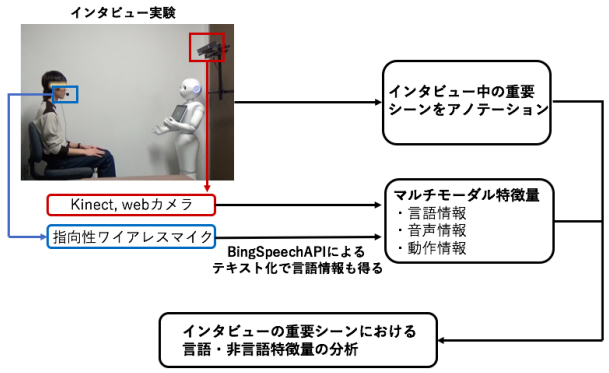

図 1: インタビュー環境と有効な特徴量の分析を行う枠組み

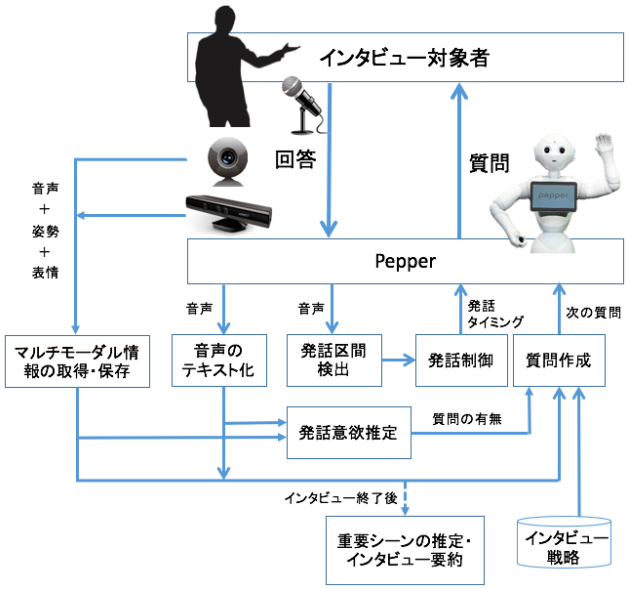

図 2: インタビューの流れ

たアノテーションが行われる．マルチモーダルデータから，重要シーンを推定するモデルが構築される．構築されたモデルを用いることで，新規のインタビューデータから重要シーンの認識が自動で行われる予定である．取得したマルチモーダルなデータを用いて，重要シーンを推定するための予備検討として，有効な特徴量の分析を行う．インタビュー後にアノテータが重要だと感じたシーンの開始時刻と終了時刻をアノテーションする．アノテーションされたシーンについて，言語特徴量として音声認識結果における名詞・動詞数を，非言語特徴量として音声区間における韻律情報，また対象者の上半身の動作情報を用い，重要シーンにおける特徴量の分析を行う．

4. インタビューコーパス

4.1 データセット

3 章で述べたインタビュー環境，質問内容で本研究室の 3 人の学生，2 人の教員に対してインタビューを行った．対象者はインタビュアーである Pepper と対面する場所に座った状態でインタビューに臨んだ．対象者となった 5 人に対し，全て同じ質問でインタビューを進めた．より詳しいインタビュー実験についての説明は，長澤ら [長澤 17] の論文中で述べる．

インタビュー実験の結果，各対象者によってインタビューの収録時間に大きな差が出ることが判明した．実際の収録時間は，それぞれ 5 分 35 秒，8 分 54 秒，9 分 0 秒，13 分 8 秒，15 分 47 秒である．収録時間の短いデータでは，ほぼ全ての質問に対して一言で回答していることがわかった．本研究の最終目的はインタビュー時における重要シーンを推定し，そこからインタビュー内容の要約を行うことであるので，そのようなデータに対して重要シーンのアノテーションを行うと，全発話中のほとんどがアノテーション区間に含まれてしまう．したがって，本研究では，収録時間が長い 13 分 8 秒 (以下，A 実験と呼ぶ)，15 分 47 秒 (以下，B 実験と呼ぶ) のインタビュー実験におけるデータを使用する．

4.2 重要シーンの定義とアノテーション

本研究での重要シーンの定義について説明する．本研究では，重要シーンを「インタビューを振り返るときに必要なシーン」と定義する．また，質問への回答に実質的に同じ内容を話し続けていた場合，そういったシーンを重要シーンであると判断する．そういったシーンは対象者が伝えたかった内容である可能性が高いためである．

アノテーションは，本論文の著者 1 名で行った．重要シーンの定義に従い，そのシーンの開始秒，終了秒をアノテーションした．なお，同一内容を話し続けていた場合，その内最も重要である 1 発話だけをアノテーションし，他を棄却した．

4.3 マルチモーダル特徴量

4.3.1 マルチモーダル情報の抽出

本研究では，Julius[河原 05] を用いて対象者の発話ターン内の発話区間を検出し，それに従って指向性マイクから取得した音声データを分割した．また，分割した各音声データ (以下，発話断片と呼ぶ) について，BingSpeechAPI を用いて音声認識を行い，テキスト化した．

アノテーション区間と時間的に 50 %以上重なる発話断片を重要発話断片とし，それ以外を非重要発話断片とする．なお，アノテーション区間が長く，複数の発話断片が 1 つのアノテーションに対応している場合，その全ての発話断片を重要発話断片とみなす．この指針で発話断片を重要発話断片と非重要発話断片に分類した結果，A 実験ではそれぞれ 24 個，46 個，B 実験では 20 個，96 個となった．どちらも非重要発話断片のほうが多い結果となっている．

4.3.2 言語特徴量への変換

言語特徴量として，音声認識結果より得られるテキスト情報から抽出した名詞数と動詞数を採用した．各発話断片の音声認識結果に対して Mecab を用いて形態素解析を行い，名詞数，動詞数を求めた．1 つのアノテーション区間に複数の発話断片が対応している場合，それぞれの発話断片の名詞数，動詞数を足し合わせたものを，そのアノテーション区間における名詞数，動詞数とした．

4.3.3 動作特徴量への変換

動作特徴量として，頭部，肩部の動作の変化量の絶対値の平均値を求めた．また，その平均値を用いて，各部位の変化量の絶対値の分散値を求めた．

4.3.4 音声特徴量への変換

各発話断片について，Speech feature extraction code[*1] を用いてピッチ，エネルギー，MFCC を求めた．導出したそれらの値について，それぞれの最大値，最小値，平均値を特徴量として用いた．

*1 Speech feature extraction code, http://groupmedia.mit.edu/data.php

5. 分析

5.1 言語特徴量

各アノテーション区間における名詞数，動詞数と，一度も重要発話断片だと判断されなかった発話断片のそれぞれの特徴量について，ウィルコクソンの順位和検定を用いて分析した．検定は，A 実験，B 実験で得られたデータをマージした状態で行われた．結果を表 1 の言語特徴量の列に示す．名詞数，動詞数での結果はどちらも，p 値が 0.01 よりも低い結果となっている．また，重要発話断片に含まれる名詞数，動詞数はそれぞれ平均 9.67 個，4.88 個であった．非重要発話断片における名詞数，動詞数はそれぞれ平均 7.26 個，3.54 個であったので，いずれの結果においても重要発話断片のほうが名詞数，動詞数ともに多くなっていることがわかった．これは各アノテーション区間における名詞数，動詞数のデータと，各非重要発話断片における名詞数，動詞数のデータの間には，有意な差があることを示している．これは，非重要発話断片には，フィラーしか入っていないデータなど，名詞数，動詞数の観点ではどちらも明らかに 0 になってしまうデータが，重要発話断片に比べて多く存在していることが要因として考えられる．

表 1: 言語特徴量/動作特徴量でのウィルコクソンの順位和検定の結果

	言語特徴量	動作特徴量
5 %以下		
1 %以下	名詞数 動詞数	頭部動作変化量 (X, Y, Z 軸) 肩部動作変化量 (右手，左手)

表 2: 音声特徴量でのウィルコクソンの順位和検定の結果

	音声特徴量
5 %以下	2, 6, 7, 11 次元最大 MFCC
	3, 6, 7, 8, 11 次元最小 MFCC
	4, 7, 9, 10, 12 次元平均 MFCC
1 %以下	最小ピッチ
	平均ピッチ
	最大エネルギー
	平均エネルギー
	1, 3, 4, 5, 8, 13 次元最大 MFCC
	7, 10, 12 次元最小 MFCC
	1, 3 次元平均 MFCC

5.2 動作特徴量

重要発話断片における頭部・肩部動作変化量の絶対値の平均値，分散値と，非重要発話断片におけるそれぞれの特徴量について，ウィルコクソンの順位和検定を用いて分析した．検定は，A 実験，B 実験で得られたデータをマージした状態で行われた．結果を表 1 の動作特徴量の列に示す．表 1 の通り，平均値，分散値の両方の特徴量において，有意な差があることが確認できた．[Okada 13] によると，人は説明を行う場合に，身振り手振りを含めた様々な非言語情報を表出することが示されている．今回の結果はそれを反映しているといえる．

5.3 音声特徴量

重要発話断片におけるピッチ，エネルギー，MFCC と，非重要発話断片におけるそれぞれの特徴量について，ウィルコクソンの順位和検定を用いて分析した．検定は，A 実験，B 実験で得られたデータをマージした状態で行われた．結果を表 2 に示す．表 2 の通り，最大ピッチと最小エネルギーを除いた全ての特徴量において有意な差があることを確認できた．MFCC については，1 から 13 次元の全てで最大値，最小値，平均値のいずれかが有効であることがわかった．特に 3 次元の MFCC については最大値，最小値，平均値のいずれも有効であることがわかる．

6. 結論

本研究では，インタビュー時に取得されたデータから重要シーンを自動的に推定し，そこからインタビューの要約を行うための，言語・非言語特徴量の分析を行った．分析の結果，重要シーンと非重要シーンでは，言語特徴量，非言語特徴量のいずれにおいても，それぞれ有意な差があることが確認された．

今後の課題として，機械学習によるモデルの構築が挙げられる．本研究による成果をもとに，モデルを構築することで，人の手に依らずに重要シーンを判別するモデルが作成できる．また，インタビュー実験の課題として，対象者によってはインタビューの収録時間が短くなってしまい，本研究に使用出来るデータが少なくなったことが挙げられる．実験前にロボットと人との間にアイスブレークを入れ，実験参加への緊張をほぐしたり，ロボットとの会話に慣れてもらう工夫を行うことで，この状況の改善が見込まれると考えられる．また，本実験では研究内容に特化した質問が質問リストの大半を占めているため，私生活についてなど，平易な質問の割合を増やすことで，対象者によって，答えやすい設定に変更することも検討する．

参考文献

[中野 15] 中野幹生, 駒谷和範, 船越孝太郎, 中野有紀子, ”対話システム (自然言語処理シリーズ)”, コロナ社, 2015

[Tur 11] Gokhan Tur, Renato De Mori, ”SPOKEN LANGUAGE Understanding: Systems for Extracting Semantic Information from Speech”, pp.357, WILEY, 2011.

[Nishida 08] Nishida, T., ”Conversational informatics: an engineering approach”, Vol. 9, Wiley. com, 2008

[二瓶 17] 二瓶芙巳雄, 高瀬裕, 中野有紀子, ”非言語情報に基づくグループ議論における重要発言の推定 -グループ議論の要約生成に向けて-”, 電子情報通信学会論文誌 A, pp34-44, 2017.

[Nihei 14] Fumio Nihei, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Huang, Shogo Okada, ”Predicting Influential Statements in Group Discussions using Speech and Head Motion Information”, ICMI, 2014.

[岡田 16] 岡田将吾，松儀良広，中野有紀子，林佑樹，黄宏軒，高瀬裕，新田克己, ”マルチモーダル情報に基づくグループ会話におけるコミュニケーション能力の推定”, 人工知能学会論文誌 Vol. 31 No.6, 2016

[河原 05] 河原達也, 李晃伸,”連続音声認識ソフトウエア Julius”, 人工知能学会誌, Vol.20, No.1, pp.41–49, 2005.

[長澤 17] 長澤史記, 石原卓弥, 岡田将吾, 新田克己, ”ユーザの態度推定に基づき適応的なインタビューを行うロボット対話システム構築への一検討”, 人工知能学会全国大会, 2017.

[Okada 13] Okada, Shogo and Bono, Mayumi and Takanashi, Katsuya and Sumi, Yasuyuki and Nitta, Katsumi, ”Context-based Conversational Hand Gesture Classification in Narrative Interaction”, Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp303-310, 2013