音響的, 言語的特徴及び対話行為を用いた話題に対する興味推定*

☆吉田 理貴, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲 (奈良先端大)

1 はじめに

対話システムを構築する上で, 話題に対するユーザの興味を推定する事は, 話を盛り上げる, ユーザに対して何かしらの推薦を適切に行う [1] など, ユーザとシステムが良好な関係を築くために重要な要素技術の一つである. このような興味推定を行う手法として, 語の共起情報 [2] や聴衆の反応 [3] などの情報に着目した手法が提案されている. さらに, 画像, 音響, 言語などの情報を組み合わせた統合的な枠組みも提案されている [4]. 本稿では, カメラ販売対話において, ユーザである顧客の興味推定に取り組む. 音響・言語情報に加えて, さらに対話行為と話者の特性を表す素性を導入し, それぞれの情報の相乗効果について調査する.

2 線形分類による話題に対する興味推定

様々な情報を興味推定に取入れる枠組みとして線形分類を用いる. 興味推定に用いる素性ベクトル $\boldsymbol{x}$ と重みベクトル $\boldsymbol{w}$ を基に, 式 (1) の分類関数を定義する.

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} \tag{1}$$

$f(\boldsymbol{x}) \geq 0$ ならば正クラス (興味あり) に, $f(\boldsymbol{x}) < 0$ ならば負クラス (興味なし, もしくはわからない) に分類する. 本稿では, 発話単位で興味推定を行う. 重みベクトルの学習には SVM を用いる.

3 興味推定のための素性

上記の分類器の入力として音響的特徴量, 言語的特徴量, 対話行為, 話者の特性を考慮した素性を用いる. 発話から素性ベクトルを抽出する流れを図 1 で示す.

3.1 音響的特徴量

音響的特徴量として, 基本周波数 (F_0) と対数パワーを用いる. 各発話に対して, 対数 F_0 の平均と分散, および, 対数パワーの平均と分散を算出する. 話者の違いがもたらす影響を取り除くため, これらの値は発話者ごとに Z スコア（平均 0, 分散 1）に正規化する.

3.2 言語的特徴量

言語的特徴量として, 単語頻度ベクトルを用いる. 書き起こされた発話文に対して, 形態素解析を行い, 各形態素の頻度を計算し, 単語頻度ベクトルを求める. また, 発話の長さが興味状態に関与する可能性があるため, 発話内の形態素数も素性として加える, なお, 本稿では, 書き起こしは人手で行う.

3.3 対話行為

対話行為を表す指標として, ISO 国際標準規格 (ISO/DIS 24617-2)[5] で規定された対話行為タグを用いる. 各発話に対して, 必ず一つのタグのみが付与される一般目的機能 (General purpose functions: GPF) タグを付与する. 例えば, GPF タグの一つである “Request” は, 話し手が聞き手に対して, 述べられた作法の実現要求を意図する際の発話に付与される. このようなタグが全部で 25 種類存在し, 本稿では, その内 17 種類が対話データに付与された. また, GPF タグでは, 相づちは合意や確認などに分類されるが, 自分が相手の発話を聞いていることを示す相づちや発話の間を保つための相づちなどは, 区別されずに別のタグとひとまとめにされる. しかし, 相づちは対話の興味状態推定で重要な要素の一つであると考え, 相づちタグを独自に定義する. 本稿では, 人手で GPF タグを付与する.

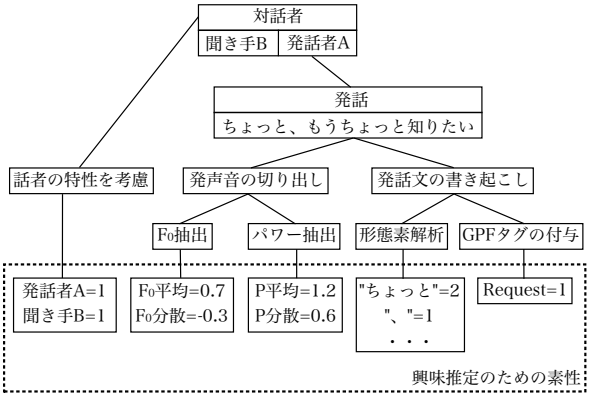

Fig. 1 対話からの素性ベクトル抽出例

Table 1 実験に用いたデータセット

対話の組		ラベル		発話数
客	店員	興味あり	その他	
客 1	店員 1	46	33	79
客 1	店員 2	34	17	51
客 1	店員 3	5	21	26
客 2	店員 1	58	79	137
客 2	店員 2	53	79	132
総数		196	229	425

3.4 話者の特性を考慮した素性

音響的特徴量や言語的特徴量, 対話行為は各発話者の性格などによって, 意図する意味合いが異なることは容易に想像できる. しかし, 発話データだけでは, これらの意味合いの違いまでを考慮することは困難である. そこで, 発話者に応じて意味合いが異なるような発話にも対応できるように, 誰が誰に対して発話をしているのかという話者情報も素性として用いる.

4 興味推定の判定実験

4.1 実験設定

本研究の実験を行うために, 平岡らにより分析されたカメラ販売コーパス [6] の店員役 3 名と客役 2 名のコーパスを利用した. 客役 2 名が自身の発話データに対して, 2 節で述べた興味状態ラベルを付与した. F_0 抽出には STRAIGHT[7] を用いた. パワーの抽出には, Snack[8] を用いた. 形態素解析エンジンには MeCab[9] を使用した. 実験に用いたデータセットを表 1 に示す. 興味状態の推定率の評価には, 424 発話を学習用のデータにし, 残り 1 発話をテスト用のデータにする 425 分割交差検定を用いた.

*Estimation of interest in topics using acoustic, linguistic, and dialogue act features. by YOSHIDA, Riki, NEUBIG, Graham, SAKTI, Sakriani, TODA, Tomoki, NAKAMURA, Satoshi (NAIST)

Table 2　各特徴量による話題に対する興味推定率

特徴量	興味推定率	
	話者情報なし	話者情報あり
なし	53.88%	60.94%
音響	60.47%	65.88%
言語	66.35%	66.12%
対話	62.59%	68.47%
音響+言語	65.65%	65.88%
音響+対話	64.00%	68.47%
言語+対話	67.06%	66.82%
音響+言語+対話	67.06%	65.88%
F_0+言語+対話	68.24%	66.12%
人間による推定	68.47%	

また, 本研究では音響, 言語の枠組みに対話行為と話者の特性を追加し興味推定を行ったが, 発話データに対する GPF タグの付与は人手で行ったため, GPF タグの付与誤りが少ない. しかし, システムによる興味推定をリアルタイムで行うことを想定すると GPF タグをシステムが予測する必要があり, これには人であれば付与しない GPF タグが付与されると考えられる. そこで, 人手で付与した GPF タグを, 17 種類の GPF タグと相づちタグの合計 18 種類のタグにランダムに置き換えることで話題に対する興味推定率がどのように変化するのかを調べた.

4.2　実験結果と考察

表 2 に各特徴量を用いた際の発話者の話題に対する興味推定率を示す. “人間による推定” とは, 発話者と別の人間に同じ対話データを聞いてもらい, 興味状態の推定を行ってもらった数値である. 話者情報なしの場合に, 最も推定率が高かった素性の組み合わせが, F_0 を単体で利用して, 言語, 対話と組み合わせたものであったため, この結果も表 2 に載せた. Schuller ら [4] は, 今回使用した実験セットとは違う枠組みではあるが, 画像, 音響, 言語などの情報を組み合わせ, SVM による学習で 63%の精度で興味推定を行えたと報告しており, 今回の研究で同程度の推定精度が出せている. また, システムが常に話者情報を利用できるとは限らないが, 話者情報がなくとも特徴量を組み合わせる事で 68.24%と人間による推定率に近い推定精度を出せる事がわかった. 次にフィッシャーの正確確率検定の 1%水準を用いて手法間の違いを比較した. 話者情報なしの場合で手法間の比較を行うと, 特徴量なしと音響以外の他の特徴量の間に有意差が確認された. 話者情報ありの場合は特徴量なしと他の特徴量の間に有意差は見られなかったが, 音響+対話の組み合わせで p 値が 0.03 と推定率の向上傾向が見られた. このことから今回使用した特徴量は興味状態を推定する上で有効であると言える. 話者情報の違いによる手法間の違いの比較も行ったが, 有意差を確認することはできなかった. 今回の実験では話者情報なしの場合に特徴量を組み合わせる事で推定率が向上する傾向が見られたが, 話者情報ありの場合には特徴量を組み合わせると推定率が低下する事が多かった. これは使用したデータが小さいために適切な学習が行えていないと考えられる.

4.3　GPF タグの付与誤りによる興味推定率の推移

図 2 は横軸が GPF タグの変換率であり, 100% は全ての GPF タグを別のタグに入れ替えた事を表す. これより, 25% 程度の付与誤りでも推定精度が低下し, 対話行為の認識精度の重要性がわかった.

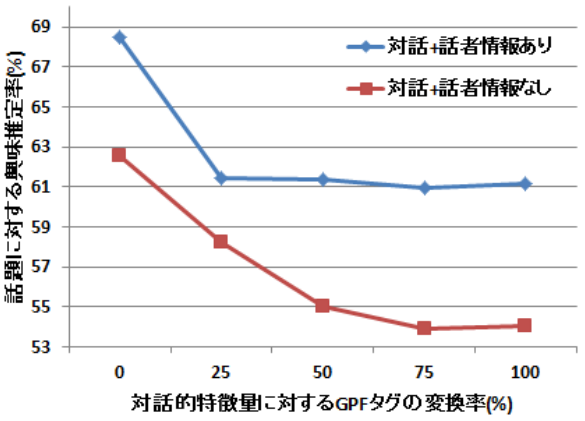

Fig. 2　ノイズデータ挿入による興味推定率の推移

5　おわりに

本稿では, 従来法の枠組みの中で, さらに対話行為と話者の特性を表す素性を導入し, それぞれの情報の相乗効果について調査することを目的に議論を進めてきた. 結果的に今回使用した特徴量は興味推定率を向上させる事に役立ち, 話者情報なしでも, 特徴量を組み合わせる事で人間に近い推定率を行えることを示した. しかし, 今回調査に用いたデータは小さく, 先行研究 [4] では人間による興味推定で 84%の精度が報告されており, これと比較すると今回の結果では推定精度も低い. 今後は, さらなる推定精度の向上を目標にデータと手法の改良と改善を目指す.

参考文献

[1] 河原達也, 川嶋宏彰, 平山高嗣, 松山隆司. 対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジェ. 情報処理, Vol. 49, No. 8, pp. 912–918, Aug 2008.

[2] 稲葉通将, 鳥海不二夫, 石井健一郎. 語の共起情報を用いた対話における盛り上がりの自動判定. 電子情報通信学会論文誌. D, 情報・システム, Vol. 94, No. 1, pp. 59–67, Jan 2011.

[3] 河原達也, 須見康平, 緒方淳, 後藤真孝. 音声会話コンテンツにおける聴衆の反応に基づく音響イベントとホットスポットの検出. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3363–3373, Dec 2011.

[4] Björn Schuller, Ronald Müller, Florian Eyben, Jürgen Gast, Benedikt Hörnler, Martin Wöllmer, Gerhard Rigoll, Anja Höthker, and Hitoshi Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, Vol. 27, No. 12, pp. 1760–1774, 2009.

[5] Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pp. 430–437, 2012.

[6] 平岡拓也, Sakriani Sakti, Graham Neubig, 戸田智基, 中村哲. 説得対話システム構築のための対話コーパス分析. 日本音響学会 2013 年春季研究発表会 (ASJ), 東京, 3 2013.

[7] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, Vol. 27, No. 3, pp. 187–207, 1999.

[8] K Sjolander. Tcl/tk snack toolkit, 2004.

[9] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析. 情報処理学会自然言語処理研究会 SIGNL-161, Vol. 47, pp. 89–96, 2004.