The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

LSTMを用いた本心でない発話の自動検出

Automatic Detection of Insincere Utterances with LSTM

見尾　和哉[*1]　石野　亜耶[*2]　目良　和也[*1]　竹澤　寿幸[*1]
Kazuya MIO　Aya ISHINO　Kazuya MERA　Toshiyuki TAKEZAWA

[*1] 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

[*2] 広島経済大学メディアビジネス学部
Faculty of Media Business, Hiroshima University of Economics

We propose a method of automatic detection of insincere utterances from voice and facial expression. Proposed method utilizes Long short-term memory (LSTM) to consider time series variation of the voice and the facial expression instead of support vector machine (SVM). The experimental results indicated that proposed method could improve recall (0.73) and F-measure (0.65) from SVM baseline.

1. はじめに

近年，人間とコミュニケーションを行う対話システムが盛んにサービス化している．対話システムがユーザとより円滑なコミュニケーションを行うためには，ユーザの感情を理解する必要がある．しかし，表出された感情が常に本心であるとは限らないため，発話が本心か否かというような抑圧された感情まで推定する技術が必要である．

そこで，本研究では，抑圧された感情まで推定するシステムの構築を目的に，本心でない発話を自動検出する手法を提案する．提案手法では，機械学習には LSTM（Long short-term memory）を使用し，特徴量として発話中の音声や表情の情報を利用する．

2. 先行研究

本研究の先行研究として，Uemura ら[Uemura 2017]の研究がある．Uemura らは，発話中の音声や表情から得られる特徴量を使用して，本心でない発話を自動検出する手法を提案している．音声特徴量としては，音声の時系列データから，openSMILE[Eyben 2010]により算出された音量の最大値や最小値などの 384 個の特徴を利用している．表情特徴量としては，発話が終わった瞬間の静止顔画像から，オムロンの OKAO Vision[オムロン]を用いて算出された 5 感情（無，喜，驚，怒，悲）の推定感情の尤度を利用している．このように，Uemura らの手法では，音声，表情ともに事前に定義された特徴量を使用して，SVM によって本心でない発話を検出している．本研究では，より細かい音声や表情の時系列での変化を把握するため，機械学習に LSTM を用いた手法を提案する．

3. LSTMを用いた本心でない発話の検出手法

本研究では，音声と表情の時系列での変化を考慮し，本心でない発話の検出を行うため，LSTM を用いた手法を提案する．LSTM とは，RNN（Recurrent Neural Network）を改良した時系列データを扱うことができるモデルである．

LSTM を用いた本心でない発話の自動検出手法のイメージ図を図 1 に示す．図 1 の縦長の長方形は，LSTM の入出力となる特徴ベクトルを模式的に表したものである．

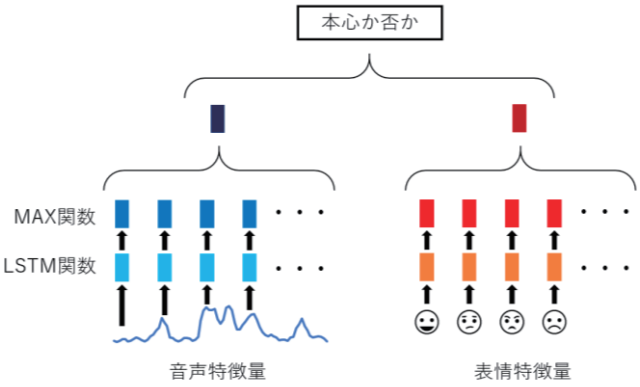


図 1　本心でない発話の自動検出手法のイメージ図

LSTM に与える特徴量について説明する．まず，音声特徴量について説明する．発話中にマイクにより録音したデータから，openSMILE を使って，25msec の窓に対して「音圧」，「基本周波数（F0）」，「自己相関関数から算出される声らしさ」の値を算出する．この窓を 10msec ごとにシフトさせ時系列の音声特徴量を得る．

次に，表情特徴量について説明する．発話中にビデオカメラにより撮影された動画から，OKAO Vision により顔部位の特徴点（左目頭，左目尻，右目頭，右目尻，鼻左，鼻右，口上，口元左，口元右）を検出し，その座標を時系列に並べたデータを表情特徴量として使用する．表情特徴量として，OKAO Vision により推定された感情を利用せず，顔部位の特徴点の座標を利用するのは，Ekman[Ekman 2009]の squelched expression（抑圧された表情）のように，顔の部位ごとの感情表出やその程度を考慮するためである．

提案手法では，音声特徴量$v_{voice,1}$と表情特徴量$v_{face,1}$を，それぞれ時系列に並べて LSTM 関数に入力し，100 次元のベクトルに変換し，次元ごとに MAX 関数を適用し$v_{voice,2}$，$v_{face,2}$を得る．$v_{voice,2}$と$v_{face,2}$を連結後，線形関数を適用し，2 次元のベクトル$v_{voice+face}$へ変換する．最後に，$v_{voice+face}$の中で最も値の大きい次元に対応するラベルを予測ラベルとする．

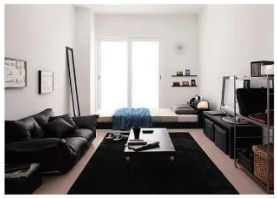
また，LSTM を双方向に拡張した BiLSTM（Bidirectional LSTM）を用いた手法でも同様の実験を行う．

連絡先：目良和也，広島市立大学大学院情報科学研究科，
mera@hiroshima-cu.ac.jp

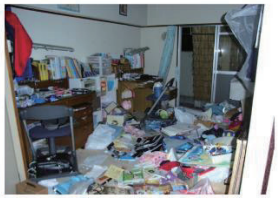
4. 実験

4.1 実験に使用したデータ

実験には，Uemura らの研究で使用したデータを利用した．Uemura らのデータ収集手順を説明する．まず，大学生 10 人に，図 2 に示すような，本心を誘発するであろう画像と，本心を誘発しないであろう画像各 20 枚を，1 枚ずつ見せた．



本心を誘発する画像



本心を誘発しない画像

図 2 データ収集に用いた画像

画像を見て，例え本意でなくても必ず褒めてもらい，その際に音声をマイクで録音，表情をビデオカメラで撮影しておき，後に特徴量として機械学習に使用した．そして発話した直後に，本心であるか否かを回答させた．これを，褒め台詞を指定する「台詞固定」と，自由な言葉で褒めてもらう「台詞自由」の 2 パターンでデータを収集し，実験に使用した．実験に使用したデータを表 1 に示す．

表 1 実験に使用したデータ

パターン	本心でない発話（正例）	本心である発話（負例）	合計
台詞固定	201	166	367
台詞自由	167	202	369

4.2 比較手法

提案手法の有効性を確認するため，以下に示す手法で実験を行った．

- SVM 手法（比較手法）: Uemura らと同様に SVM を利用した手法
- LSTM 手法（提案手法）: LSTM を利用した手法
- BiLSTM 手法（提案手法）: BiLSTM を利用した手法

LSTM 手法と BiLSTM 手法のモデルパラメータの最適化手法には Adam を使用した．隠れ層の数は 2，バッチサイズは 50，エポック数は 100 とした．評価尺度には，精度，再現率，F 値を使用し，5 分割交差検定を行った．

4.3 実験結果

実験結果を表 2 に示す．まず，台詞固定のパターンについて考察する．比較手法である SVM 手法より，提案手法である LSTM 手法は，精度は 0.02 ポイント，再現率は 0.12 ポイント，F 値は 0.07 ポイント改善することができた．また，提案手法である BiLSTM 手法では，LSMT 手法よりもさらに再現率を 0.03 ポイント改善することができた．精度の向上はわずかであったが，再現率は大幅に向上させることができた．これは，比較手法である SVM 手法では，音声，表情ともに事前に定義された特徴量を使用していたが，提案手法では，LSTM を利用することで，より細かい音声や表情の時系列での変化を把握することができたためであると考えらえる．BiLSTM では，双方向の変化を把握することができるため，より再現率が向上したと考えられる．よって，台詞固定のパターンでは，提案手法の有効性が確認できた．

次に，台詞自由のパターンについて考察する．比較手法である SVM 手法より，提案手法である LSTM 手法や BiLSTM 手法では，精度，再現率，F 値ともに低下してしまった．台詞自由のパターンでは，様々な台詞が混在しているため，音声や表情の時系列での変化を細かく把握する手法は，有効でないと考えられる．台詞自由のパターンに対応させるためには，まずは似た特徴を持つ発話をクラスタリングし，そのクラスタごとに提案手法を適用する必要がある．

5. まとめ

本研究では，機械学習として LSTM を使用することで，本心でない発話を自動検出する手法を提案した．実験の結果，台詞固定のパターンにおいては，SVM を用いた比較手法と比較し，提案手法では再現率を 0.12 ポイント向上させることができた．台詞自由のパターンにおける精度の改善が今後の課題である．

謝辞

本研究は国立研究開発法人科学技術振興機構（JST）の研究成果展開事業「センター・オブ・イノベーション（COI）プログラム」の助成を受けたものです．

参考文献

[Uemura 2017] Joji Uemura, Kazuya Mera, Yoshiaki Kurosawa, and Toshiyuki Takezawa: Suppressed Negative-Emotion-Detecting Method by using Transitions in Facial Expressions and Acoustic Features , Proc. Emotions , Decisions and Opinions 2017, pp. 122-127, 2017.

[Eyben 2010] Florian Eyben , Martin Wöllmer , and Björn Schuller: openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor , Proc. ACM Multimedia Conference – MM, pp.1459-1462, 2010.

[オムロン] OMRON Japan, OKAO Vision | オムロン人画像センシングサイト，https://plus-sensing.omron.co.jp/technology, (2019 年 2 月 8 日アクセス).

[Ekman 2009] Paul Ekman: Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage, W.W. Norton, 2009.

表 2 実験結果

手法	台詞固定			台詞自由		
	精度	再現率	F 値	精度	再現率	F 値
SVM 手法（比較手法）	0.57	0.58	0.57	0.54	0.53	0.53
LSTM 手法（提案手法）	0.59	0.70	0.64	0.52	0.38	0.44
BiLSTM 手法（提案手法）	0.59	0.73	0.65	0.43	0.43	0.43