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

コメディドラマにおける字幕と表情を用いた笑い予測

Predicting Laughters in Comedy Drama with Subtitles and Facial Expression

萱谷 勇太 *1 Yuta KAYATANI　大谷 まゆ *2 Mayu OTANI　Chenhui Chu*3　中島 悠太 *3 Yuta NAKASHIMA　竹村 治雄 *1 Haruo TAKEMURA

*1大阪大学 大学院 情報科学研究科
Graduate School of Information Science and Technology, Osaka University

*2株式会社サイバーエージェント
CyberAgent, Inc.

*3大阪大学 データビリティフロンティア機構
Institute for Datability Science, Osaka University

In this paper, we propose a model to predict whether an utterance leads to laughter in a comedy TV show. And we counts for facial expression that actors/presenters make. The model with subtitles and facial expression constructed as input was able to obtain accuracy, precision, recall, f-score more than model which input subtitle only or model which input only facial expression.

1. はじめに

笑いは，コミュニケーションにおいて非常に重要な役割を持つ．その笑いを引き起こすジョークは，プレゼンテーションにおいても聴講者の関心を示し，引き込むことができる．それだけでなく，ジョークは普段の会話においても会話を盛り上げ，その場を和ませることができる．実際に，これらのジョークによる笑いが生み出す効果は，暗黙的に周知されており，人々は会話の中で笑いを誘発しようとしている．

現在，機械による笑いの予測に向けた研究も進められており，例えば Bertero ら [1] の研究が挙げられる．Bertero らは，笑いが起こる発話はその発話以前の発話が起因となっていると仮定し，時系列情報を扱うことを得意とする LSTM を使用したモデルを構築した．この Bertero らの手法は字幕のみを入力としており，登場人物の表情などの視覚情報について考慮されていない．しかし，我々は笑いを引き起こすことにおいて視覚情報は重要な手がかりであると考えた．例えば，図 1 は登場人物の発言が終了した時点の表情の変化によって笑いを誘発している．そこで，我々は Bertero らと同様に，笑いを予測することができるモデルを考案し，Bertero らが複数の発話を入力していることに対して，1 つの発話のみを用い，その上で，字幕だけでなく，視覚情報である発話者の表情の特徴量である Action Unit[2] を使用し，それが笑い予測において重要であることを示す．結果として，字幕だけでなく，視覚情報である表情も組み合わせることで，最も高い精度を得ることができた．

図 1: 表情の例（The BIg Bang Theory, Season 3, Espode 12 (CBS) より.）

表 1: Action Unit

AU No.	内容	AU No.	内容
1	眉の内側を上げる	14	えくぼを作る
2	眉の外側を上げる	15	唇両端を下げる
4	眉を下げる	17	オトガイを上げる
5	上まぶたを上げる	20	唇両端を横に引く
6	頬を上げる	23	唇を固く閉じる
7	瞼を緊張させる	25	顎を下げずに唇を開く
9	鼻にシワを寄せる	26	顎を下げて唇を開く
10	上唇を上げる	45	瞬く
12	唇両端を上げる		

2. 提案手法

提案手法として，モデルへの入力を表情・字幕とし，その二つの入力に対するモデルの出力によって，笑いが起こった確率を得る．

2.1 表情の取得

本研究では，表情を取得する手法として，OpenFace[3] を用いる．OpenFace では，目・鼻・口などの顔のパーツの形状や位置，頭の向き，視線の向き，顔の表情（Action Unit）などを検出することができる．Action Unit とは，顔面筋肉の解剖学的知見を基礎とした，44 の動作単位の事を指し，例えば，笑顔を表す時，Action Unit は 6 + 12 となる．OpenFace では 17 種類の Action Unit の強度 (0 - 5) が取得できる．OpenFace で得られる Action Unit は表 1 の通りである．

2.2 ネットワーク構成

本研究におけるネットワークモデルを図 2 に示す．二つの入力における詳細は以下の通りである．

- 字幕

  字幕に関しては，単語分散表現である GloVe[4] を使用する．この時，発話ごとに単語のベクトル値の平均値を計算したものを使用し，また入力の発話は 1 つのみとする．

連絡先: 萱谷 勇太, 大阪大学 大学院 情報科学研究科, yuta.kayatani@lab.ime.cmc.osaka-u.ac.jp

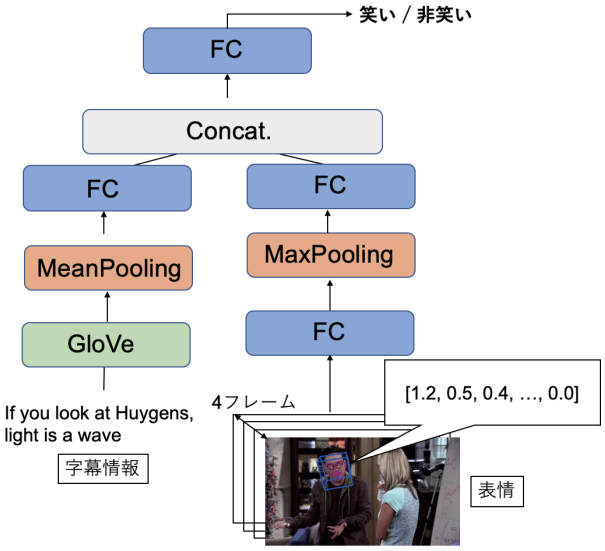

図 2: ネットワークモデル図

表 2: データの内訳

エピソード数	159
発話数	84191
笑いの割合	43.2%
非笑いの割合	56.8%

- 表情

  表情に関しては OpenFace を用いて出力された Action Unit の強度情報を使用する．この時，発話終了時点の時間のフレームだけでなく,「発話終了時点の時間 + 3 秒内のフレーム」,「発話終了時点の時間 − 3 秒内のフレーム」を入力としたモデルをそれぞれ構築した．それぞれは，時系列方向へのマックスプーリングを行うことで 1 次元に圧縮しこれを使用する．この時，発話者の Action Unit ではなく，その発話の時間において OpenFace により出力される，信頼度の最も高い登場人物の Action Unit において，マックスプーリングを行う．

3. 実験

3.1 実験設定

本研究では，笑いを予測するタスクのデータセットとして海外コメディドラマである The Big Bang Theory を使用した．

データにおける笑い・非笑いのラベリング方法としては発話が終わった時間から 1 秒以内の効果音から笑いが検出された時，笑いが起こった（ラベルを 1）とした．また，データにおける 80%をトレーニングセット，10%をバリデーションセット），10%をテストセットとした．データの内訳を表 2 に示す．

また，モデルのパラメータ設定として，エポック数を 80，バッチサイズを 32，最適化手法を AdaDelta とした．

3.2 実験結果

字幕のみ・表情のみ・字幕と表情を使用した場合の実験結果を表 3 に示す．字幕のみ・表情のみを用いた結果に比べて，字幕と表情の両方を用いた場合，−3 秒，+3 秒両方において性能が向上し，+3 秒の時が最も高い性能を達成することができた．この結果から，笑い予測のタスクにおいて字幕だけでなく表情も組み合わせることは精度の向上において有効であることが確認できた．

表 3: 実験結果

手法	Acc	Pre	Rec	F
字幕のみ	63.5	61.6	63.5	61.5
表情のみ	62.2	55.9	62.2	51.3
字幕 + 表情 (−3 秒)	65.7	63.9	65.7	62.0
字幕 + 表情 (+3 秒)	**66.1**	**64.5**	**66.2**	**63.1**

3.3 考察

字幕と表情を用いた場合において，+3 秒・−3 秒の両方について評価を行ったが，+3 秒時点の方がモデルの精度が高かったことから，笑い予測のタスクにおいては発話が終了した後の表情が重要であることが認識できる．しかし，予測という観点において，笑いが起こった後の情報を学習に使用することは適切ではないと考えられる．だが，−3 秒でのモデルでも性能が向上していることから，表情の組み合わせは有効であり，適切なモデルを選定することでさらなる精度の向上が見込まれる．

4. 結論と今後の課題

本論文では，笑い予測のために字幕だけでなく表情を用いたモデルの提案を行った．結果としては字幕だけ，表情だけを用いたモデルよりも両方を用いたモデルのほうが全ての評価指標において高い数値を得ることができることを示した．今後の課題としては，本論文では字幕に付与する情報として表情のみを扱っているが，他にも発話者の骨格情報や姿勢の情報も使用することなどが考えられる．

本研究の一部は科研費 No. 18H03264 による。

参考文献

[1] Dario Bertero and Pascale Fung. A long short-term memory framework for predicting humor in dialogues. In *NAACL-HLT*, pp. 130–135, 2016.

[2] Paul Ekman. Pictures of facial affect. *Consulting Psychologists Press*, 1976.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, pp. 1–10. IEEE, 2016.

[4] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pp. 1532–1543, 2014.