Humor in Chinese Videos

Zixiaofan Yang, Lin Ai, Julia Hirschberg

COMS 6998

April 26, 2019

Outline

- Introduction
- Related Work
- Data collection
- Unsupervised Humor Label Generation
- Feature Extraction
- Experimental Analysis
- Ongoing Work

Why do we study humor?

- Understand human interaction
- Detect when people are being humorous rather than serious to evaluate the content of what they say
- Synthesize humorous speech (e.g. games, advertisements)

What is humor?

1. Producer + Perceiver
2. Positive emotional reactions (laughter)
3. Highly individualistic & cultural specific

Lack of multimedia data annotated with humor

Humor Detection in Text

- 16k one-liners (Mihalcea and Strapparava, 2005)
  - Humor-Specific Stylistic Features: alliteration/rhyme, antonymy, adult slang
    - *"A clean desk is a sign of a cluttered desk drawer"*
- One-liners + 1k news article from "The Onion" (Mihalcea and Pulman, 2007)
  - Human-centeredness and negative polarity
    - *"Take my advice; I don't use it anyway"*
- The New Yorker Cartoon Caption Contest (Radev et al, 2015)
  - Negative sentiment, human-centeredness
    - *"If that 's theseus , I 'm not here."*

Humor Detection in Text

- Extract humor anchor in one-liners (Yang et al., 2015)
  - The subset of candidates that provides the maximum decrement of humor scores
    - *"The one who invented the door knocker got a No-bell prize."*
- 1k tweets (Zhang and Liu, 2014)
  - Phonetic + morpho-syntactic + lexico-semantic + pragmatic + affective features
    - *"I generally avoid temptation unless I can't resist it. - Mae West #quote #humor"*
- TED talk trancripts (Chen and Lee, 2017)
  - Sentences containing or immediately followed by markup '(Laughter)'
    - *"If you're a dog and you spend your whole life doing nothing other than easy and fun things, you're a huge success! (Laughter) "*

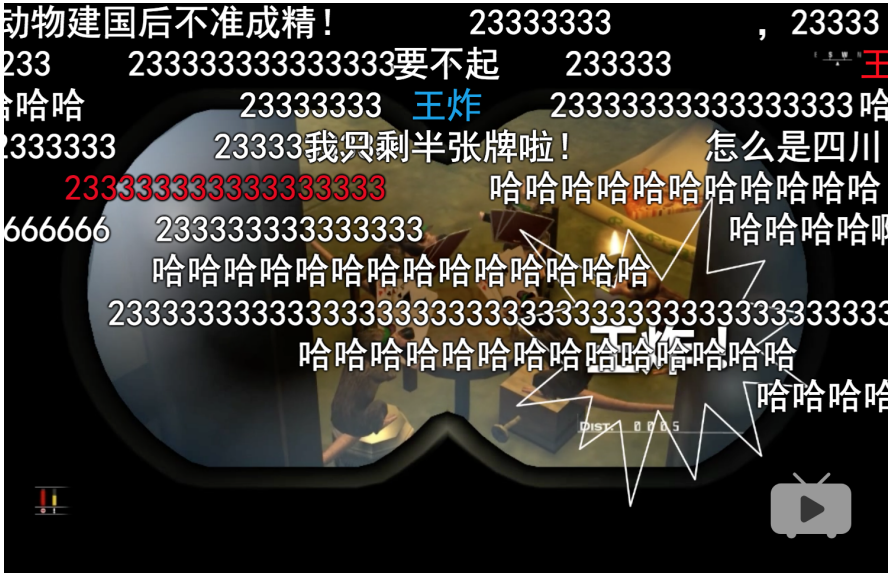
Multimodal Humor Detection

- TV sitcoms
  - Use canned laughters to label humor
    - FRIENDS (Purandare and Litman, 2006)
    - The Big Bang Theory (Bertero and Fung, 2016)
    - Seinfeld (Bertero and Fung, 2016)
  - No study has shown that canned laughter actually represents the audience's perception of humor.

Fig. 1: Example from The Big Bang Theory:
LEONARD: I did a bad thing.
SHELDON: Does it affect me?
LEONARD: No.
SHELDON: ***Then suffer in silence.*** **LAUGH**

‘bullet curtain’ = *Time-aligned comments*

https://www.bilibili.com/



Hypothesis

Audiences tend to respond to humor in videos with laughing

A high volume of laughing comments at a given time

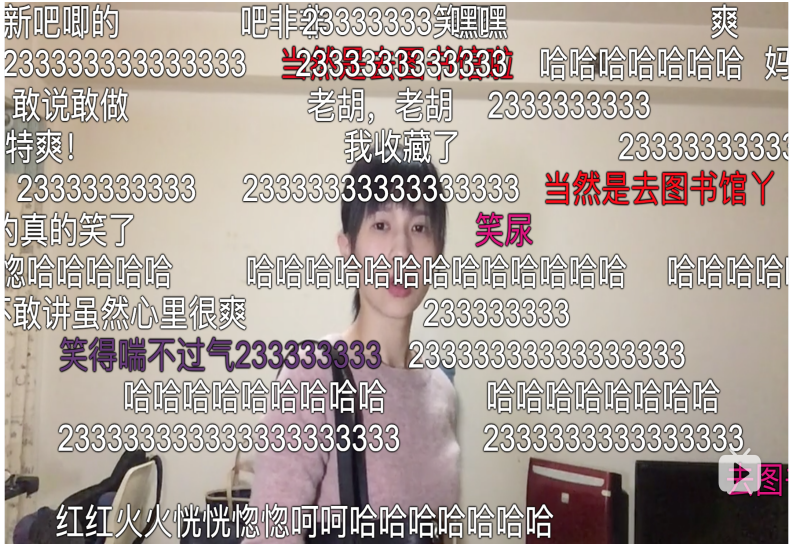
HUMOR!

- Laughing indicators
  - ‘233’ (internet meme)
  - ‘哈哈’ & ‘hh’ (onomatopoeia of laughter)

Data Collection

‘Papi酱’

- A Chinese online celebrity
- Famous for discussing trending topics in a humorous way
- 4 million subscribers, 296 million views on *bilibili.com*



Data Collection

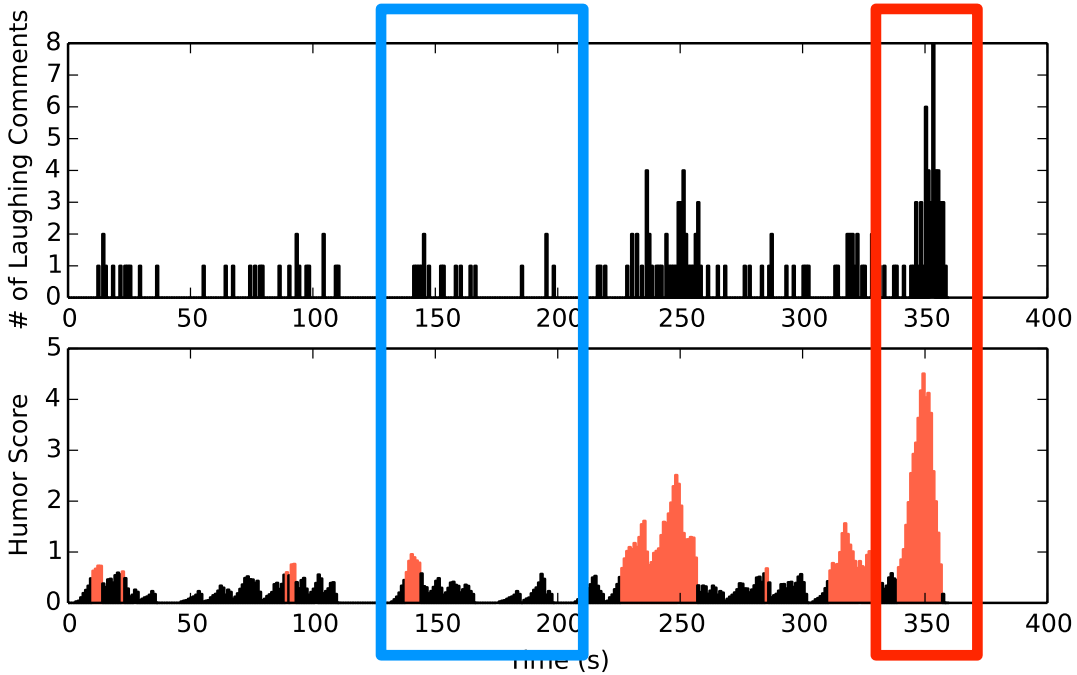
- We use all videos created by ‘Papi酱’
  - Filtered out videos containing dialects and advertisements
  - 100 videos
  - 93593 comments
    - 5064 comments with ‘233’
    - 7255 comments with ‘哈哈’
    - 730 with ‘hh’

Response Time Calculation

- Users typically don’t pause for commenting

- Response Time = reaction time + typing time
- Smooth number of laughing comments by response time
- Set threshold to distinguish humor from non-humor segments

Constructing Unsupervised Labels



Before smoothing

After smoothing

Verification: Human annotation

- Three human annotators
  - Label each second with humor/non-humor
  - Average Cohen's Kappa: 0.65
  - Fleiss' Kappa: 0.65

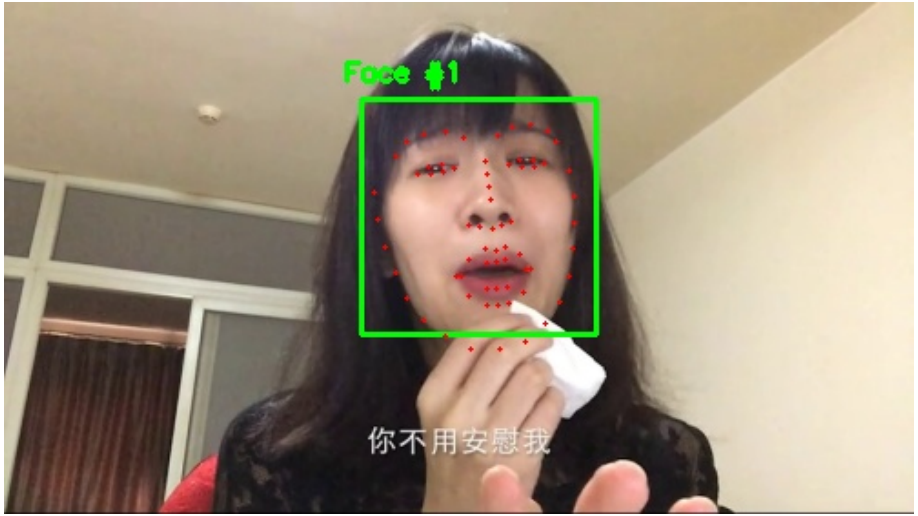
- Gold labels on test set: majority vote
  - Unsupervised labels' accuracy: 0.78

Feature Extraction

- Acoustic-prosodic
  - RMS frame energy and F0
    - 25ms frame, 10ms stride
    - Mean, max, and stddev on 5sec context window
- Transcript-based
  - Slow down and normalize; Google Speech ASR
  - Speaking rate (# character spoken each sec)
    - Range from 0~12
  - Human-centeredness and negation
- Visual
  - Frame difference every 5 frames (SSIM)

Feature Extraction - Ongoing Work

- Facial landmarks
  - dlib: outputs 68 coordinates indicating facial landmarks in static images



Feature Extraction - Ongoing Work

- AlphaPose
  - 17 coordinates marking body conjunctions

Analysis - Speech Features

- Humor expressions have
  - Sudden changes in energy
  - Higher energy and pitch
  - Sudden changes in pitch
  - Slower speaking rate
- Humor techniques
  - Surprise and Exaggeration

Feature	t	p
Energy stddev	24.19	<0.001
Energy mean	23.02	<0.001
F0 mean	22.11	<0.001
Energy max	21.46	<0.001
F0 stddev	19.59	<0.001
F0 max	12.00	<0.001
Speaking rate	-13.94	<0.001

Analysis - Speech Features

(Hamlet) In the end, surprisingly and also not surprisingly — ***everyone died!***

Analysis - Textual Features

- Human-centeredness and negation positively related with humor in one-liners (Mihalcea and Pulman, 2007) (Radev et al, 2015)
  - *"Take my advice; I don't use it anyway."*

- However, humorous punchlines in our videos are different

Feature	**t**	**p**
Human centeredness	-3.74	<0.001
Negation	-6.72	<0.001

Analysis - Visual Features

- SSIM - frame similarity
- Humor segments
  - Are unlikely to be motionless
  - But also have less complete scene-changing

Feature	t	p
SSIM max	-6.79	<0.001
SSIM min	3.72	<0.001
SSIM mean	-2.76	0.006

Analysis - Visual Features

Good news for those who are single!
In 2016 — (beautiful whirling) — you will still be a single dog.

Classification Experiment

- Data
  - 70% (16957sec) as training set
  - 30% (7398sec) as test set
    - with human annotations as gold labels
- Features
  - 384 acoustic-prosodic features from openSMILE
  - TF-IDF Unigram after text segmentation
  - Speaking rate & SSIM scores
- Random Forest Classifier
- F1 score: 0.73

Problems We Encountered

- Intensive punchlines with ~1s duration
  - Perform smoothing carefully
- Adding user weight didn't help on preventing spamming
  - More comments doesn't mean lower quality
- Non-integer video frame rate (24.95/29.95 fps)
  - # of frames in each second is sometimes different
- Google ASR predicts long-lasting characters
  - E.g. a single character starts at 4.1s, but ends at 13.5s
  - Especially when the speech is speeded up

Ongoing & Future Work

- Experiments using different segmentation methods
  - 1-second level → Inter-Pausal Units (IPU) level
- Add more visual features to capture more information
  - Facial expression, gesture, pose, etc.
- Build better model for humor classification

Thanks233!