Speaker Naming in Movies

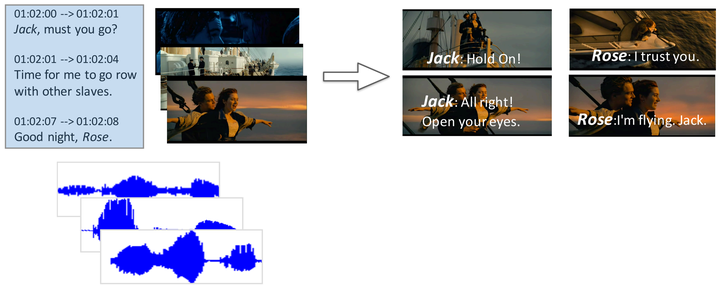
Mahmoud Azab, Mingzhe Wang, Max Smith, Noriyuki Kojima, Jia Deng, Rada Mihalcea
University of Michigan {mazab,mzwang,mxsmith,kojimano,jiadeng,mihalcea}@umich.edu

Introduction

Problem Definition:

Given a movie video and its subtitles, label each segment of the subtitles with the name of its corresponding speaker

Input **Output**



Motivation:

- Speaker naming is important for video understanding, indexing and summarization
- Existing methods are hard to generalize:
  - Use supervised approaches, cannot work on new movies
  - Rely on scripts/cast lists to get speaker names and labels
  - Ignore text, use only vision and speech

Contributions:

- Propose a novel weakly supervised unified multimodal optimization framework for speaker naming
- Construct new speaker naming dataset of 18 movies and 6 episodes of TV shows
- Achieve state-of-the-art performance on movieQA subtitles challenge

Data

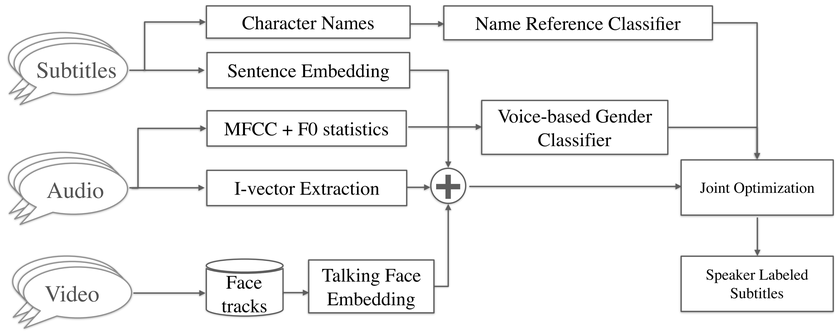
- Construct a new dataset consisting of:
  - 24 videos (18 movies & 6 episodes of a TV show)
  - 31,019 turns
  - 21.99 hours of dialog
  - 437 different character names

with each subtitle segment manually labeled with its corresponding speaker name

- Publicly available at: http://lit.eecs.umich.edu/downloads.html

Framework & Model

Framework Overview:



Names Identification and Reference Classifier:

Target Labels

- [Sheldon, Dr. Cooper, Sheldon Cooper] —> Sheldon
- [Mrs. Cooper, Mary Cooper, Mary] —> Mary Cooper
- [Leonard] —> Leonard
- [Howard, Howard Wolowitz] —> Howard

Weak Labels

```
- Hey, Sheldon!            2
- Did you see Leonard?     3
- I am Penny.              1
```

Talking Face Embedding:

Joint Optimization:

- Approach speaker naming as a transductive learning problem with constraints
- Model the objective function as a linear combination of different losses

Matrix representing assignment of names to subtitles

$$f^* = \arg\min_f \lambda_1 L_{initial}(f) + \lambda_2 L_{MI}(f) + \lambda_3 L_{neg}(f) + \lambda_4 L_{gender}(f) + \lambda_5 L_{dis}(f)$$

- Assign names based on first person reference ("I am Jack")
- Assign same name to speakers with similar facial, textual, acoustic features

Multi-instance constraint:
- 2nd person reference is *some* speaker in the conversation

```
s1: I think we can do it.
s2: I don't think so, Jack.
```

Negative constraint:
- 2nd and 3rd person ref.

```
s: I am flying, Jack.
```
→ s cannot be Jack.

Gender constraint:
- Voice and name genders must match

- More 1st and 2nd person reference → more likely to speak more
- Main characters speak more

Experiments & Results

Intrinsic Evaluation:

- Fine-tuning parameters using 4 movies and testing on the rest

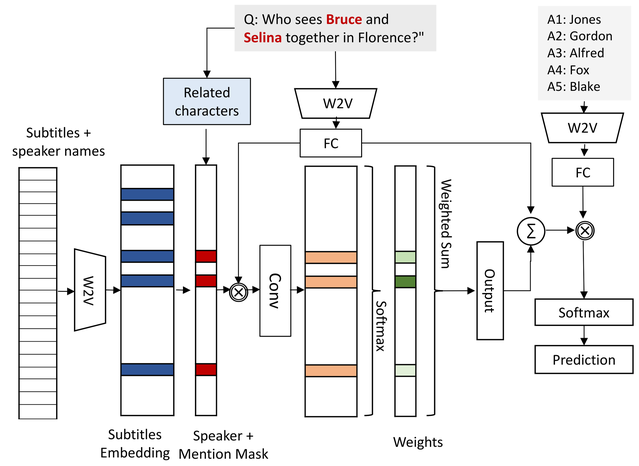
	Precision	Recall	F-score
B1: MFMC	0.0910	0.2749	0.1351
B2: DRA	0.2256	0.1819	0.1861
B3: Gender-based DRA	0.2876	0.2349	0.2317
Our Model (Skip-thoughts)*	0.3468	0.2869	0.2680
Our Model (TF-IDF)*	0.3579	0.2933	0.2805
Our Model (iVectors)	0.2151	0.2347	0.1786
Our Model (Visual)*	0.3348	0.2659	0.2555
Our Model (Visual+iVectors)*	0.3371	0.2720	0.2617
Our Model (TF-IDF+iVectors)*	0.3549	0.2835	0.2643
Our Model (TF-IDF+Visual)*	0.3385	0.2975	0.2821
Our Model (all)*	**0.3720**	**0.3108**	**0.2920**

- Replacing different auxiliary classifiers input with ground-truth

	Precision	Recall	F-score
Our Model	0.3720	0.3108	0.2920
Voice Gender (VG)	0.4218	0.3449	0.3259
VG + Name Gender (NG)	0.4412	0.3790	0.3645
VG + NG + Name Ref	0.4403	0.3938	0.3748

Extrinsic Evaluation:

- We build a full QA model to explore the effectiveness of the speaker naming model on movieQA task



Method	Subtitles val	Subtitles test
SSCB-W2V (Tapaswi et al., 2016)	24.8	23.7
SSCB-TF-IDF (Tapaswi et al., 2016)	27.6	26.5
SSCB Fusion (Tapaswi et al., 2016)	27.7	-
MemN2N (Tapaswi et al., 2016)	38.0	36.9
Understanding visual regions	-	37.4
RWMN (Na et al., 2017)	40.4	38.5
C-MemN2N (w/o SN)	40.6	-
SC-MemN2N (Ours)	**42.7**	**39.4**

This research is based upon work funded by a Samsung research grant