Automatic Composition of Broadcast News Summaries using Rank Classifiers trained with Acoustic and Lexical Features

Taufiq Hasan[1], Mohammed Abdelwahab[2], Srinivas Parthasarathy[2], Carlos Busso[2] and Yang Liu[2]

[1]Research and Technology Center, Robert Bosch LLC, Palo Alto, CA, USA

[2]The University of Texas at Dallas, Richardson, TX, USA

Introduction

- Speech summarization poses unique challenges due to errors in ASR transcripts, sentence end-point detection, music/speech separation, etc.
- Real applications demand that summarized speech satisfies the listener
- Combining sentences into a satisfying summary is highly challenging
- Conventional summarization performance metrics (e.g. ROUGE) do not represent user's satisfaction
- This paper proposes methods of summary composition aiming to improve the perceptual quality of the summarized audio.

Training and Evaluation Data

- Training: 90 news stories from LDC RT-03 MDE corpus used for training
- Annotation: Manual summary labels via Amazon Mechanical Turk
- Evaluation: total 37 news segments with duration of 5~7 minutes selected from online podcasts (NPR, WSJ, CNN)

Features & Classifier

Acoustic features:

- Interspeech 2011 Speaker state challenge feature set
- Various functionals applied to short-term features (fundamental frequency, pitch, etc.) to obtain high level features (4368 dimension)

Lexical features:

- Features: (i) no. of words, (ii) no. of Named Entities, (iii) no. of stop-words, (iv) sentiment polarity, (v) TFIDF (Term frequency - Inverse Document Frequency) vector, and (v) bi-gram language model scores

Structural features:

- Features: (i) duration of sentence, (ii) duration of previous sentence, (iii) duration of next sentence, and (iv) position of sentence

Rank-SVM Classifier:

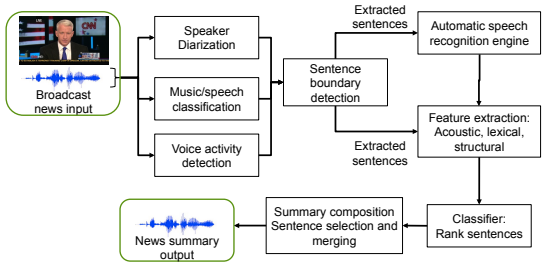
- Pair-wise approach for ranking sentences in news story
- Use rank-SVM to learn the relative importance between sentences
- Use crowd-sourced annotations for ground truth relative importance

Objective Evaluation [Rank Classifier]

Feature set	ROUGE-1	ROUGE-2	ROUGE-L
Acoustic	0.51912	0.34858	0.50261
Lexical	0.52002	0.36904	0.50680
Structural	0.51642	0.32996	0.49939
Combined	**0.55974**	**0.39472**	**0.54450**

Table: Results comparing 10% summary performance for different features using the rank-SVM classifier. ROUGE metrics are used with manual transcripts as gold standard.

Summarization system [Block diagram]



Emotion-aware summarization

- Assume emotional content in summary are more satisfying to users
- News stories are annotated for emotion using Mturk
- Activation and valence levels are used to modify sentence ranking
- Rank-classifier (Λ_{rank}) and emotion scores ($\Lambda_{emotion}$) fused for ranking:

$$\Lambda_{fusion} = \alpha\Lambda_{rank} + (1-\alpha)\,\Lambda_{emotion} \qquad \Lambda_{emotion} = \frac{\sqrt{Arousal^2 + Valence^2}}{\sqrt{2}}$$

Summary composition methods

Anchor's summary

- First part of the news audio where the anchor provides an overview
- Find transition point by observing drop in sentence rank

Trailer-like summary

- Always select the first sentence
- Chronologically combine high rank sentences until duration limit reached
- Sentences before and after the selected sentences are also included

Hot-spot summary

- The region surrounding the highest ranked sentence is identified
- Use speaker labels to keep coherence and minimize discontinuity

Order-based summary

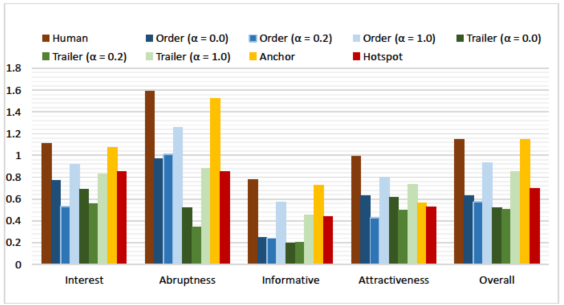
- Simply combine highest ranked sentences until duration limit is reached
- Provides most important sentences (but with possible discontinuities)

Human composed summary

- Skilled audio engineer carefully listened and crafted summaries
- Optimize coherence and listener satisfaction
- Use as gold-standard for audio summaries

Subjective Evaluation [Mechanical Turk]

Criteria	Description
Interest	How interesting is the summary?
Abruptness	How often did you notice unusual discontinuities in the summary?
Information	Does the summary provide adequate information about the story?
Attractiveness	How likely are you to listen to entire news story after hearing the summary?
Quality	How is the overall quality of the audio?



Average subjective evaluation scores for summary generation methods. Results shown in five criteria: i) interest, ii) abruptness, iii) informative, iv) attractiveness, and v) overall. ($\alpha = 1$ indicates no emotional content used)

Results

- Human generated and anchor's summary provides best performance
- Trailer method shows promise in improving 'attractiveness'
- Emotional content in summary does not provide significant benefit. This is due to lack of sufficient emotional variation in news domain

Conclusions

- We designed a fully automatic end-to-end summarization framework
- Mechanical turk was used for summary sentence annotation and perceptual evaluation of summary quality
- Subjective quality of summarization is evaluated using real podcasts
- Effect of emotion is studied in case of news summarization

ICASSP·2016 Shanghai, China
The 41st IEEE International Conference on Acoustics, Speech and Signal Processing, 20-25 March 2016

UTD BOSCH