Prosody-Based Unsupervised Speech Summarization with Two-Layer Mutually Reinforced Random Walk

Sujay Kumar Jauhar
Yun-Nung (Vivian) Chen
Florian Metze

{sjauhar, yvchen, fmetze}@cs.cmu.edu

The 6th International Joint Conference on Natural Language Processing – Oct. 14-18, 2013

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Outline

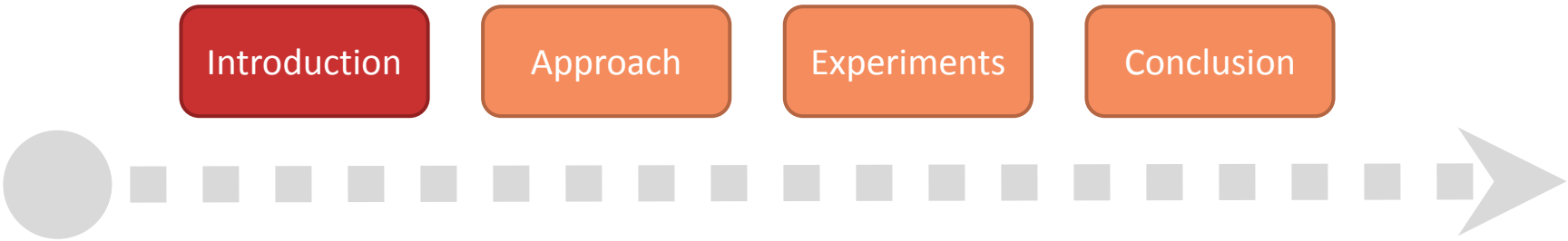
Introduction

Approach

Experiments

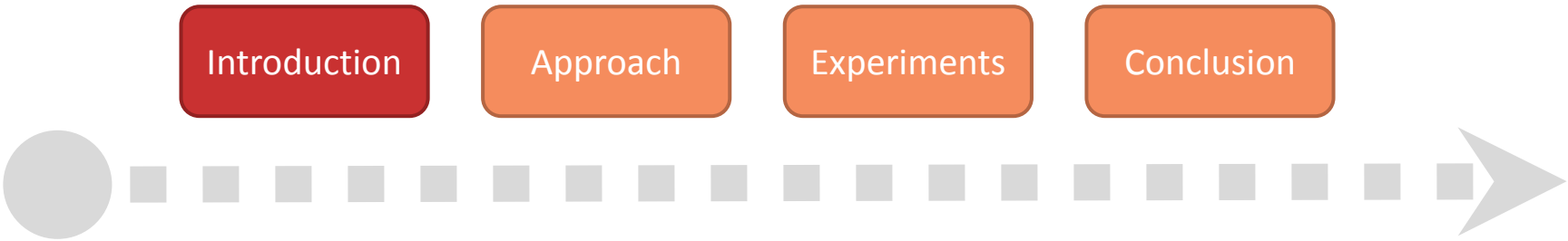
Conclusion

Outline



- Motivation
- Extractive Summarization

Outline



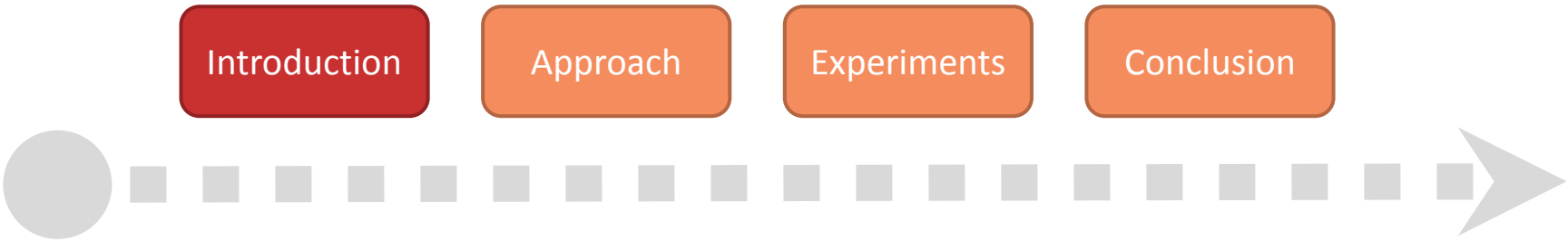
- Motivation
- Extractive Summarization

Motivation

- Speech Summarization
  - Spoken documents are more difficult to browse than texts

    → easy to browse, save time, easily get the key points
- Prosodic Features
  - Speakers may use prosody to implicitly convey the importance of the speech

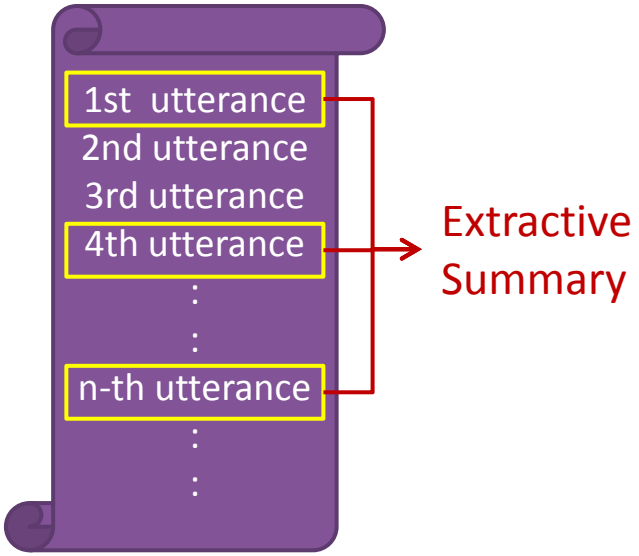
Outline



- Motivation
- Extractive Summarization

Extractive Summarization (1/2)

- Extractive Speech Summarization
  - Select the indicative utterances in a spoken document
  - Cascade the utterances to form a summary



How to select indicative utterances?

Extractive Summarization (2/2)

- Selection of Indicative Utterances
  - Each utterance U in a spoken document d is given an *importance score* I(U, d)
  - Select the indicative utterances based on I(U,d)
  - The number of utterances selected as summary is decided by a predefined ratio

utterance

term

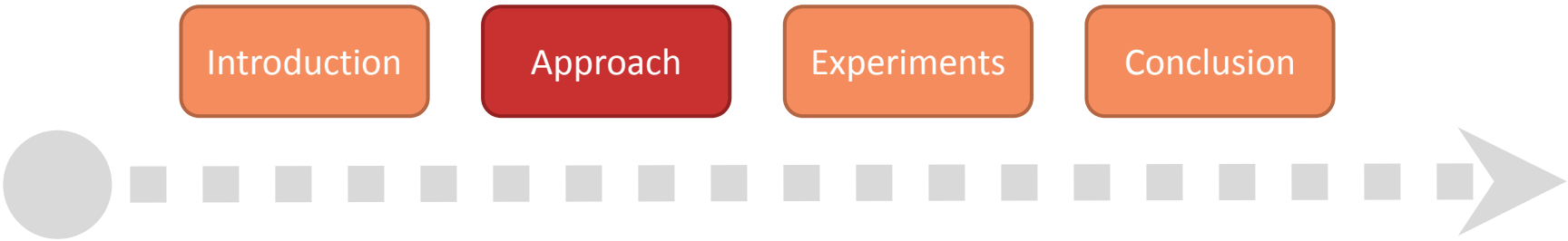
$$U = t_1 t_2 \ldots\ldots t_i \ldots\ldots t_n$$

$$\mathrm{I}(U,d) = \sum_{i=1}^{n} [s(t_i,d) \cdot \ldots\ldots] + \ldots\ldots$$

Importance score

term statistical measure (ex. TF-IDF)

Outline



- Prosodic Feature Extraction
- Graph Construction
- Two-Layer Mutually Reinforced Random Walk

Outline

Introduction | **Approach** | Experiments | Conclusion

- Prosodic Feature Extraction
- Graph Construction
- Two-Layer Mutually Reinforced Random Walk

Prosodic Feature Extraction

- For each pre-segmented audio file, we extract
  - number of syllables
  - number of pauses
  - duration time: speaking time including pauses
  - phonation time: speaking time excluding pauses
  - speaking rate: #syllable / duration time
  - articulation rate: #syllable / phonation time
  - fundamental frequency measured in Hz: avg, max, min
  - energy measured in Pa^2/sec
  - intensity measured in dB

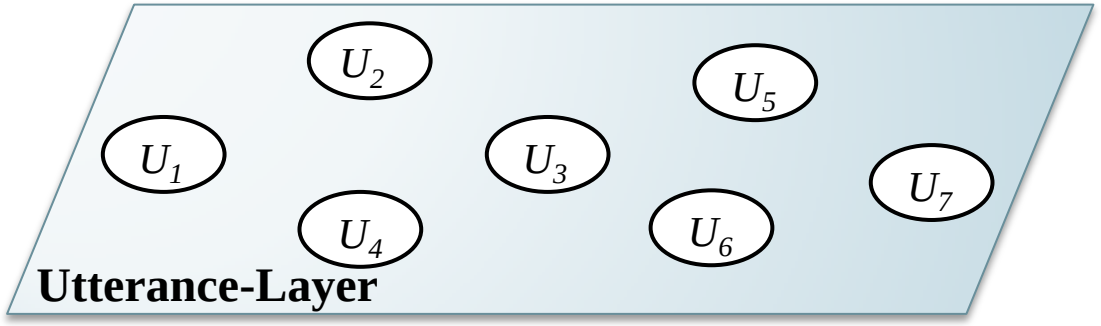
Outline

Introduction | Approach | Experiments | Conclusion

- Prosodic Feature Extraction
- Graph Construction
- Two-Layer Mutually Reinforced Random Walk

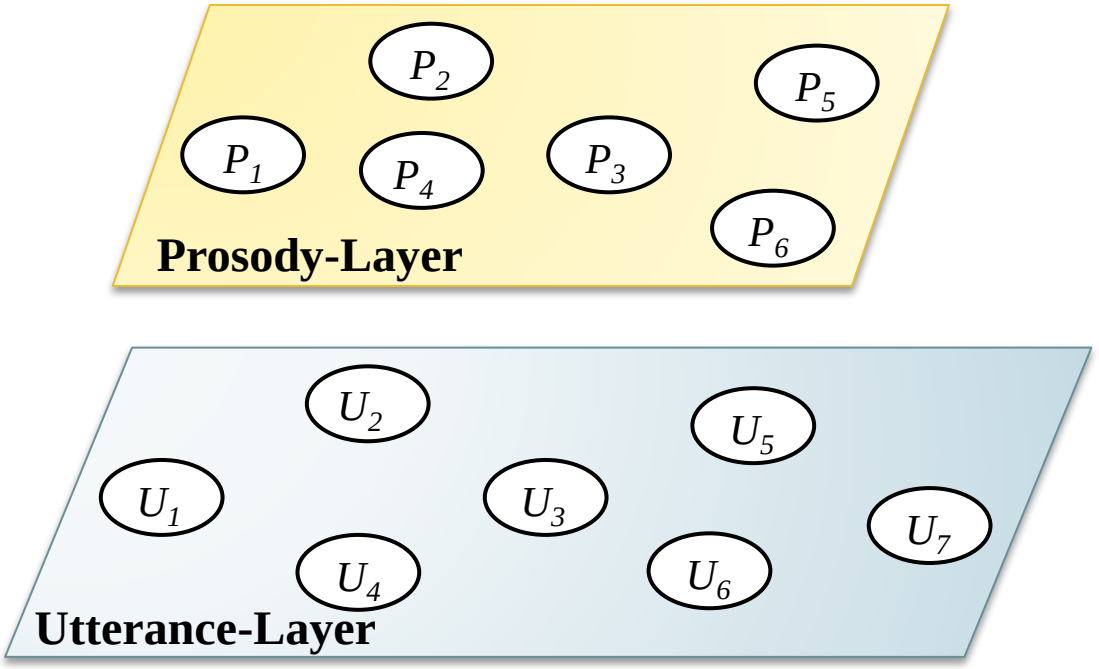
Graph Construction (1/3)

- Utterance-Layer
  - Each node is the utterance in the meeting document



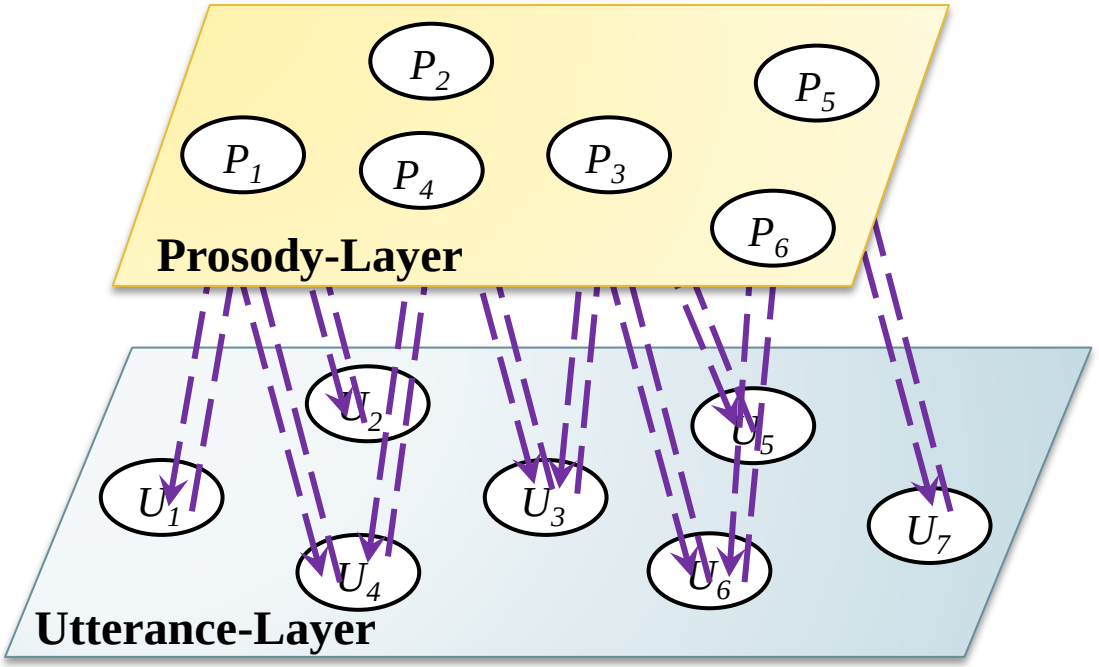
Graph Construction (2/3)

- Utterance-Layer
  - Each node is the utterance in the meeting document
- Prosody-Layer
  - Each node is a prosodic feature



Graph Construction (3/3)

- Utterance-Layer
  - Each node is the utterance in the meeting document
- Prosody-Layer
  - Each node is a prosodic feature
- Between-Layer Relation
  - The weight of the edge is the normalized value of the prosodic feature extracted from the utterance



Outline

Introduction | Approach | Experiments | Conclusion

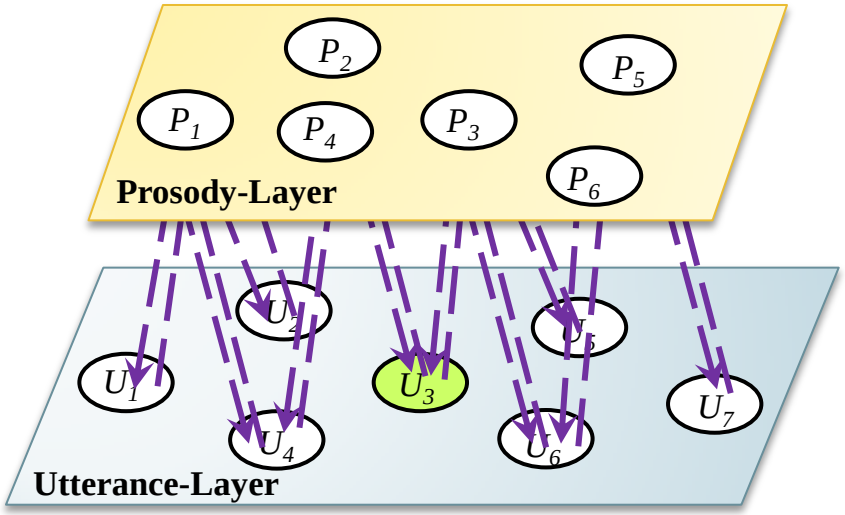
- Prosodic Feature Extraction
- Graph Construction
- Two-Layer Mutually Reinforced Random Walk

Two-Layer Mutual Reinforced Random Walk (1/2)

- Mathematical Formulation

utterance scores at (t+1)-th iteration

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases}$$



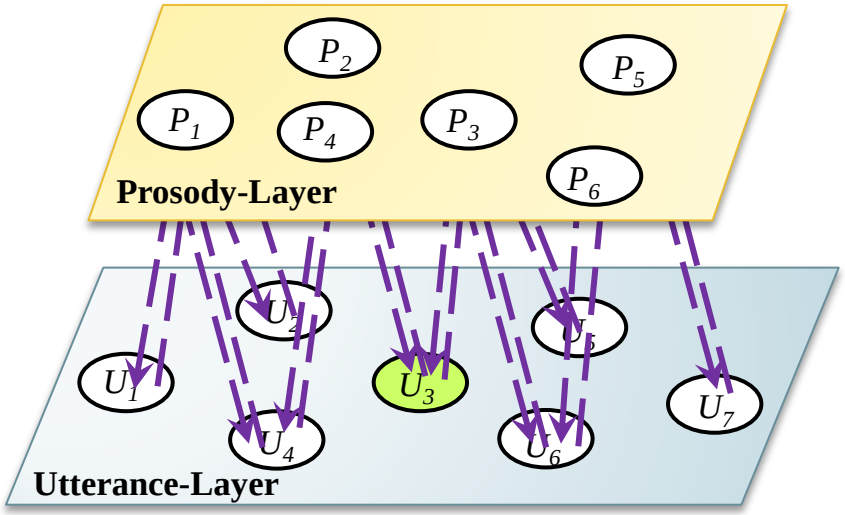
Two-Layer Mutual Reinforced Random Walk (1/2)

- Mathematical Formulation

original importance of utterances

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases}$$

- Original importance
  - Utterance: equal weight



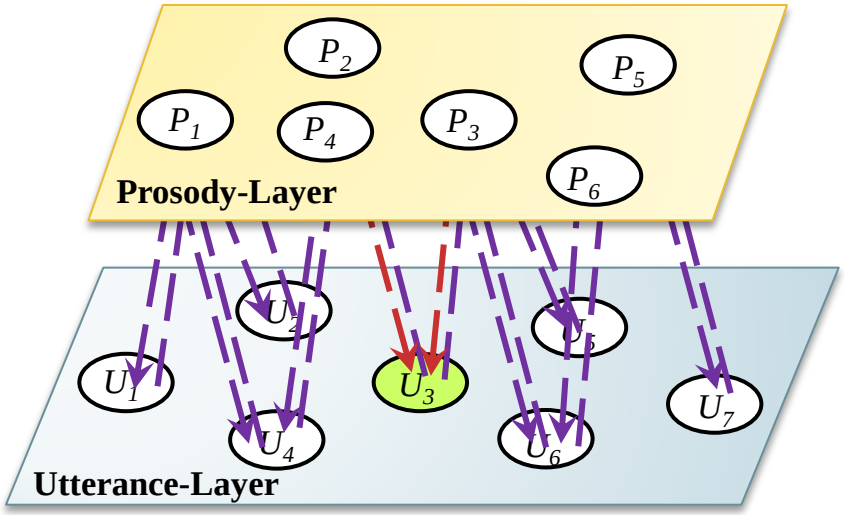
Two-Layer Mutual Reinforced Random Walk (1/2)

- Mathematical Formulation

scores propagated from prosody nodes weighted by prosodic values

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases}$$

- Original importance
  - Utterance: equal weight



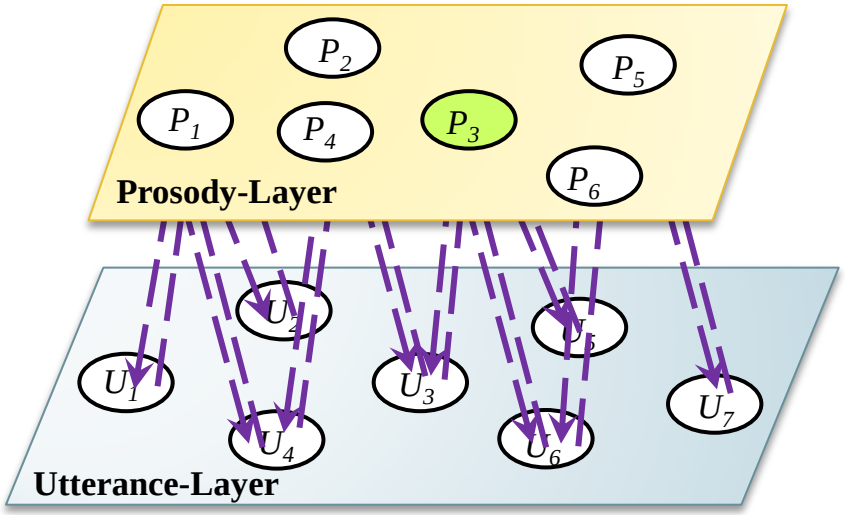
Two-Layer Mutual Reinforced Random Walk (1/2)

- Mathematical Formulation

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases}$$

prosody scores at (t+1)-th iteration

- Original importance
  - Utterance: equal weight



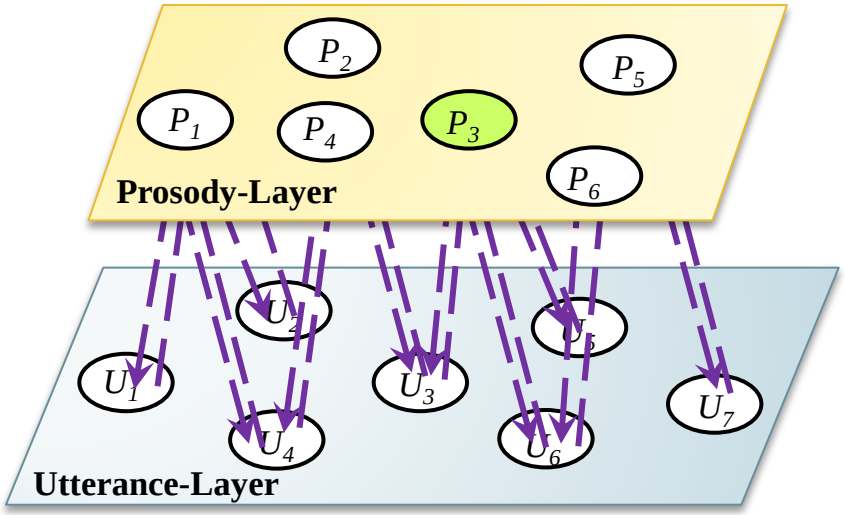
Two-Layer Mutual Reinforced Random Walk (1/2)

- Mathematical Formulation

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases}$$

original importance of prosodic features

- Original importance
  - Utterance: equal weight
  - Prosody: equal weight



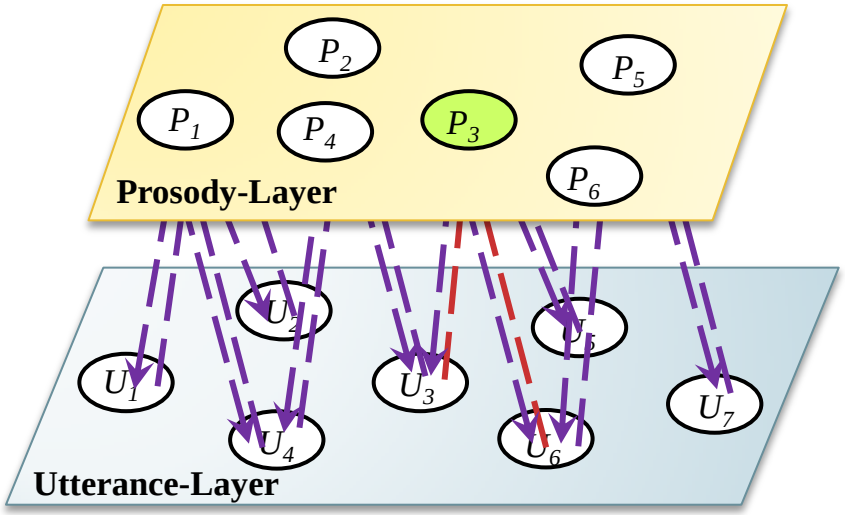
Two-Layer Mutual Reinforced Random Walk (1/2)

- Mathematical Formulation

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot \boxed{L_{PU}F_U^{(t)}} \end{cases}$$

scores propagated from utterances weighted by prosodic values

- Original importance
  - Utterance: equal weight
  - Prosody: equal weight



Two-Layer Mutual Reinforced Random Walk (2/2)

- Mathematical Formulation

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases}$$

Utterance node U can get higher score when

- More important prosodic features with higher weights corresponding to utterance U

Two-Layer Mutual Reinforced Random Walk (2/2)

- Mathematical Formulation

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases}$$

Utterance node U can get higher score when

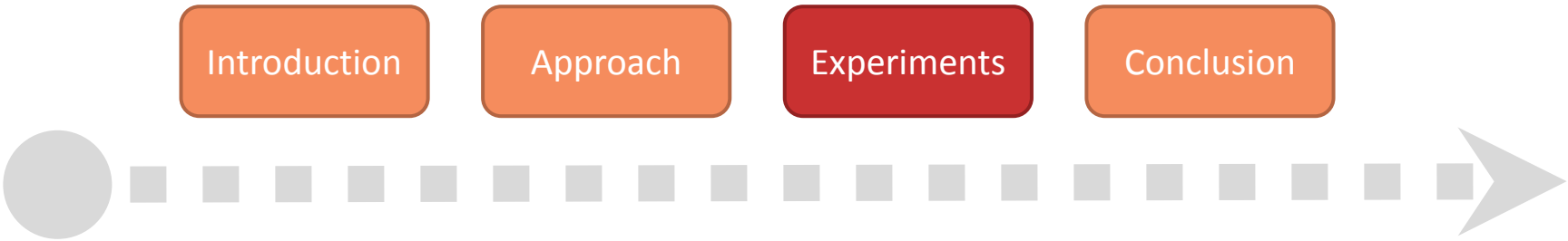
- More important prosodic features with higher weights corresponding to utterance U

Prosody node P can get higher score when

- More important utterances have higher weights corresponding to the prosodic feature P

→ Unsupervised learn important utterances/prosodic features

Outline



- Experimental Setup
- Evaluation Metrics
- Results
- Analysis

Outline

Introduction | Approach | Experiments | Conclusion

- Experimental Setup
- Evaluation Metrics
- Results
- Analysis

Experimental Setup

- CMU Speech Meeting Corpus
  - 10 meetings from 2006/04 – 2006/06
  - #Speaker: 6 (total), 2-4 (each meeting)
  - WER = 44%
- Reference Summaries
  - Manually labeled by two annotators as three "noteworthiness" level (1-3)
  - Extract utterances with level 3 as reference summaries
- Parameter Setting
  - α = 0.9
  - Extractive summary ratio = 10%, 20%, 30%

$$\begin{cases} F_U^{(t+1)} = (1-\alpha)F_U^{(0)} + \alpha \cdot L_{UP}F_P^{(t)} \\ F_P^{(t+1)} = (1-\alpha)F_P^{(0)} + \alpha \cdot L_{PU}F_U^{(t)} \end{cases}$$

REDUNDANCY

Outline

Introduction | Approach | Experiments | Conclusion

- Experimental Setup
- Evaluation Metrics
- Results
- Analysis

Evaluation Metrics

- ROUGE
  - ROUGE-1
    - F-measure of matched unigram between extracted summary and reference summary
  - ROUGE-L (Longest Common Subsequence)
    - F-measure of matched LCS between extracted summary and reference summary
- Average Relevance Score
  - Average noteworthiness scores for the extracted utterances

Outline

Introduction | Approach | Experiments | Conclusion

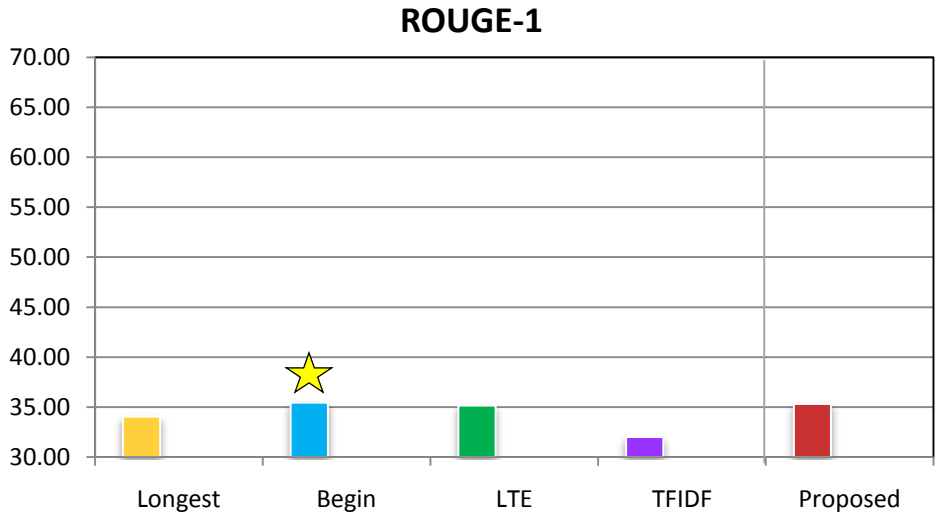
- Experimental Setup
- Evaluation Metrics
- Results
- Analysis

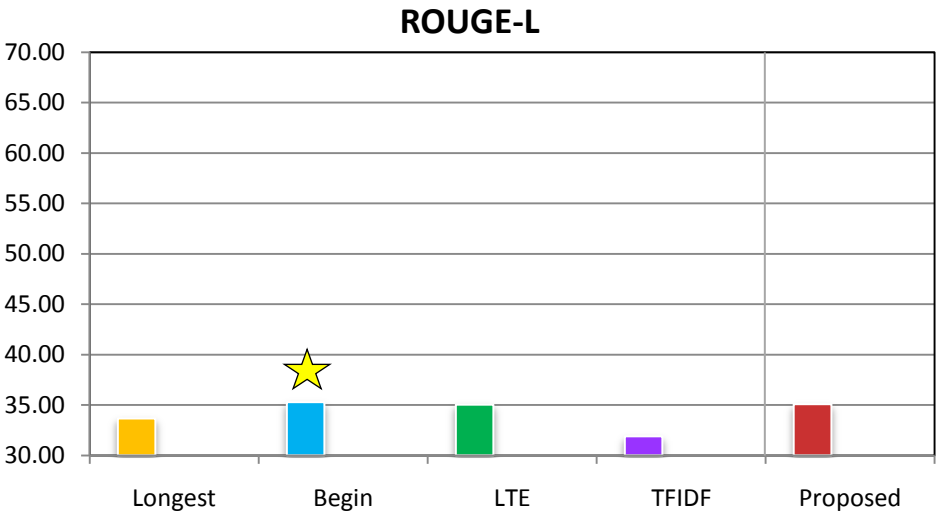
Baseline

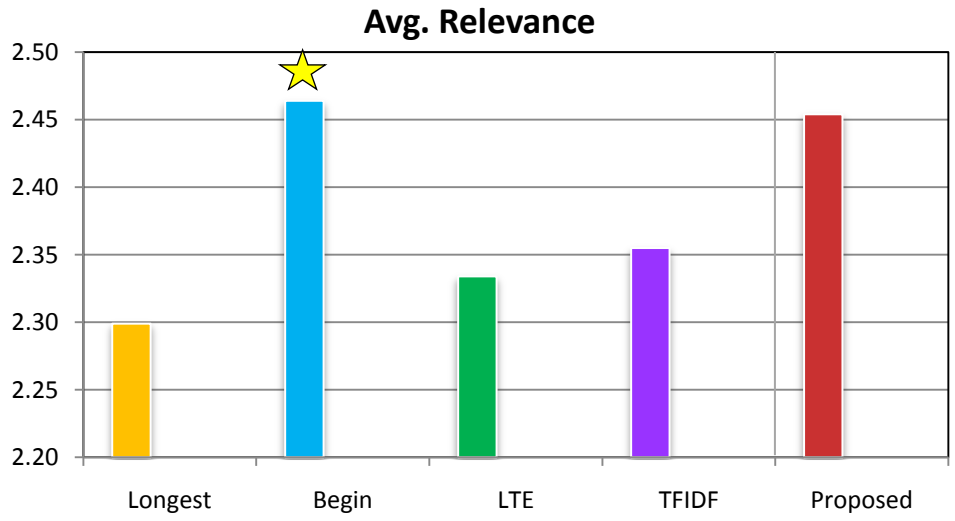
- Longest
  - the longest utterances based on #tokens
- Begin
  - the utterances that appear in the beginning
- Latent Topic Entropy (LTE)
  - Estimate the “focus” of an utterance
  - Lower topic entropy represents more topically informative
- TFIDF
  - Average TFIDF scores of all words in the utterances

10%

Results



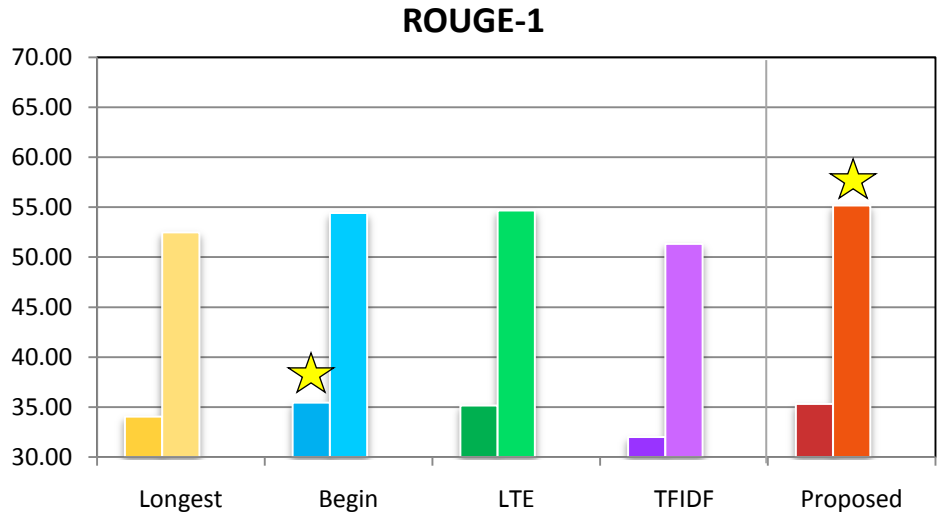


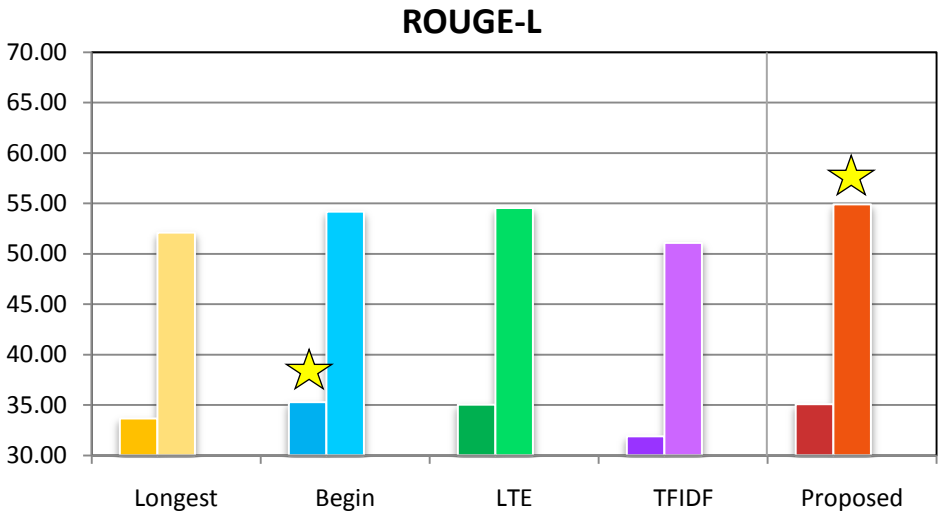


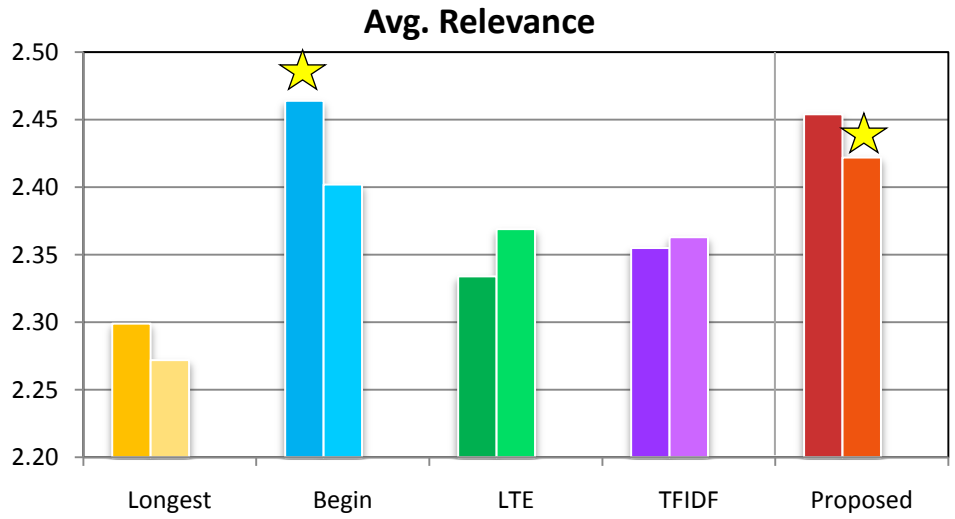
For 10% summaries, Begin performs best and proposed performs comparable results

10% & 20%

Results



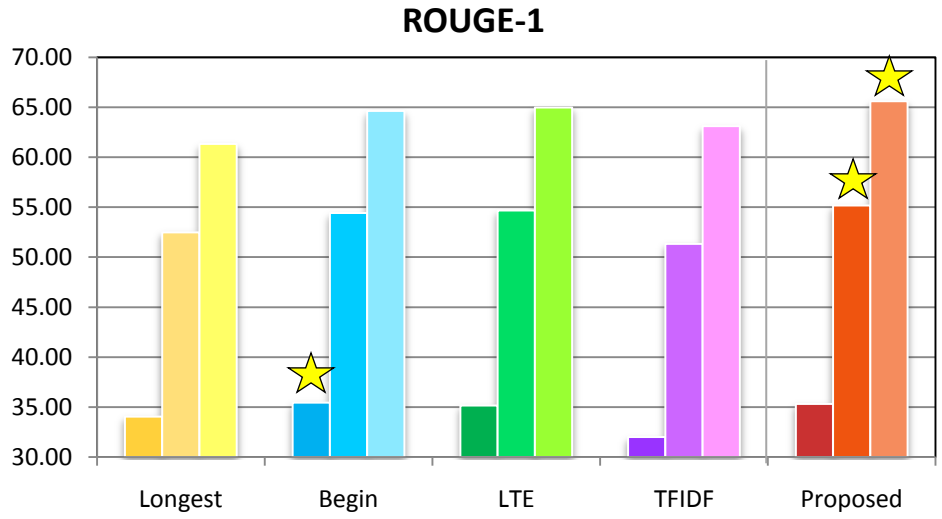


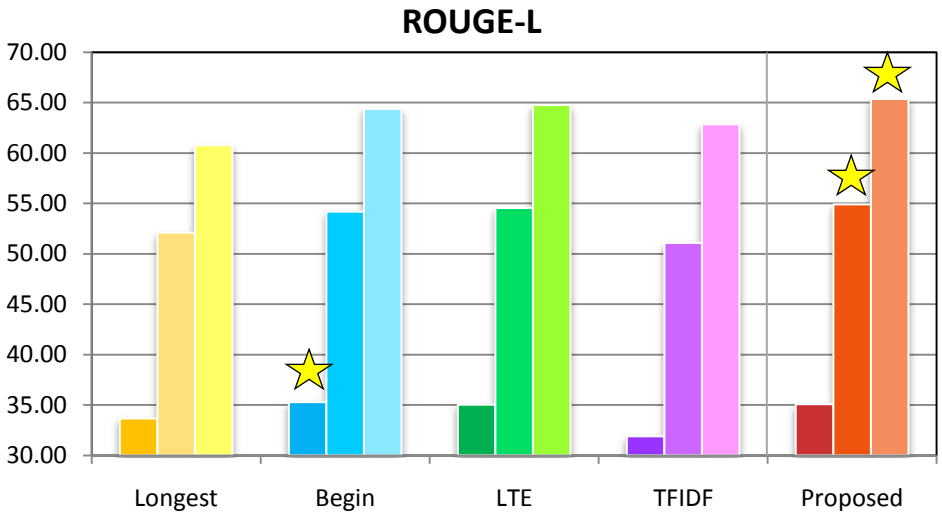


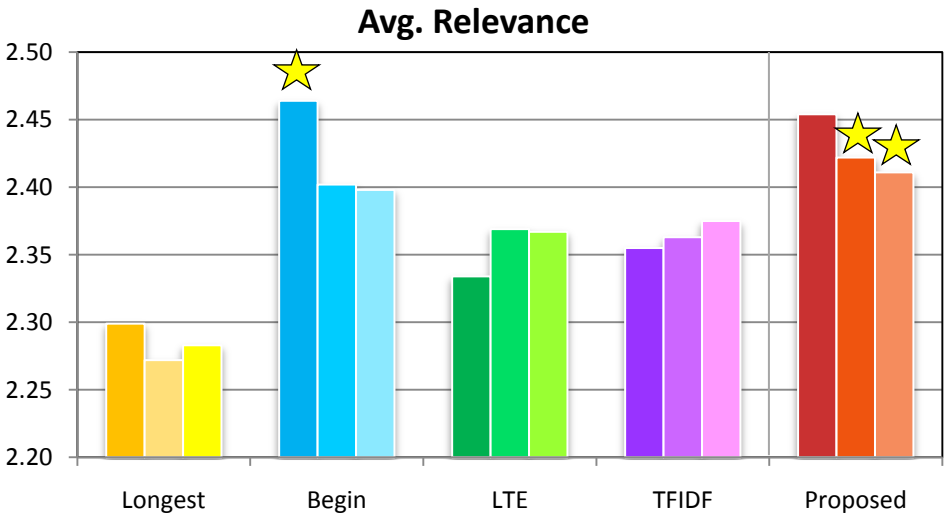
For 20% summaries, proposed approach outperforms all of the baselines

10% & 20% & 30%

Results







For 30% summaries, proposed approach outperforms all of the baselines

Outline

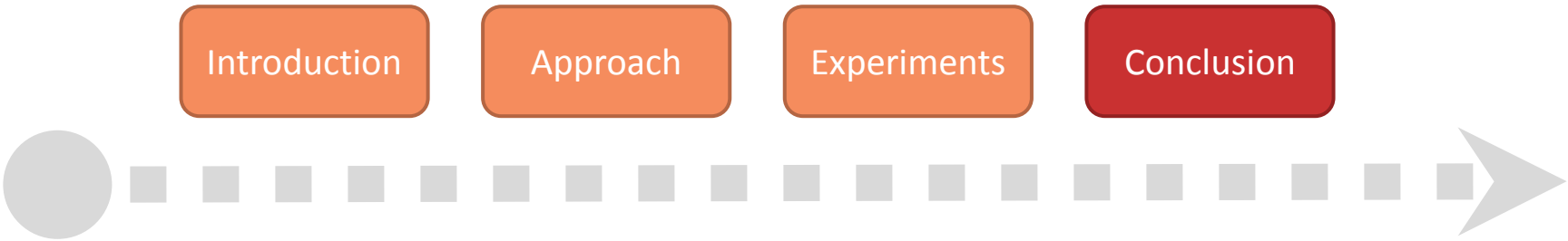
Introduction | Approach | Experiments | Conclusion

- Experimental Setup
- Evaluation Metrics
- Results
- Analysis

Analysis

- Based on converged scores for prosodic features
  - Predictive features
    - number of pauses
    - min pitch
    - avg pitch
    - intensity
  - Least predictive features
    - the duration time
    - the number of syllables
    - the energy

Outline



- Two-layer mutually reinforced random walk integrates prosodic knowledge into an unsupervised model for speech summarization
- We show the first attempt at performing unsupervised speech summarization without using lexical information
- Compared to some lexically derived baselines, the proposed approach outperforms all of them but one scenario

Thanks for your attention! ☺

Q & A