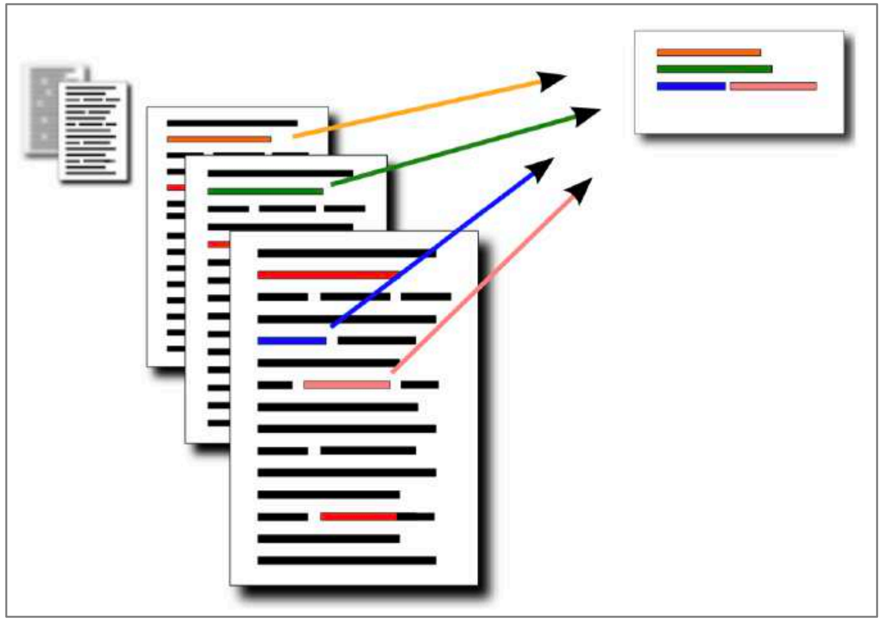
Multi-modal Summarization

Xiachong Feng

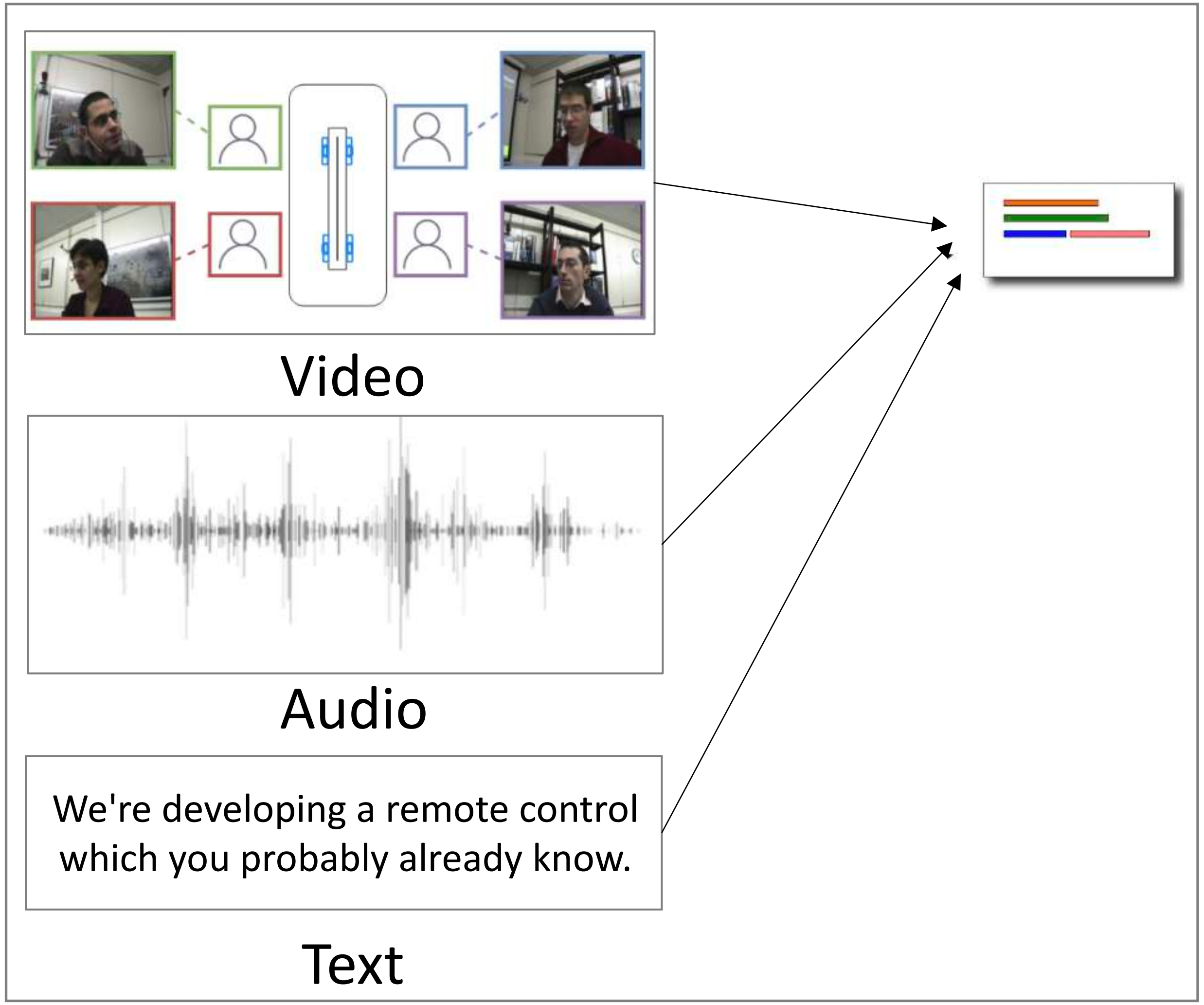
For General Summarization

1. **Brief intro to Summarization**(PPT): http://xcfeng.net/res/presentation/%E6%96%87%E6%9C%AC%E6%91%98%E8%A6%81%E7%AE%80%E8%BF%B0.pdf
2. **Brief intro to Summarization**(Notes): http://xcfeng.net/res/notes/Brief-intro-to-summarization.pdf
3. **ACL19**: http://xcfeng.net/res/presentation/ACL19%20Summarization.pdf
4. **EMNLP19**: http://xcfeng.net/res/notes/EMNLP19_Summarization.pdf
5. **ACL20**: http://xcfeng.net/res/presentation/acl2020-summarization.pdf

Task



Text Summarization



Video

Audio

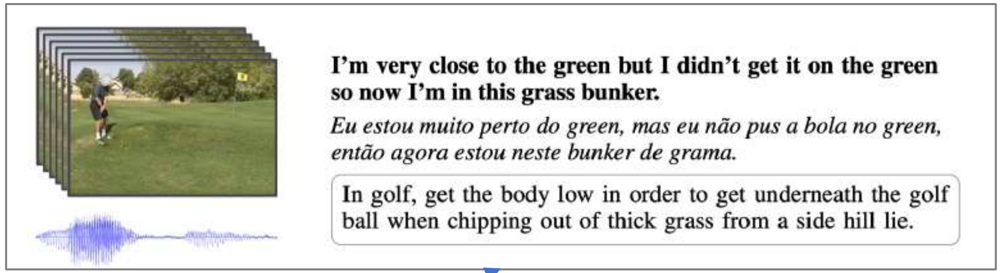
We're developing a remote control which you probably already know.

Text

Multi-modal Summarization

Classification

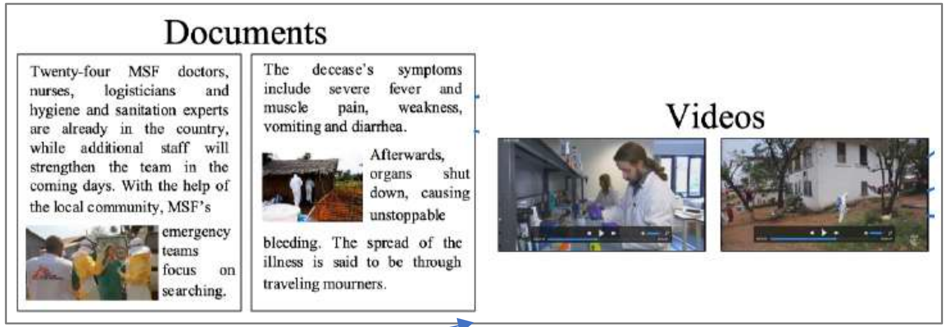
- Classification based on task type



I'm very close to the green but I didn't get it on the green so now I'm in this grass bunker.

Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.

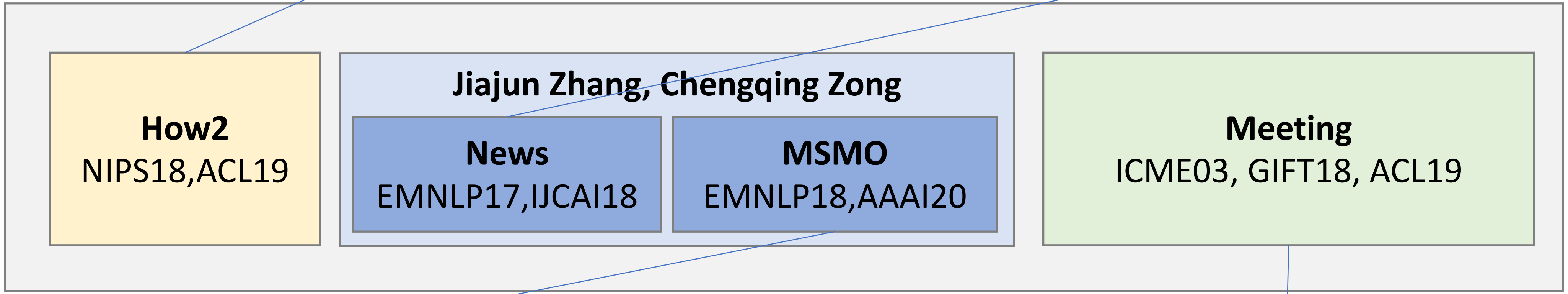
In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.



Documents

Twenty-four MSF doctors, nurses, logisticians and hygiene and sanitation experts are already in the country, while additional staff will strengthen the team in the coming days. With the help of the local community, MSF's emergency teams focus on searching.

The decease's symptoms include severe fever and muscle pain, weakness, vomiting and diarrhea. Afterwards, organs shut down, causing unstoppable bleeding. The spread of the illness is said to be through traveling mourners.



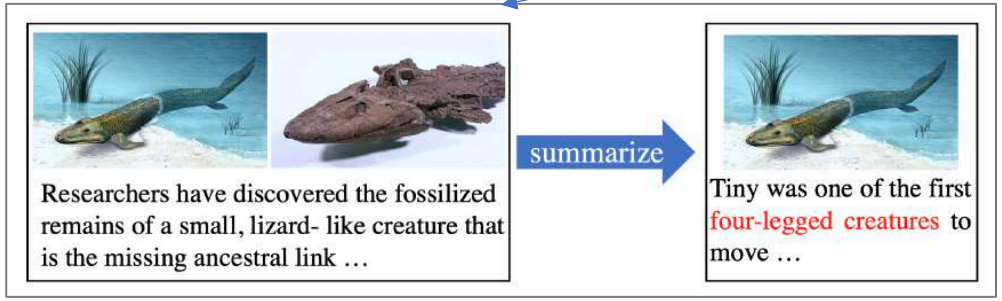
How2
NIPS18,ACL19

Jiajun Zhang, Chengqing Zong

News
EMNLP17,IJCAI18

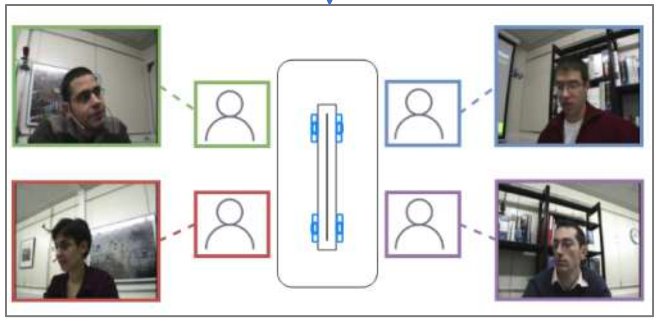
MSMO
EMNLP18,AAAI20

Meeting
ICME03, GIFT18, ACL19



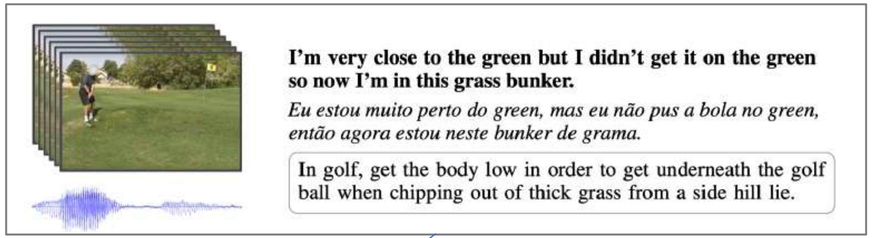
Researchers have discovered the fossilized remains of a small, lizard- like creature that is the missing ancestral link …

Tiny was one of the first four-legged creatures to move …



Classification

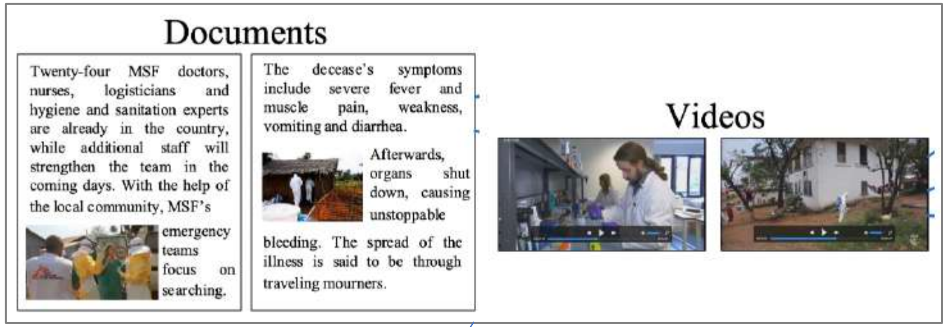
- Classification based on dataset type



I'm very close to the green but I didn't get it on the green so now I'm in this grass bunker.

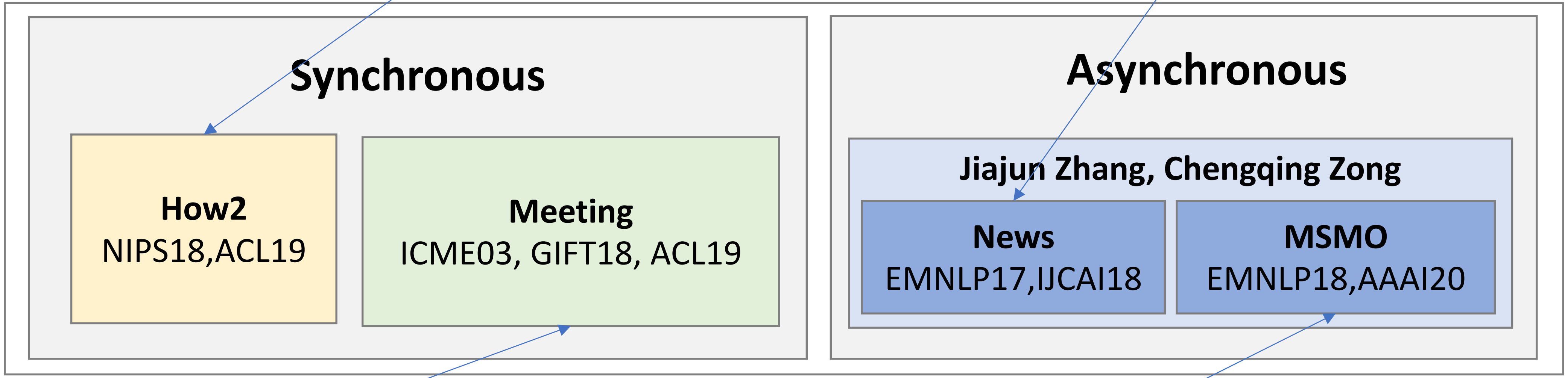
Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.

In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.



Twenty-four MSF doctors, nurses, logisticians and hygiene and sanitation experts are already in the country, while additional staff will strengthen the team in the coming days. With the help of the local community, MSF's emergency teams focus on searching.

The decease's symptoms include severe fever and muscle pain, weakness, vomiting and diarrhea. Afterwards, organs shut down, causing unstoppable bleeding. The spread of the illness is said to be through traveling mourners.



Synchronous

How2
NIPS18,ACL19

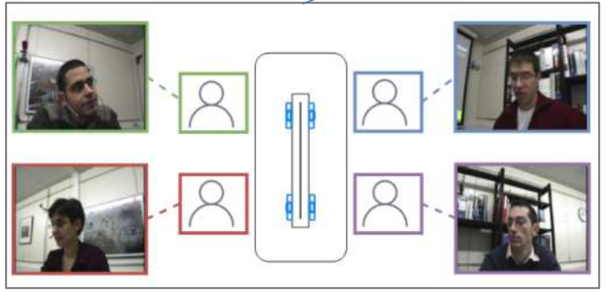
Meeting
ICME03, GIFT18, ACL19

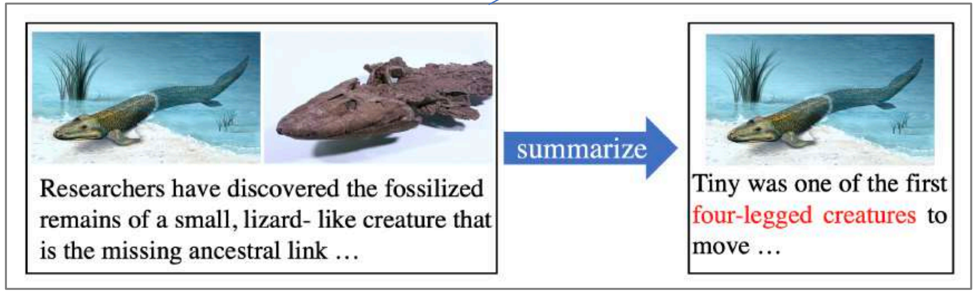
Asynchronous

Jiajun Zhang, Chengqing Zong

News
EMNLP17,IJCAI18

MSMO
EMNLP18,AAAI20





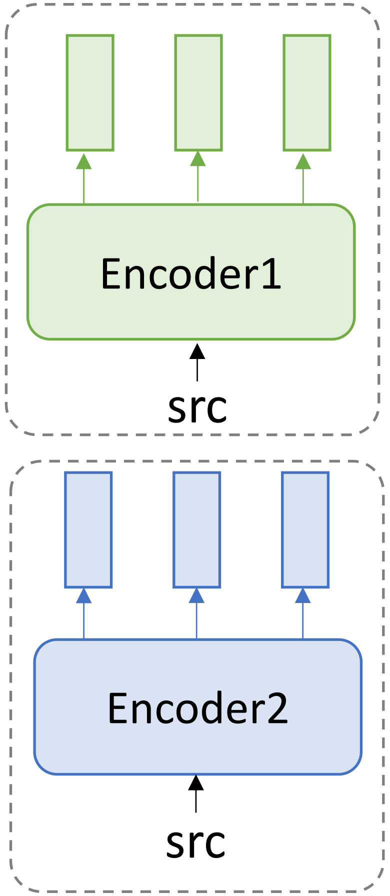
Researchers have discovered the fossilized remains of a small, lizard- like creature that is the missing ancestral link …

Tiny was one of the first four-legged creatures to move …

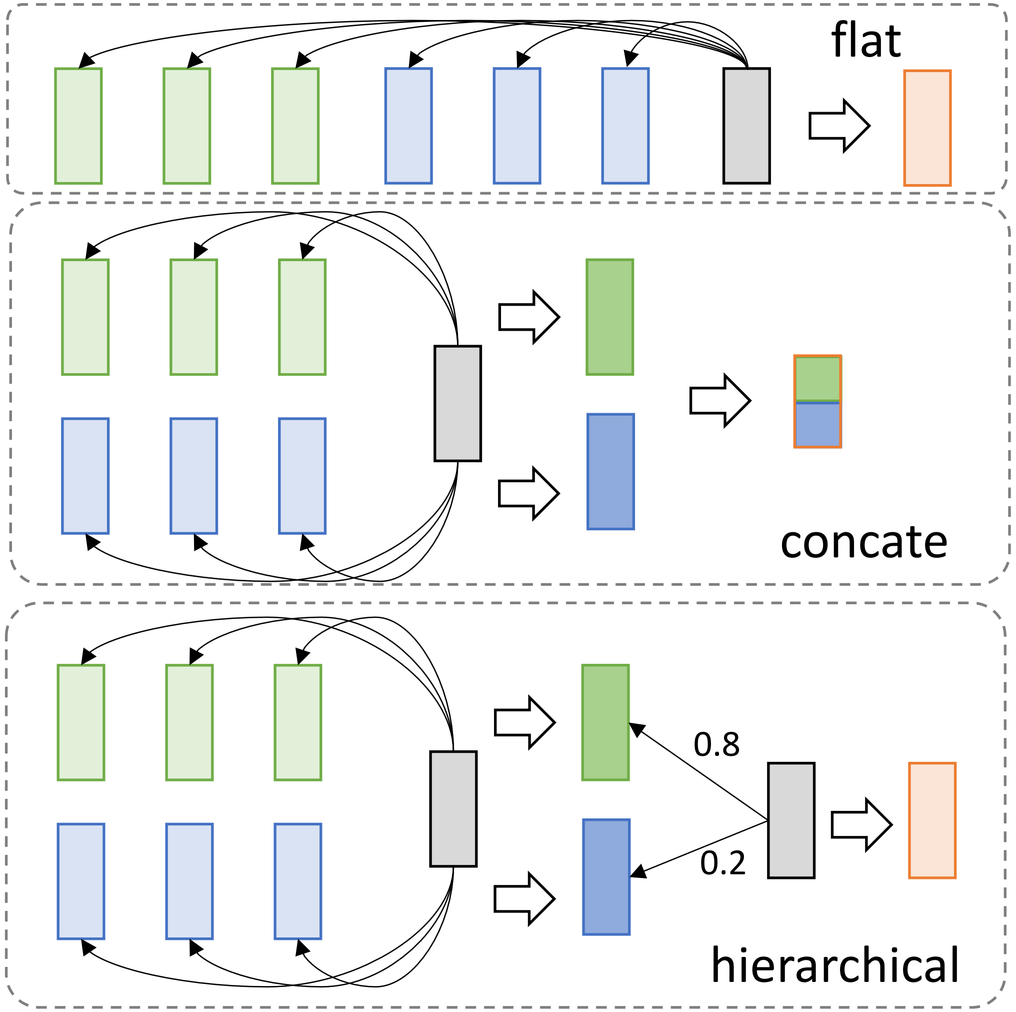
Basic Ideas

- Attention
- Embedding
- Pretrain

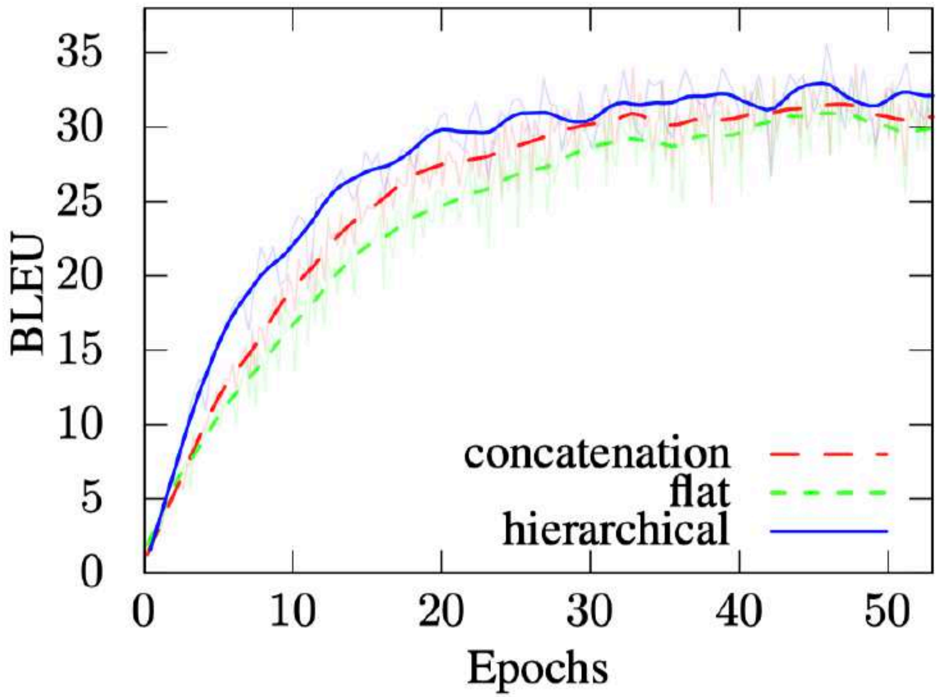
Attention Strategies for Multi-Source Sequence-to-Sequence Learning *ACL17*



Multi Encoder



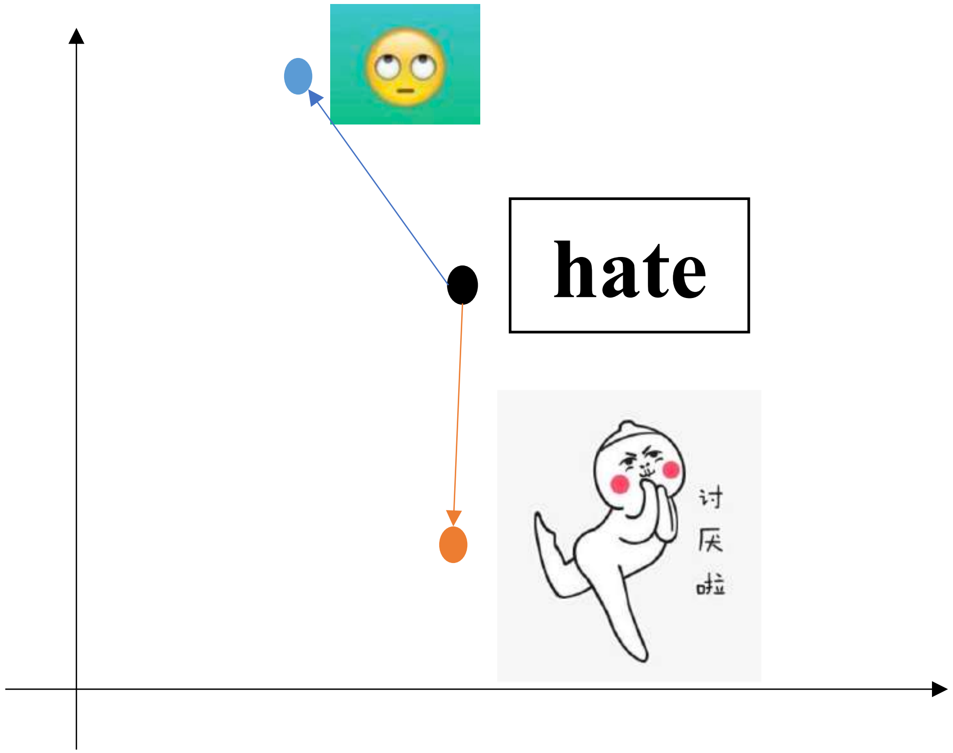
Task : WMT16 Multimodal Translation



Learning curves on validation data

Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors *AAAI19*

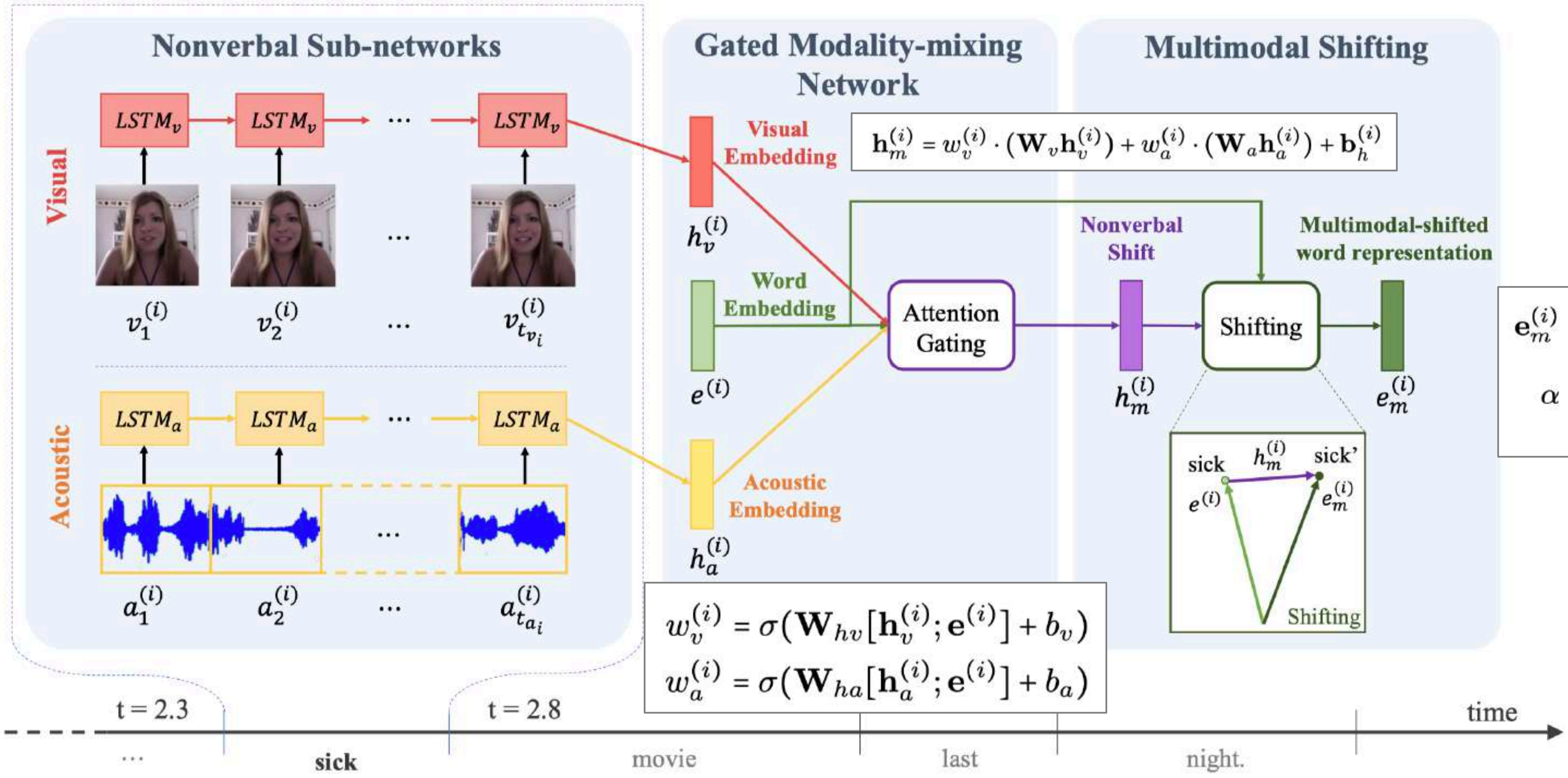
- **Core Idea :** use nonverbal information to adjust word representation



Recurrent Attended Variation Embedding Network (RAVEN)

$$\mathbf{h}_v^{(i)} = \mathrm{LSTM}_v(\mathbf{V}^{(i)})$$

$$\mathbf{h}_a^{(i)} = \mathrm{LSTM}_a(\mathbf{A}^{(i)})$$



$$\mathbf{h}_m^{(i)} = w_v^{(i)} \cdot (\mathbf{W}_v \mathbf{h}_v^{(i)}) + w_a^{(i)} \cdot (\mathbf{W}_a \mathbf{h}_a^{(i)}) + \mathbf{b}_h^{(i)}$$

$$w_v^{(i)} = \sigma(\mathbf{W}_{hv}[\mathbf{h}_v^{(i)};\mathbf{e}^{(i)}] + b_v)$$

$$w_a^{(i)} = \sigma(\mathbf{W}_{ha}[\mathbf{h}_a^{(i)};\mathbf{e}^{(i)}] + b_a)$$

$$\mathbf{e}_m^{(i)} = \mathbf{e}^{(i)} + \alpha \mathbf{h}_m^{(i)}$$

$$\alpha = \min\left(\frac{\|\mathbf{e}^{(i)}\|_2}{\|\mathbf{h}_m^{(i)}\|_2}\beta, 1\right)$$

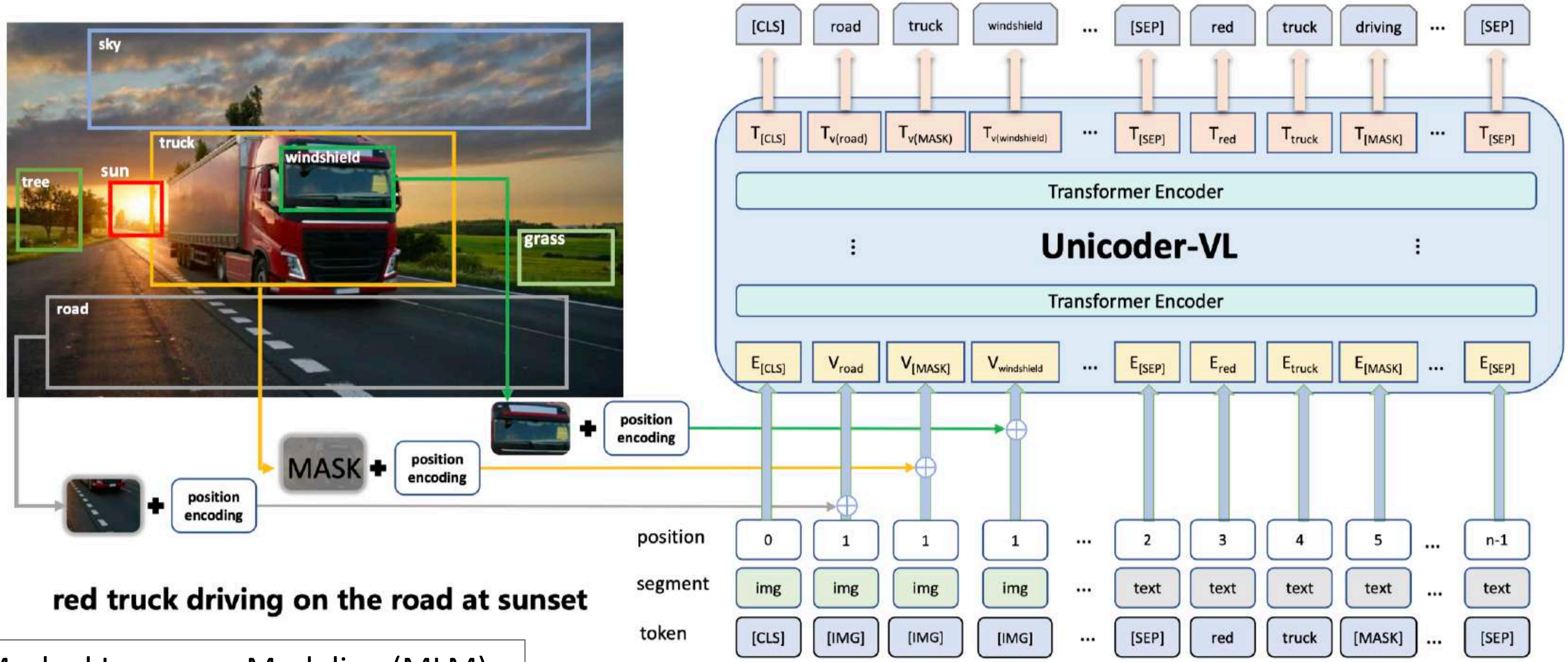
- Visual Features : facial expression analysis toolkit **FACET**(30Hz)
- Acoustic Features : **COVAREP** acoustic analysis framework(100Hz)

Experiment

Multimodal Sentiment Analysis: ✓

Multimodal Emotion Recognition: ×

Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training *AAAI20*

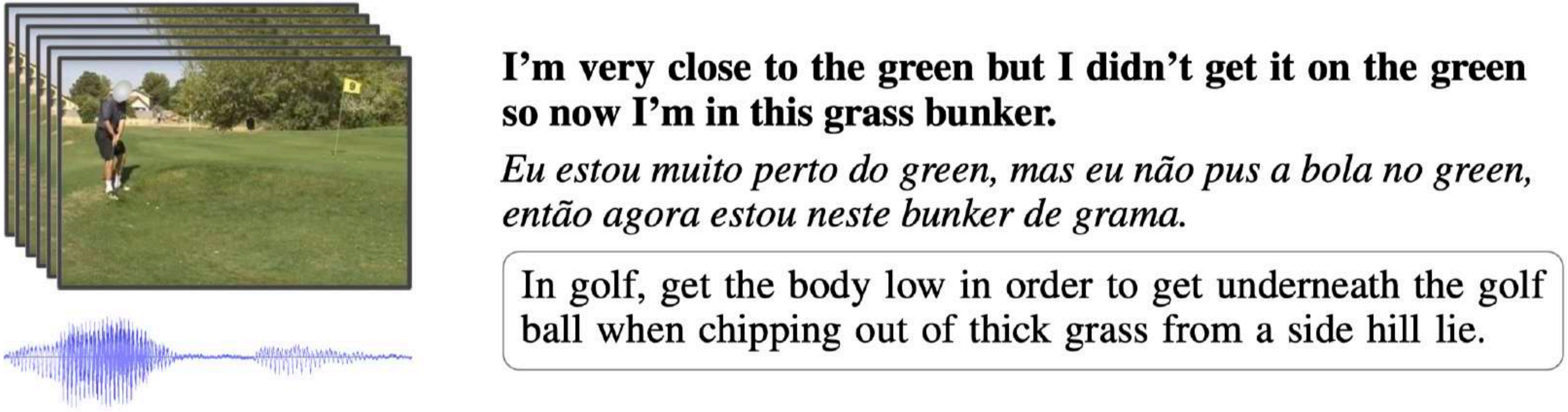


- Masked Language Modeling (MLM)
- Masked Object Classification (MOC)
- Visual-linguistic Matching (VLM)

Downstream Task : Image-text retrieval is the task of identifying an image from candidates given a caption describing its content, or vice versa.

How2

How2: A Large-scale Dataset for Multimodal Language Understanding *NIPS18*



I'm very close to the green but I didn't get it on the green so now I'm in this grass bunker.

Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.

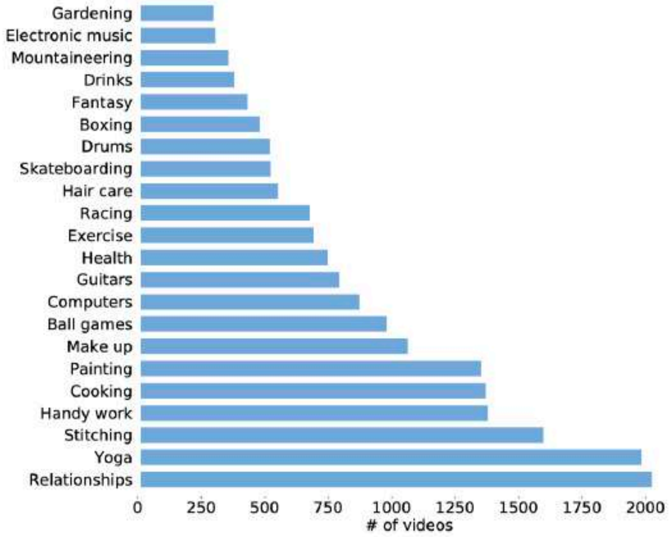
In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.

Figure 1: `How2` contains a large variety of instructional videos with utterance-level English subtitles (in bold), aligned Portuguese translations (in italics), and video-level English summaries (in the box). Multimodality helps resolve ambiguities and improves understanding.

How2: A Large-scale Dataset for Multimodal Language Understanding *NIPS18*

- 2,000 hours of short instructional videos, spanning different domains such as cooking, sports, indoor/outdoor activities, music, etc.
- Each video is accompanied by a human-generated transcript and a 2 to 3 sentence summary

Training	73993
Validation	2965
Testing	2156
Input avg	291 words
Summary avg	33 words

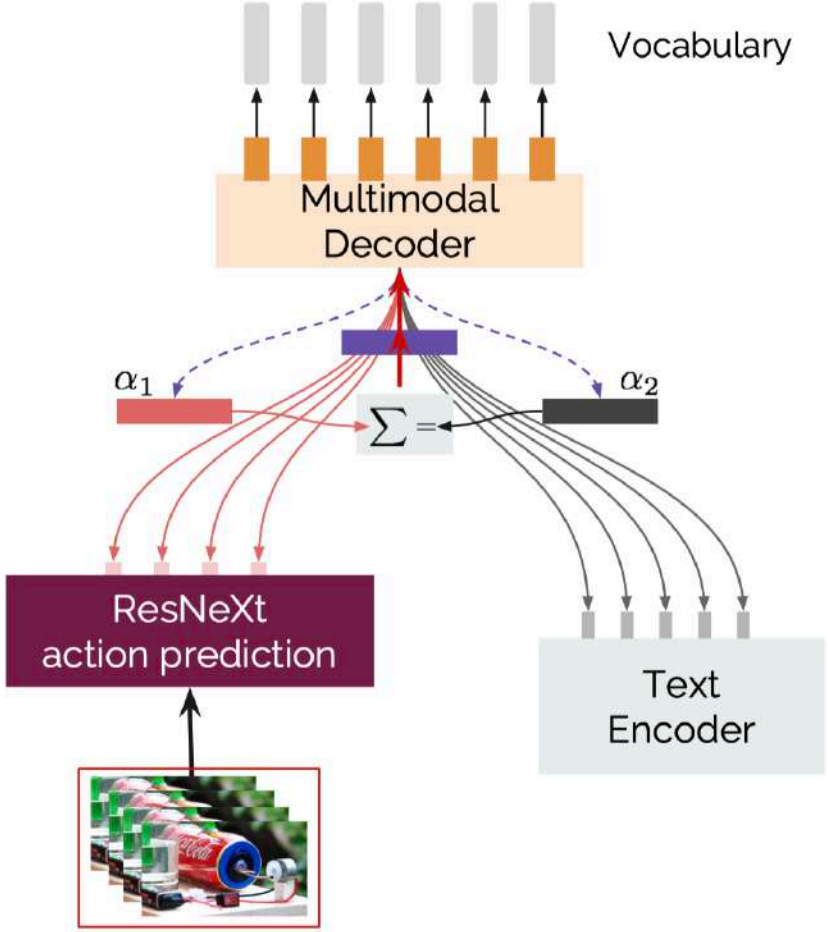


(a) Topic distribution.

(1) Multimodal Abstractive Summarization for Open-Domain Videos *NIPS18*
(2) Multimodal Abstractive Summarization for How2 Videos *ACL19*

- **Action Feature Extraction**
  - 2048 dimensions extracted every 16 non-overlapping frames using a ResNext-101 3D Convolutional Neural Network, 400 different human actions

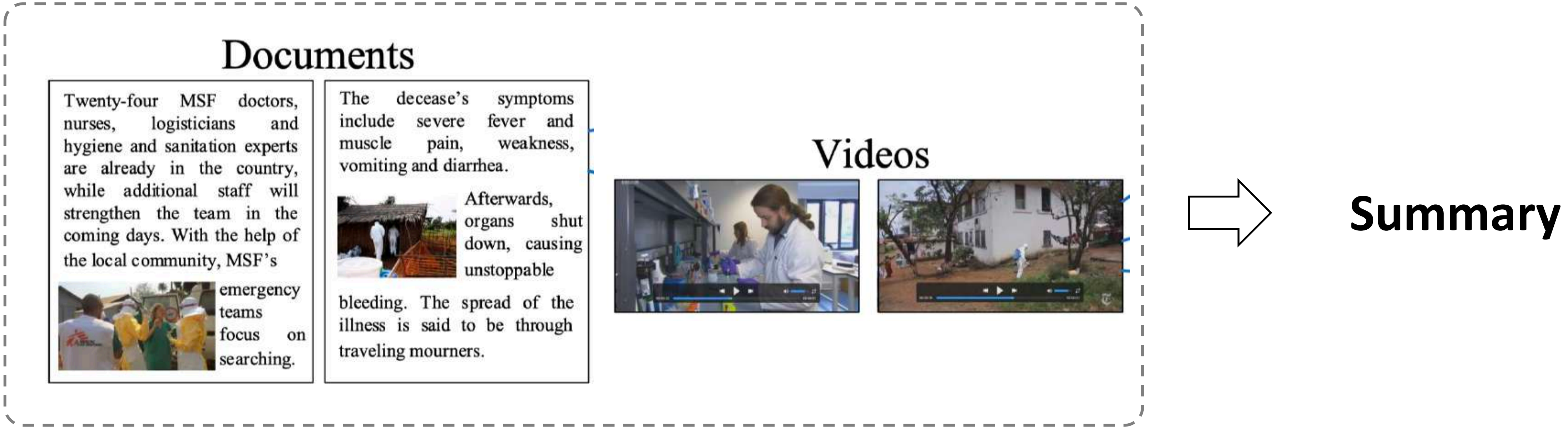
- **Result**
  - Best : Text + Action with Hierarchical Attn



News

(1) Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video *EMNLP17*

(2) Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video *IEEE18*



Documents

Twenty-four MSF doctors, nurses, logisticians and hygiene and sanitation experts are already in the country, while additional staff will strengthen the team in the coming days. With the help of the local community, MSF's emergency teams focus on searching.

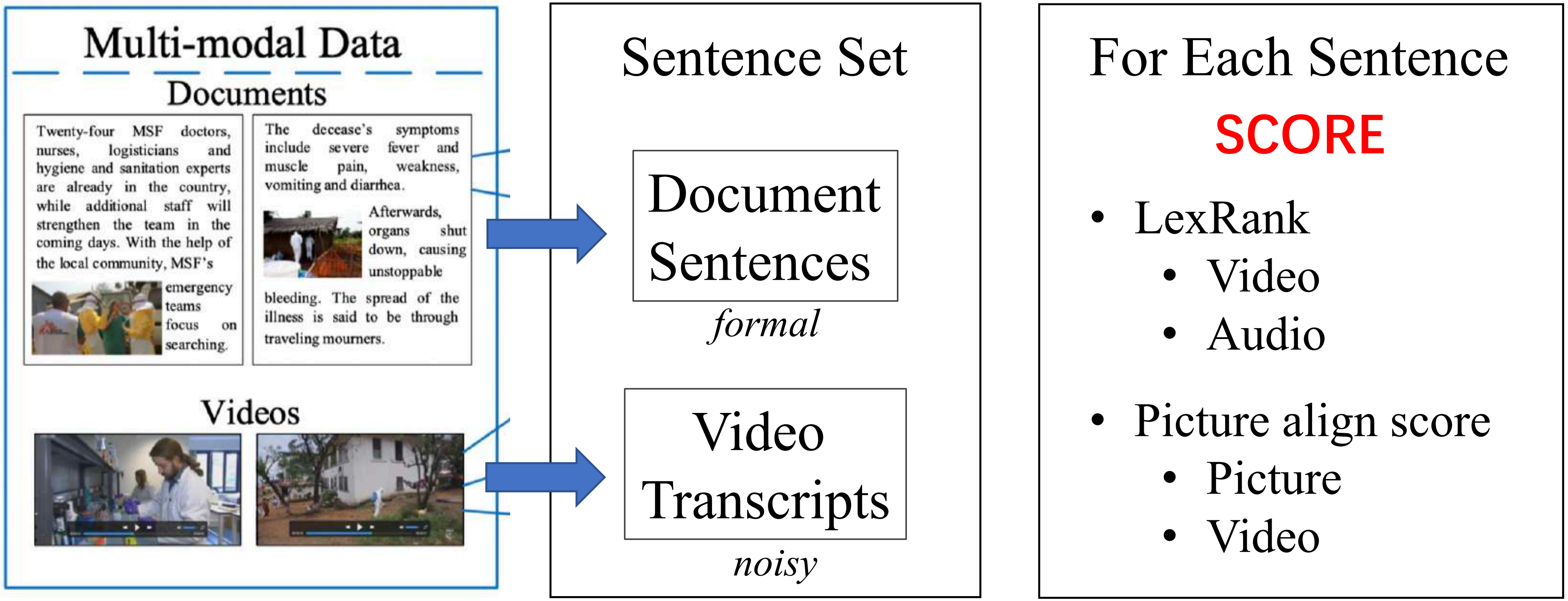
The decease's symptoms include severe fever and muscle pain, weakness, vomiting and diarrhea. Afterwards, organs shut down, causing unstoppable bleeding. The spread of the illness is said to be through traveling mourners.

Videos

Summary

Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video *EMNLP17*

- Method: Extractive Summarization



Multi-modal Data

Documents

Twenty-four MSF doctors, nurses, logisticians and hygiene and sanitation experts are already in the country, while additional staff will strengthen the team in the coming days. With the help of the local community, MSF's emergency teams focus on searching.

The decease's symptoms include severe fever and muscle pain, weakness, vomiting and diarrhea. Afterwards, organs shut down, causing unstoppable bleeding. The spread of the illness is said to be through traveling mourners.

Videos

Sentence Set

Document Sentences

formal

Video Transcripts

noisy

For Each Sentence

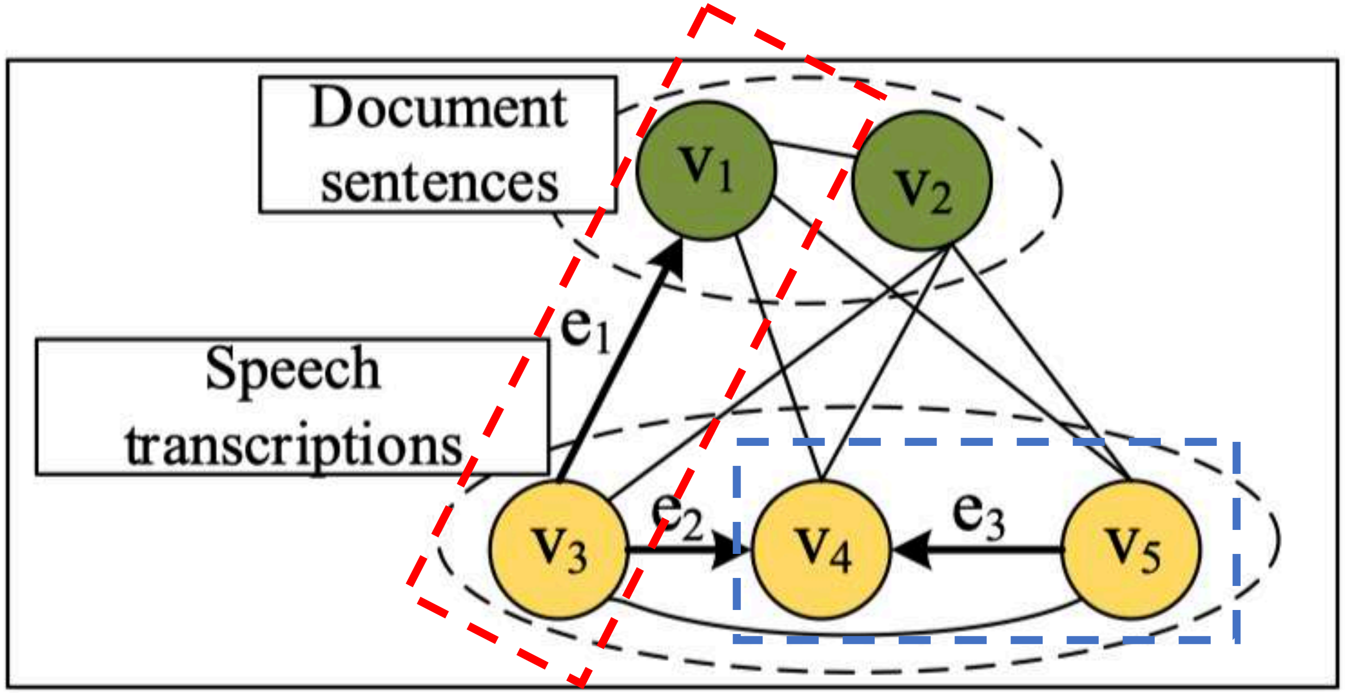
SCORE

- LexRank
  - Video
  - Audio
- Picture align score
  - Picture
  - Video

Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video *EMNLP17*

Document sentence is better then transcripts.

Video information



Hier audio score transcripts are more important

Audio information

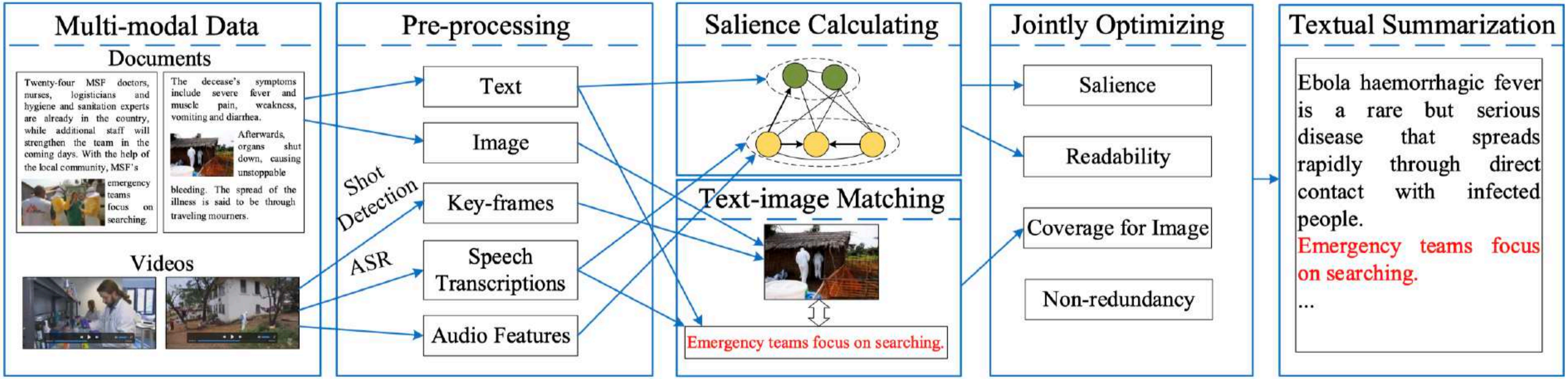
Modified LexRank

Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video *EMNLP17*

8月13日，深圳1份巴西进口冻鸡翅和西安1份厄瓜多尔进口冻虾的外包装先后被披露新冠病毒核酸检测结果呈阳性。 ♥

消费者可在购买时做好防护，在烹饪时应注意关键环节卫生，且将食材煮熟。但食品安全没有零风险，如果个人特别在意，可选择不吃。 ✖

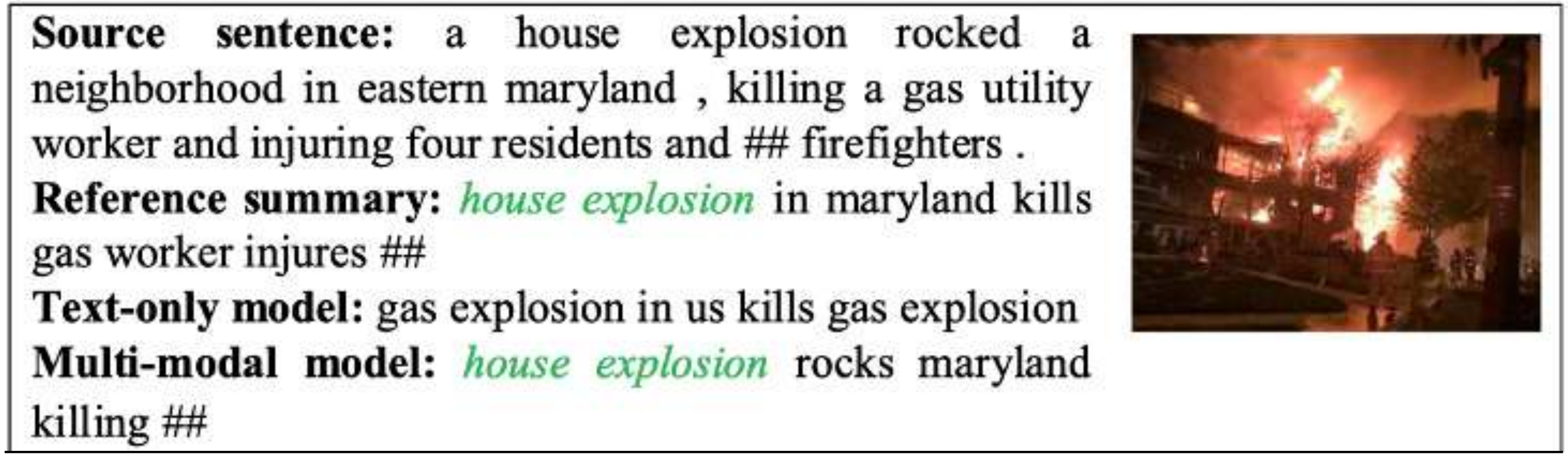
Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video *EMNLP17*


Multi-modal Data
Documents
Twenty-four MSF doctors, nurses, logisticians and hygiene and sanitation experts are already in the country, while additional staff will strengthen the team in the coming days. With the help of the local community, MSF's emergency teams focus on searching.
The decease's symptoms include severe fever and muscle pain, weakness, vomiting and diarrhea.
Afterwards, organs shut down, causing unstoppable bleeding. The spread of the illness is said to be through traveling mourners.
Videos
Pre-processing
Text
Image
Key-frames
Speech Transcriptions
Audio Features
Shot Detection
ASR
Salience Calculating
Text-image Matching
Emergency teams focus on searching.
Jointly Optimizing
Salience
Readability
Coverage for Image
Non-redundancy
Textual Summarization
Ebola haemorrhagic fever is a rare but serious disease that spreads rapidly through direct contact with infected people.
Emergency teams focus on searching.
...

- Dataset
  - 25 Chinese topics
  - 25 English topics
  - Each topic 20 documents
  - Human generate summaries

Multi-modal Sentence Summarization with Modality Attention and Image Filtering *IJCAI18*

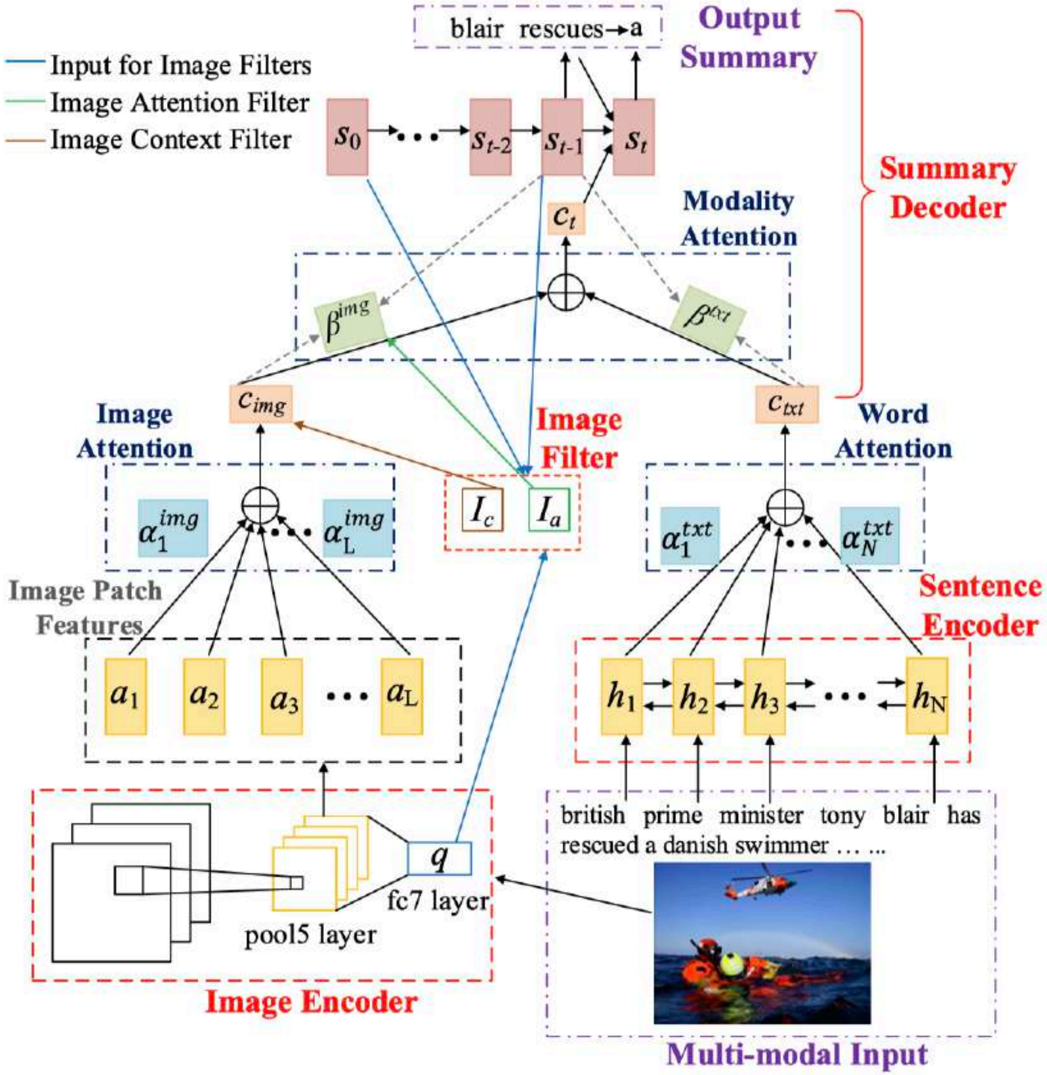
- **Task**

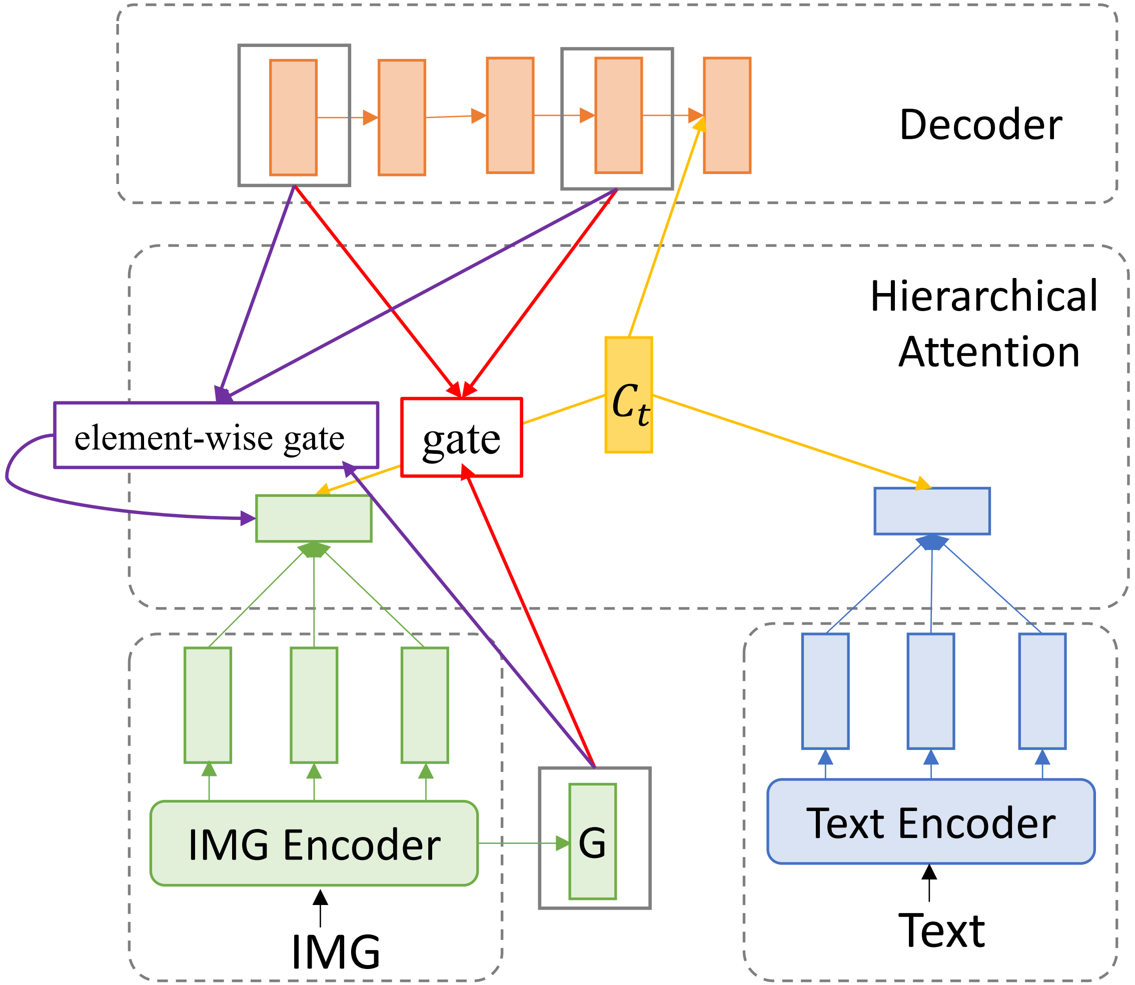


Source sentence: a house explosion rocked a neighborhood in eastern maryland , killing a gas utility worker and injuring four residents and ## firefighters .
Reference summary: *house explosion* in maryland kills gas worker injures ##
Text-only model: gas explosion in us kills gas explosion
Multi-modal model: *house explosion* rocks maryland killing ##

- **Dataset**
  - Gigaword + Image Search (top 5) + human selection (top 1)
  - train(62000) valid(2000) test(2000)

Multi-modal Sentence Summarization with Modality Attention and Image Filtering *IJCAI18*

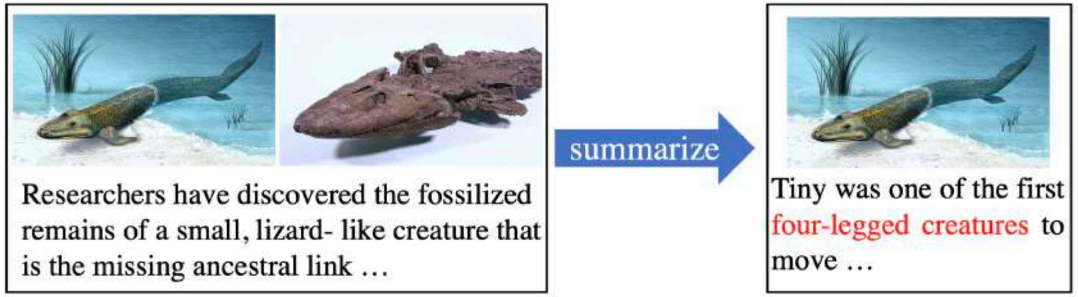




MSMO

MSMO: Multimodal Summarization with Multimodal Output *EMNLP18*

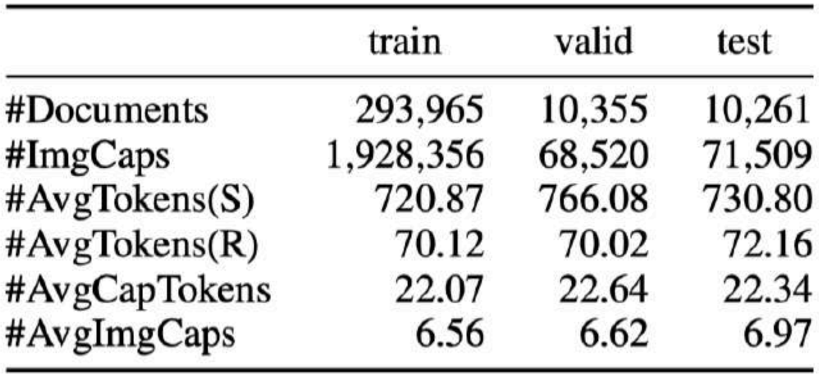
- Task ：MSMO



Researchers have discovered the fossilized remains of a small, lizard- like creature that is the missing ancestral link …

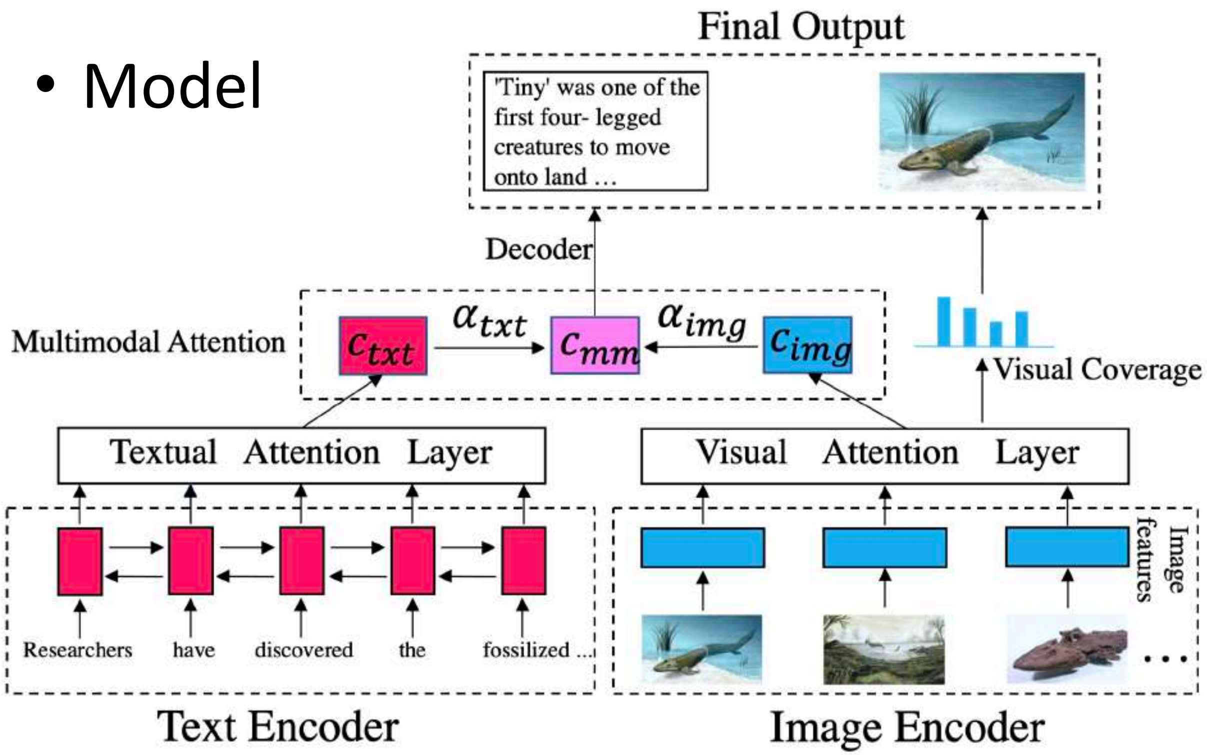
Tiny was one of the first four-legged creatures to move …

- Dataset: From Daily Mail, annotators select up to three images as ref imgs.



	train	valid	test
#Documents	293,965	10,355	10,261
#ImgCaps	1,928,356	68,520	71,509
#AvgTokens(S)	720.87	766.08	730.80
#AvgTokens(R)	70.12	70.02	72.16
#AvgCapTokens	22.07	22.64	22.34
#AvgImgCaps	6.56	6.62	6.97

- Model

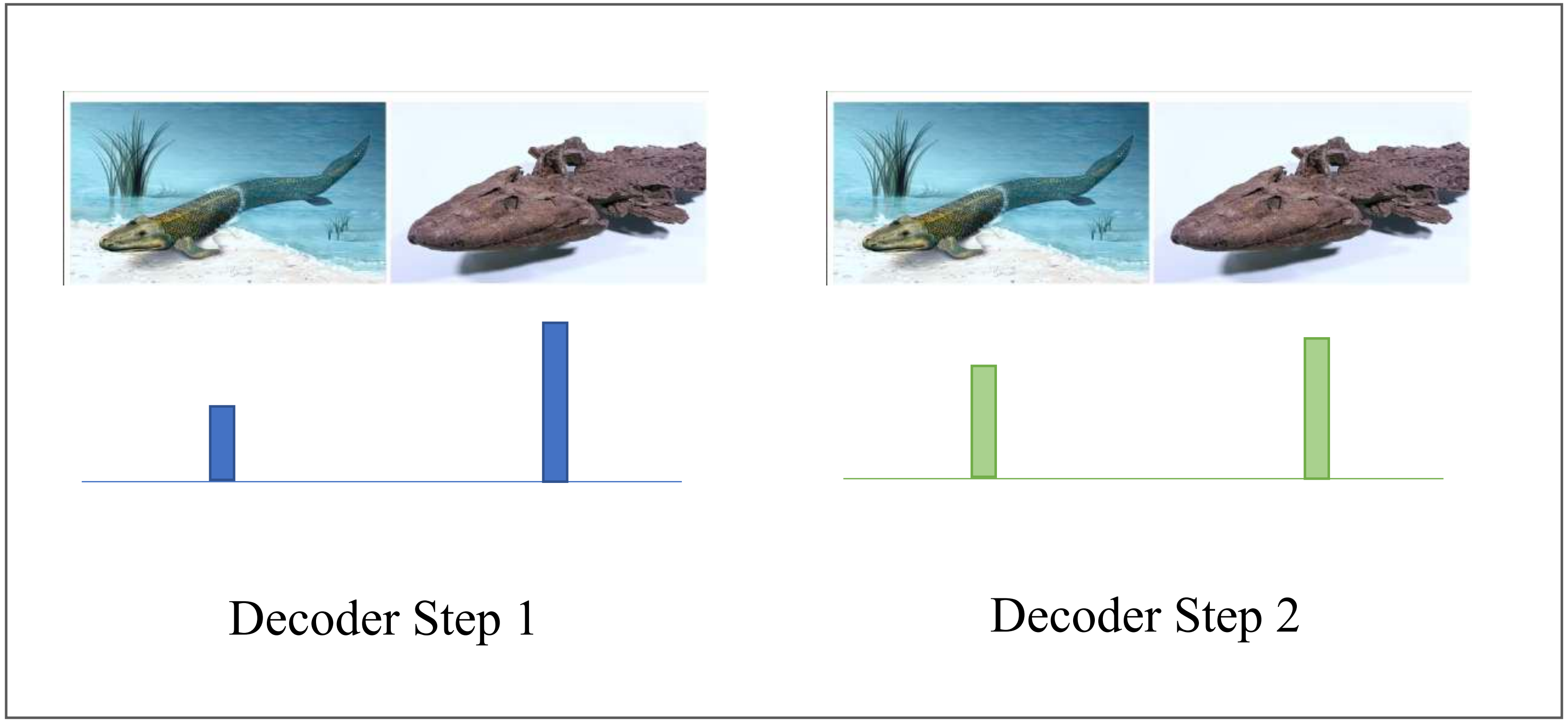


- Evaluation: MMAE

ROUGE+ Image precision + Image-Text Relevance

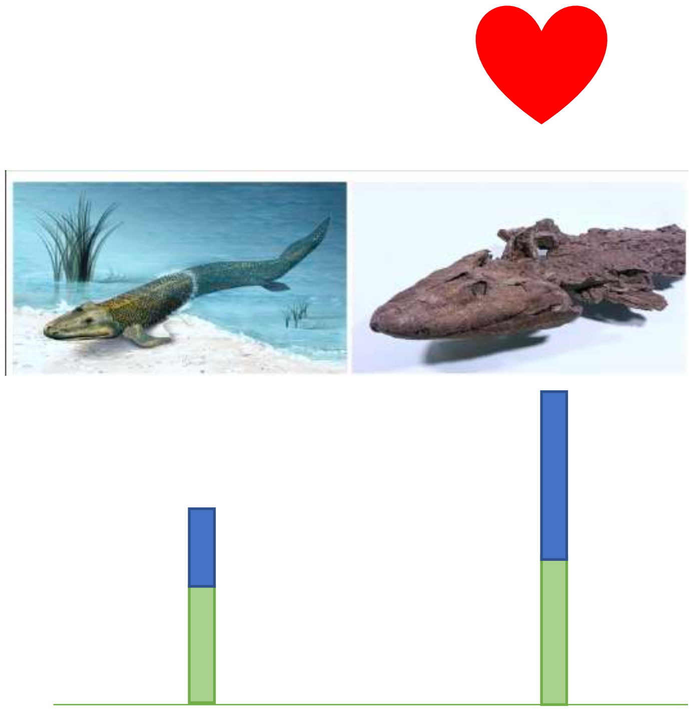
MSMO: Multimodal Summarization with Multimodal Output *EMNLP18*

- Visual Coverage



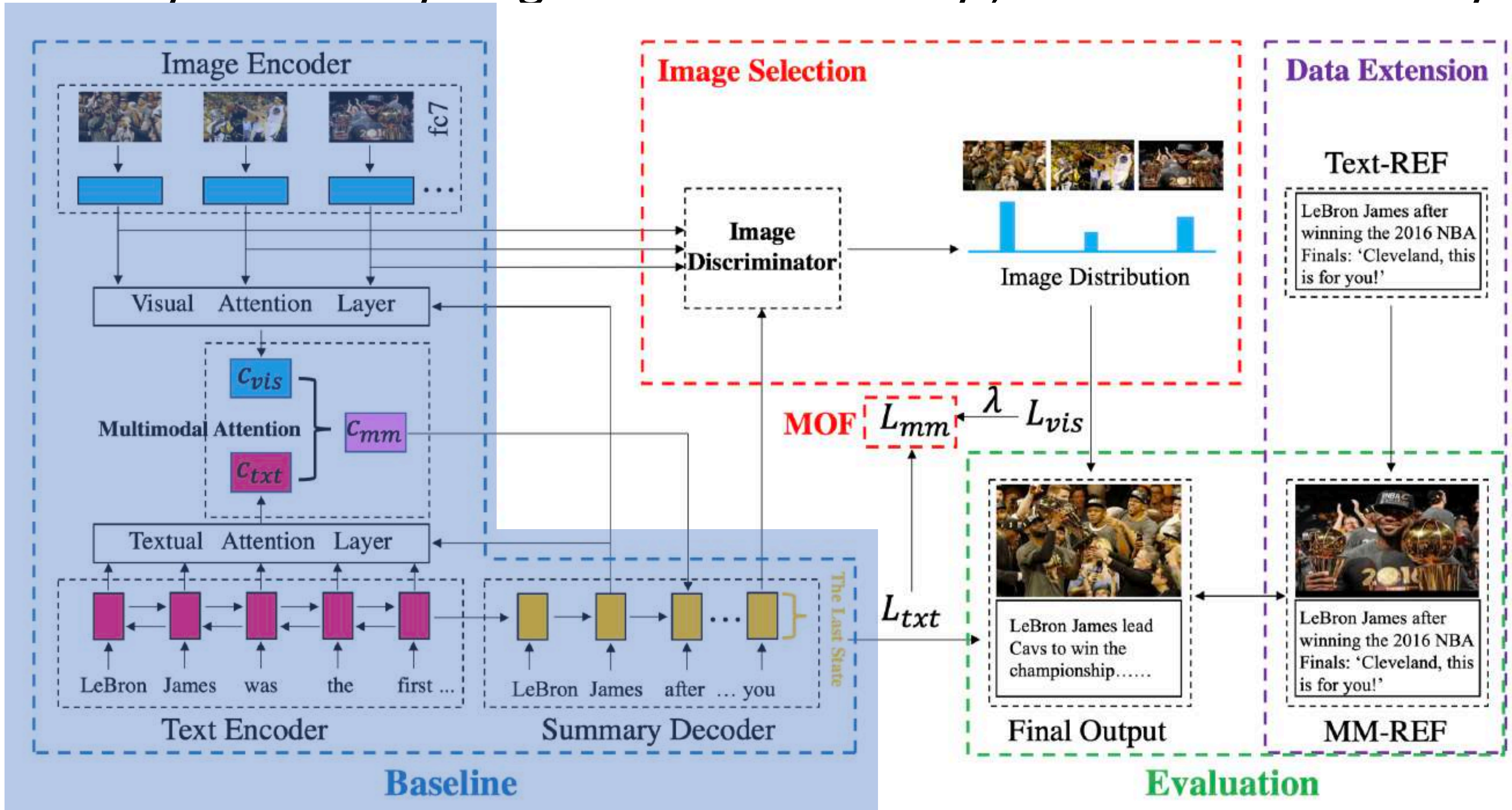
Decoder Step 1

Decoder Step 2



Multimodal Summarization with Guidance of Multimodal Reference *AAAI20*

- Problem
  - Only trained by target of text modality , lead to the modality-bias problem.

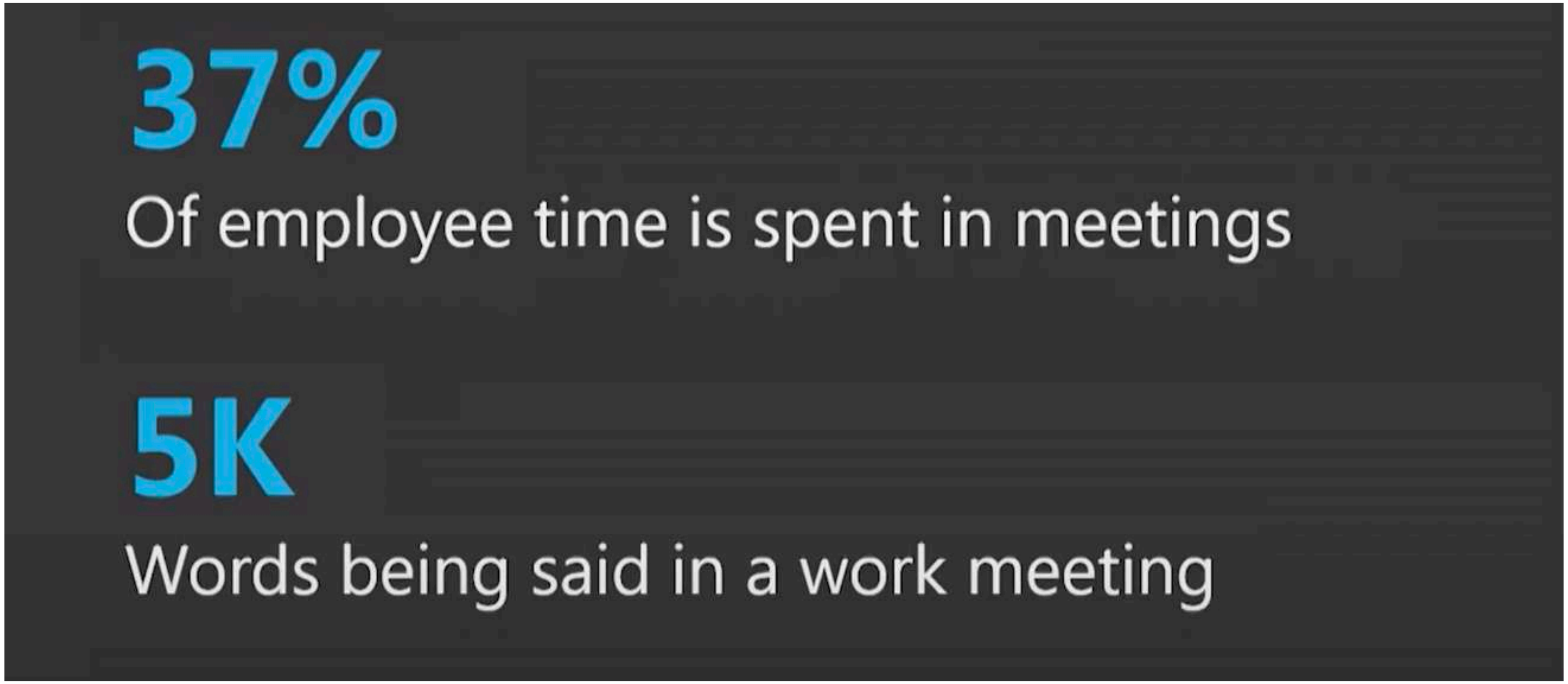


Data Extension

- ROUGE-ranking.
- Order-ranking.

Meeting

Communication Load is Overwheming



37%

Of employee time is spent in meetings

5K

Words being said in a work meeting

https://www.youtube.com/watch?v=Pd5Y7pfbU4w&list=LLFh1maHZbOfktmNXYft0-rA&index=4&t=0s

Multimodal Summarization of Meeting Recordings *ICME2003*

- **Why summarize meetings?**
  - Frequently, meetings last hours and have many lengthy boring parts.
- **Why use multi-modal information?**
  - Language analysis-based abstraction techniques may not be sufficient to capture significant visual and audio events in a meeting, such as a person entering the room to join the meeting or an emotional discussion.
- Meeting sequences generally contain a very small amount of visual activity. Also, because our meeting recorder captures omni-directional video, there is no camera motion and therefore no scene breaks.

Multimodal Summarization of Meeting Recordings *ICME2003*

Audio

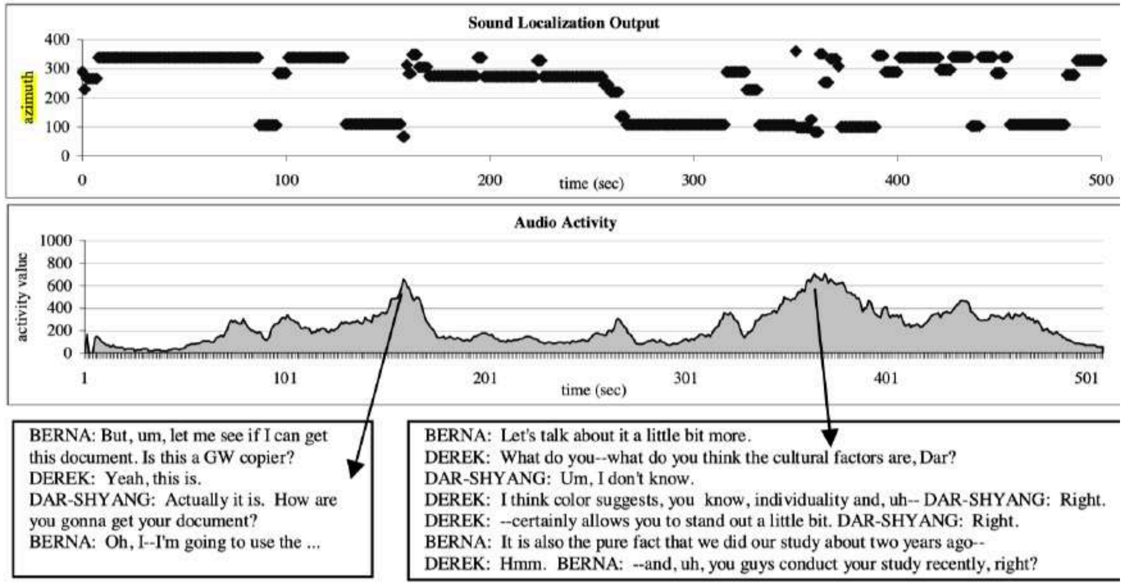
- sound localization output
- magnitude of the audio signal

Video

- luminance changes in a small window

Text

- IF-IDF



BERNA: But, um, let me see if I can get this document. Is this a GW copier?
DEREK: Yeah, this is.
DAR-SHYANG: Actually it is. How are you gonna get your document?
BERNA: Oh, I--I'm going to use the ...

BERNA: Let's talk about it a little bit more.
DEREK: What do you--what do you think the cultural factors are, Dar?
DAR-SHYANG: Um, I don't know.
DEREK: I think color suggests, you know, individuality and, uh-- DAR-SHYANG: Right.
DEREK: --certainly allows you to stand out a little bit. DAR-SHYANG: Right.
BERNA: It is also the pure fact that we did our study about two years ago--
DEREK: Hmm. BERNA: --and, uh, you guys conduct your study recently, right?

Figure 1. Sound localization and audio activity output for a staff meeting recording.

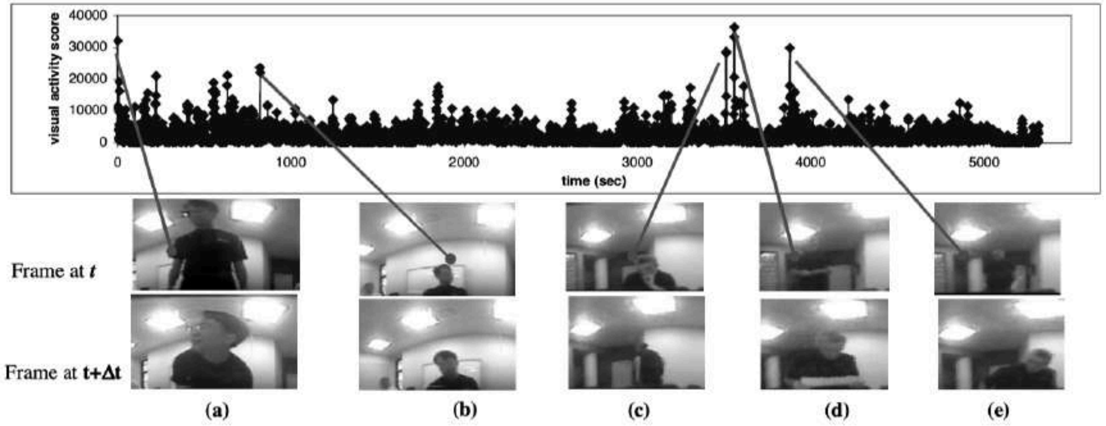
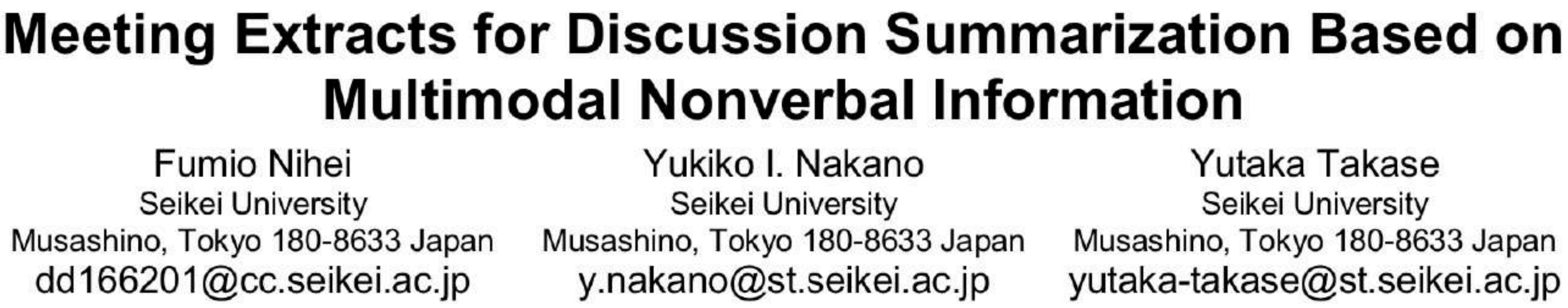


Figure 2. Examples of high visual activity scores corresponding to significant visual events.

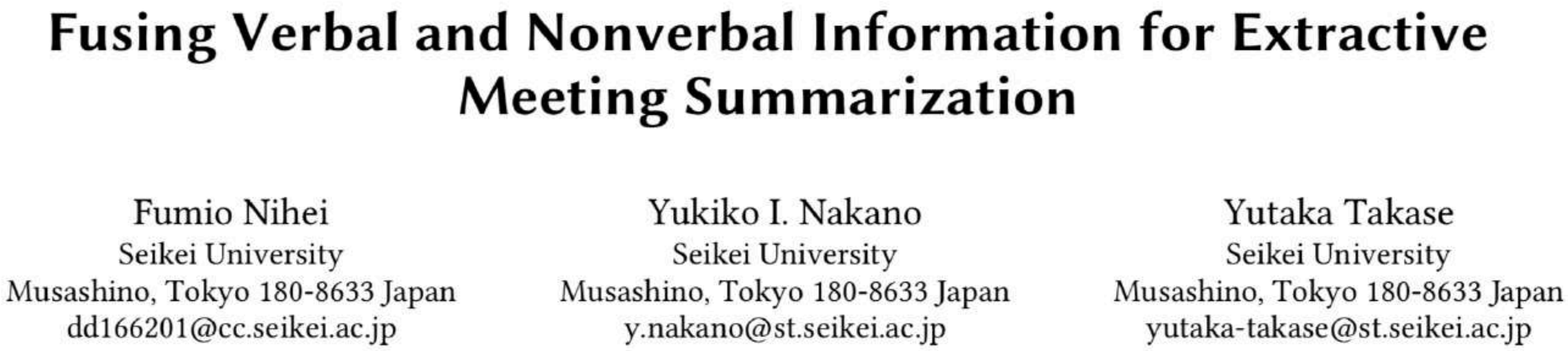


Meeting Extracts for Discussion Summarization Based on Multimodal Nonverbal Information

Fumio Nihei
Seikei University
Musashino, Tokyo 180-8633 Japan
dd166201@cc.seikei.ac.jp

Yukiko I. Nakano
Seikei University
Musashino, Tokyo 180-8633 Japan
y.nakano@st.seikei.ac.jp

Yutaka Takase
Seikei University
Musashino, Tokyo 180-8633 Japan
yutaka-takase@st.seikei.ac.jp

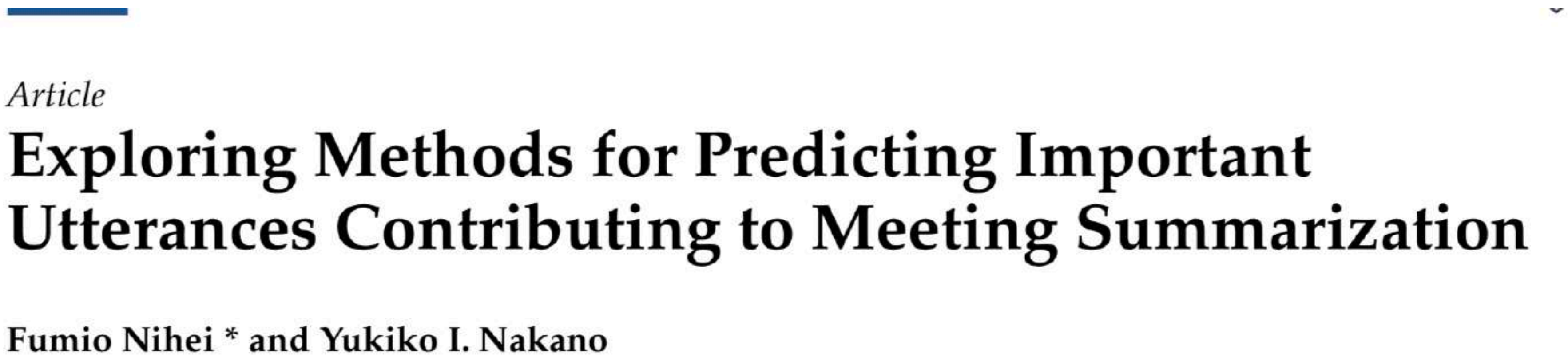


Fusing Verbal and Nonverbal Information for Extractive Meeting Summarization

Fumio Nihei
Seikei University
Musashino, Tokyo 180-8633 Japan
dd166201@cc.seikei.ac.jp

Yukiko I. Nakano
Seikei University
Musashino, Tokyo 180-8633 Japan
y.nakano@st.seikei.ac.jp

Yutaka Takase
Seikei University
Musashino, Tokyo 180-8633 Japan
yutaka-takase@st.seikei.ac.jp

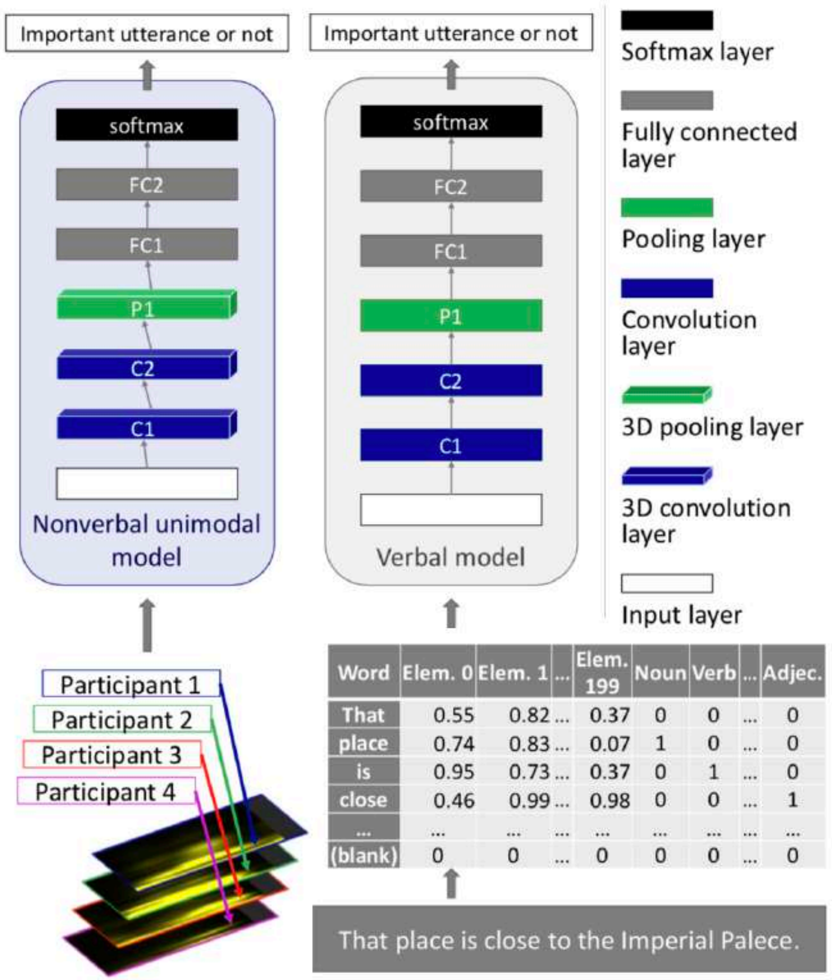


Article

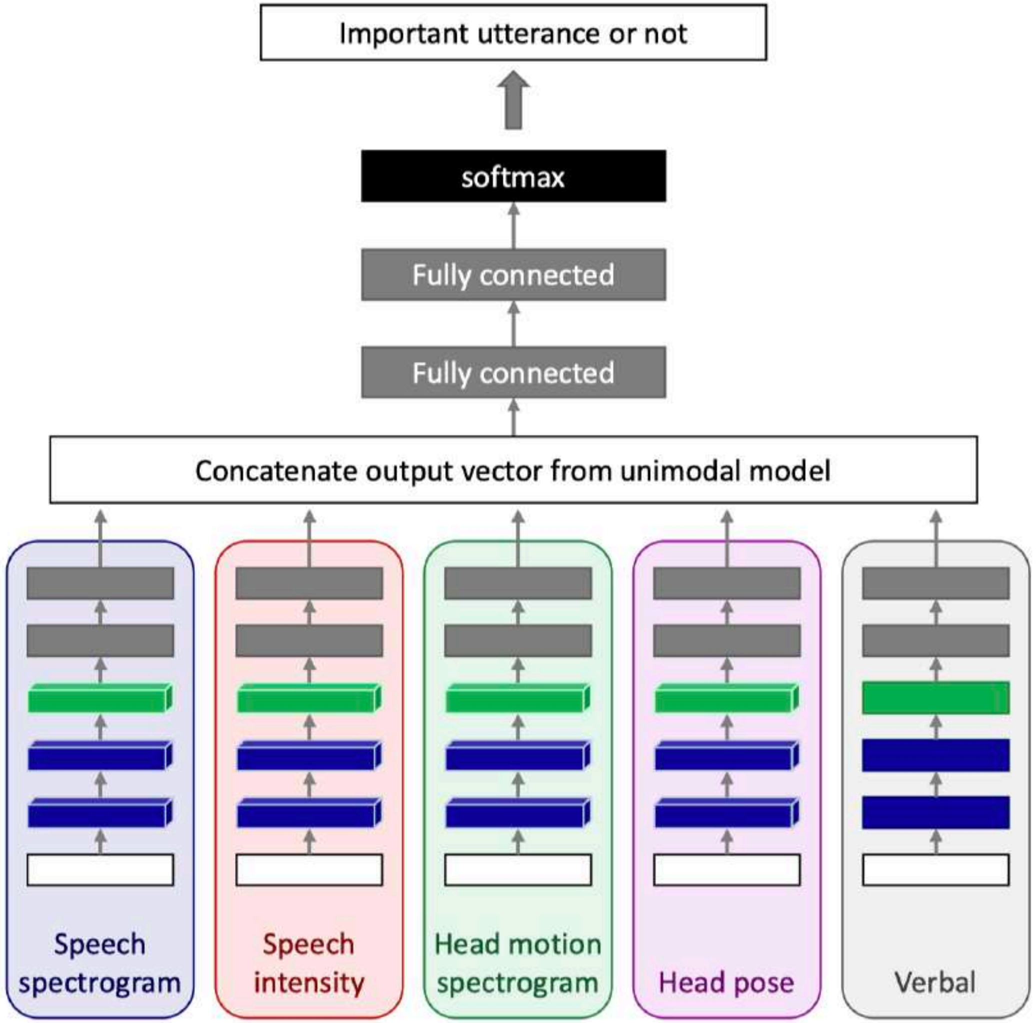
Exploring Methods for Predicting Important Utterances Contributing to Meeting Summarization

Fumio Nihei * and Yukiko I. Nakano

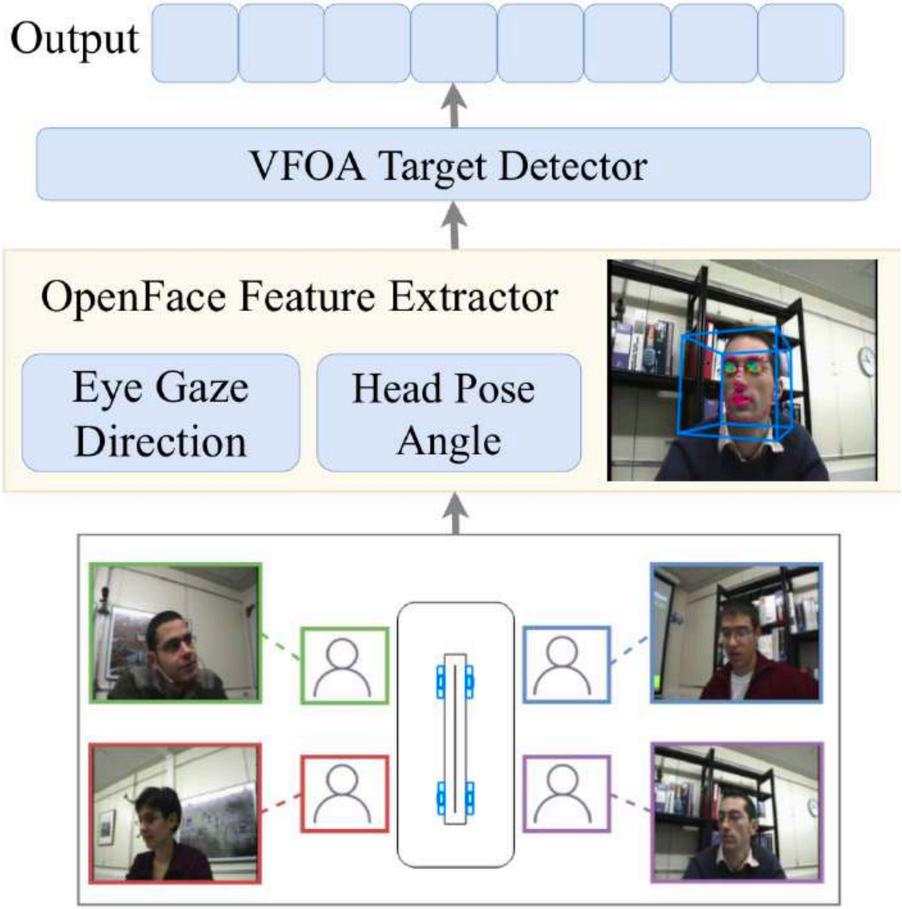
Fusing Verbal and Nonverbal Information for Extractive Meeting Summarization *GIFT18*



Word	Elem. 0	Elem. 1	...	Elem. 199	Noun	Verb	...	Adjec.
That	0.55	0.82	...	0.37	0	0	...	0
place	0.74	0.83	...	0.07	1	0	...	0
is	0.95	0.73	...	0.37	0	1	...	0
close	0.46	0.99	...	0.98	0	0	...	1
...	...	...	...	...	...	...	...	...
(blank)	0	0	...	0	0	0	...	0



Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization *ACL19*



- Input:5-dimensional feature vector.
  - head pose angle (*roll*, *pitch* and *yaw*)
  - eye gaze direction vector (*azimuth* and *elevation*)
- Trained on the VFOA annotation
  - p0,...,p|P|, *table*, *whiteboard*, *projection screen* and *unknown*
- For utterance u_i , the VFOA vector $f_i \in \mathbb{R}^{|P|*|F|}$

Conclusion

- Model is simple. [Encoder + Decoder + Hierarchical attention]
- Modality Interaction is less.
- Priori knowledge is important.
- Multimodal output (text, pic, video) can help a larger set of people.
- Data privacy is not considered in Meeting Summarization.

Thanks!