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

Haoran Li[1,2], Junnan Zhu[1,2], Cong Ma[1,2], Jiajun Zhang[1,2] and Chengqing Zong[1,2,3]
[1] National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
{haoran.li, junnan.zhu, cong.ma, jjzhang, cqzong}@nlpr.ia.ac.cn

August 16th, 2017

Outline

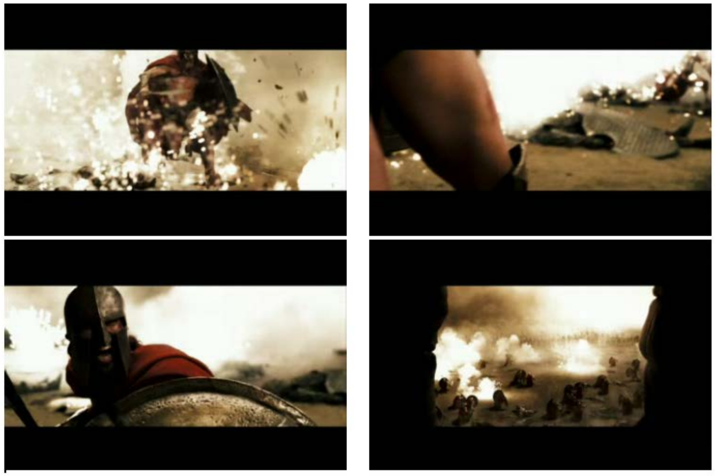
1. Introduction
   - What is Multi-modal?
   - What is Asynchronous?
2. Model
   - Model overview
   - Salience for Text
   - Coverage for Visual
   - Objective Functions
3. Experiment
   - Dataset
   - Experimental Results
4. Conclusion and Future Works

Outline

1. Introduction
   - What is Multi-modal?
   - What is Asynchronous?
2. Model
   - Model overview
   - Salience for Text
   - Coverage for Visual
   - Objective Functions
3. Experiment
   - Dataset
   - Experimental Results
4. Conclusion and Future Works

Introduction

- What is Multi-modal?



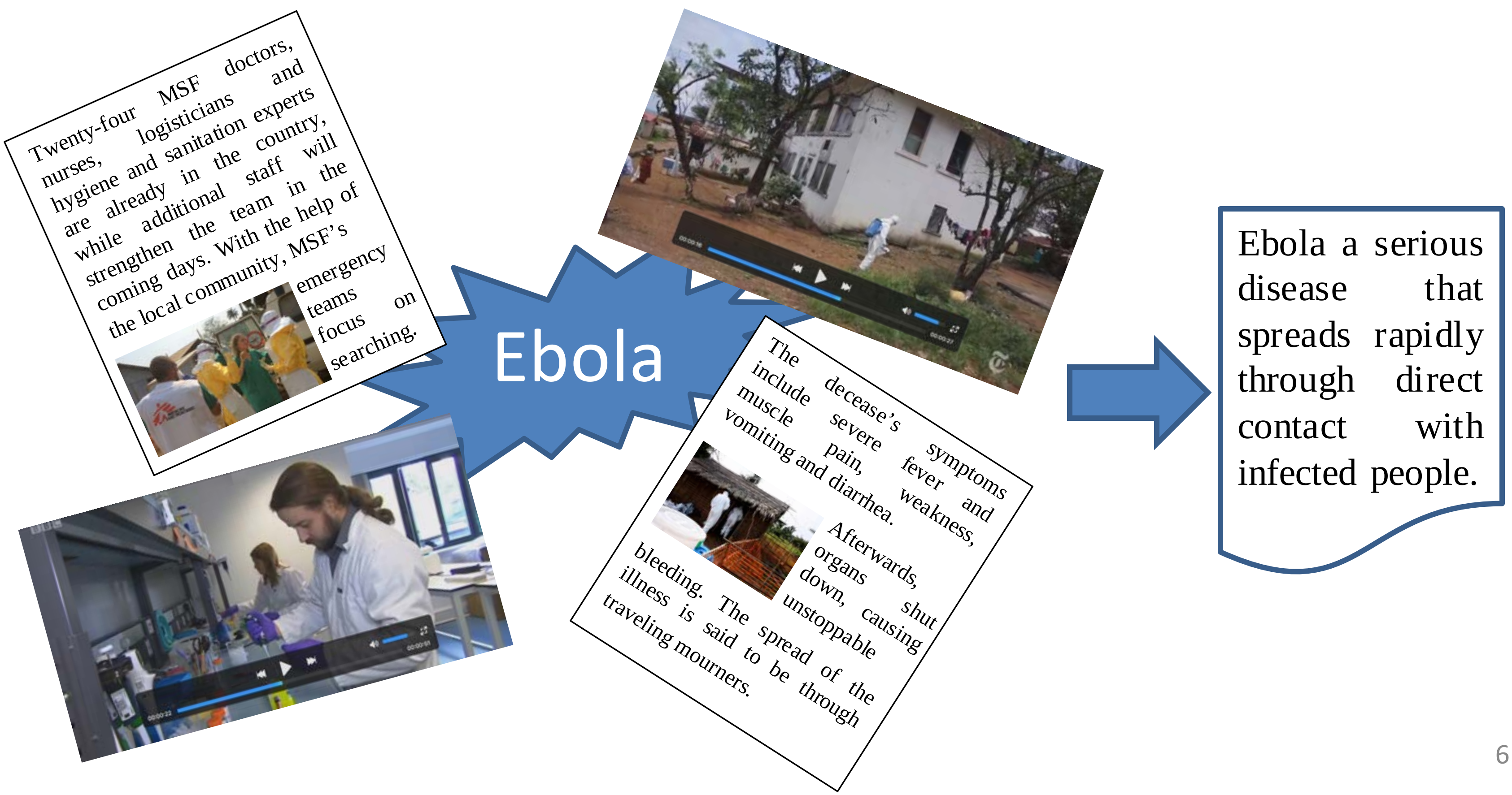
Introduction

- What is Asynchronous?
  - Synchronous V.S. Asynchronous?
  - Synchronous: images are paired with text descriptions, videos are paired with subtitles, …

Movie Summarization [Evangelopoulos et al. TMM 2013]

Introduction

- What is Asynchronous?
  - Asynchronous

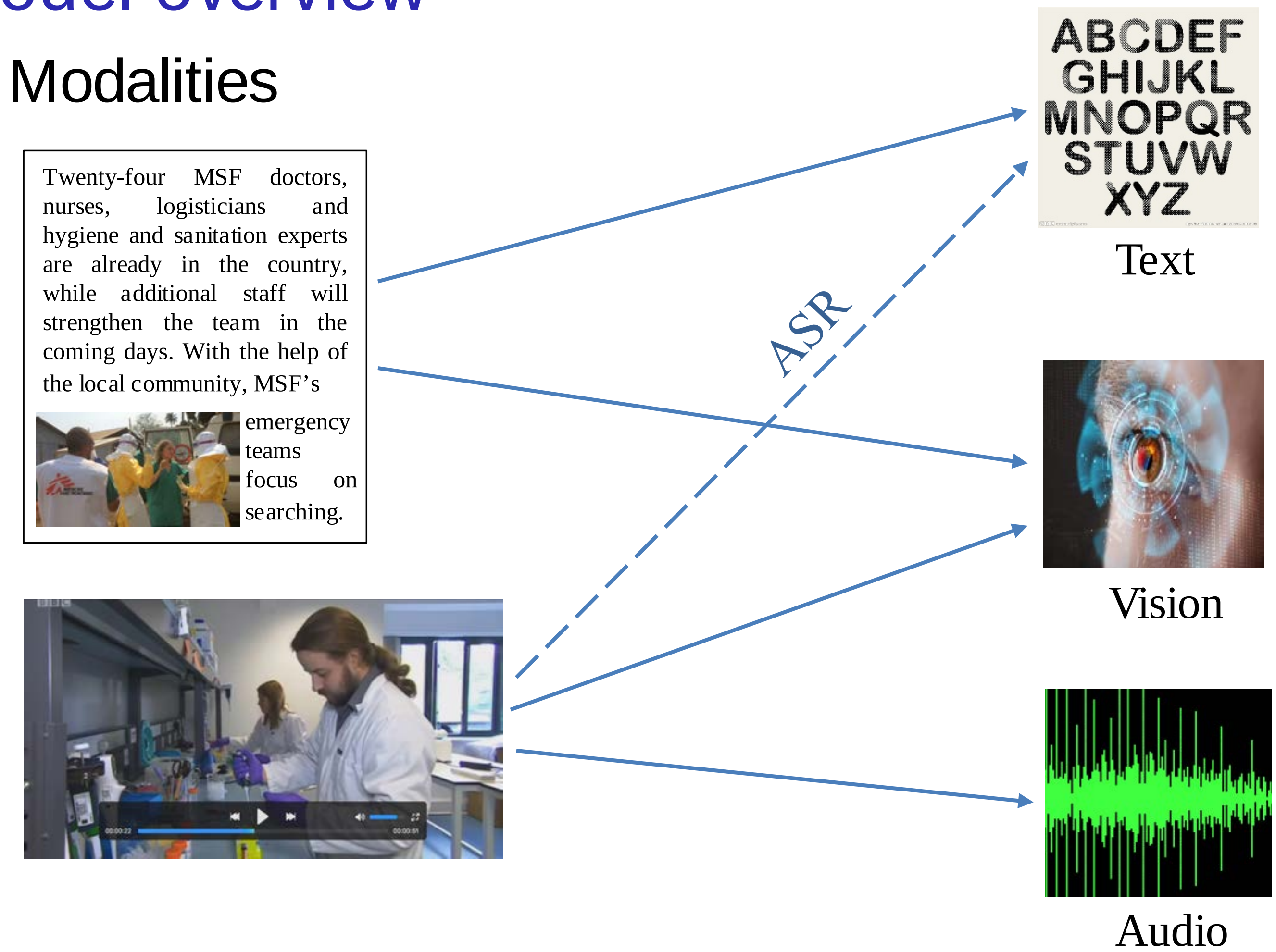


Outline

1. Introduction
   - What is Multi-modal?
   - What is Asynchronous?
2. Model
   - Model overview
   - Salience for Text
   - Coverage for Visual
   - Objective Functions
3. Experiment
   - Dataset
   - Experimental Results
4. Conclusion and Future Works

Model

- Model overview
  - Modalities



Model

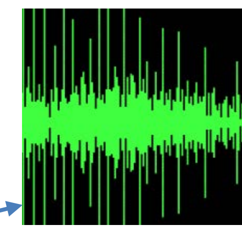
➢ Model overview

- Bridge the semantic gaps between multi-modal content.

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.

Text

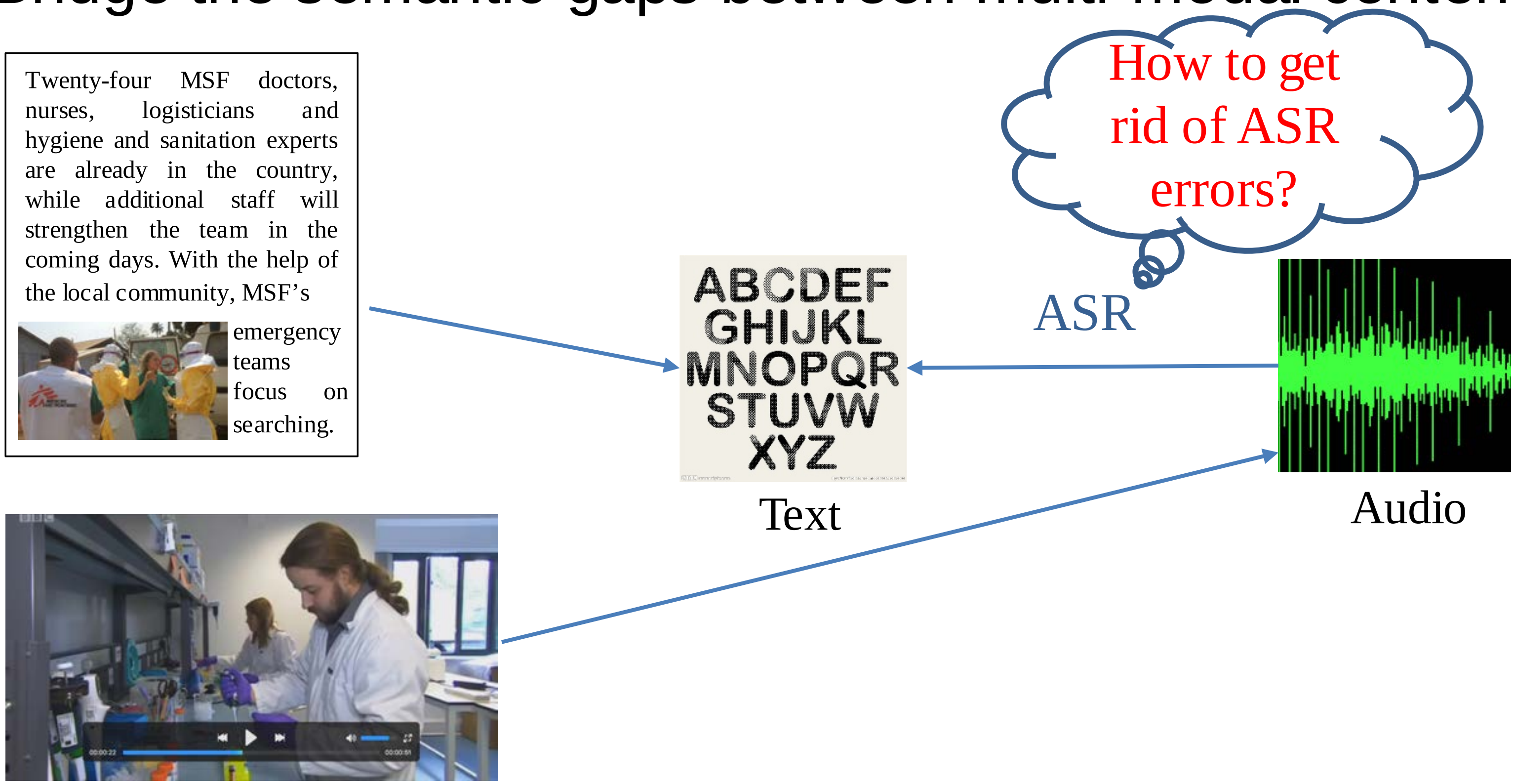


Audio

Model

➢ Model overview

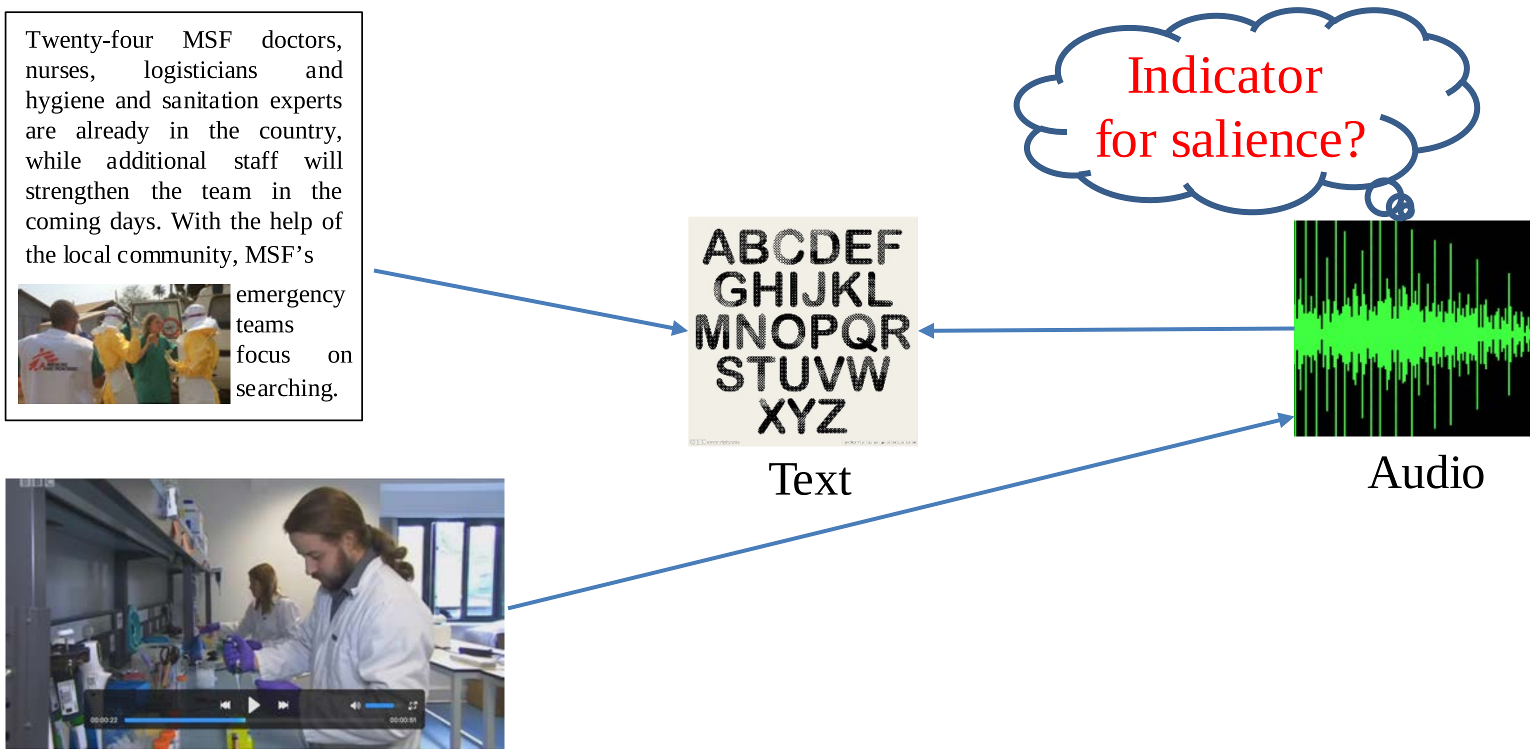
- Bridge the semantic gaps between multi-modal content.



Model

➢ Model overview

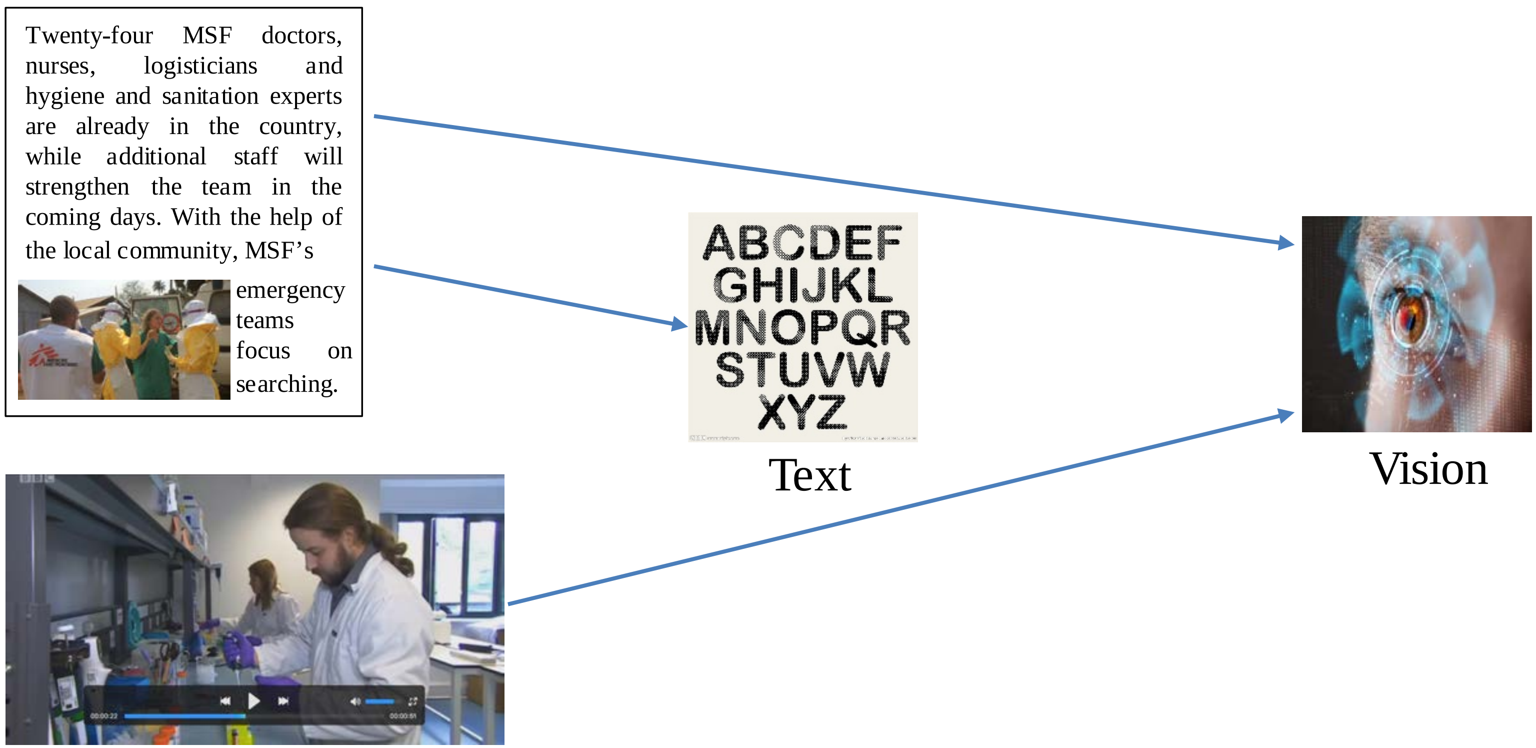
- Bridge the semantic gaps between multi-modal content.



Model

➢ Model overview

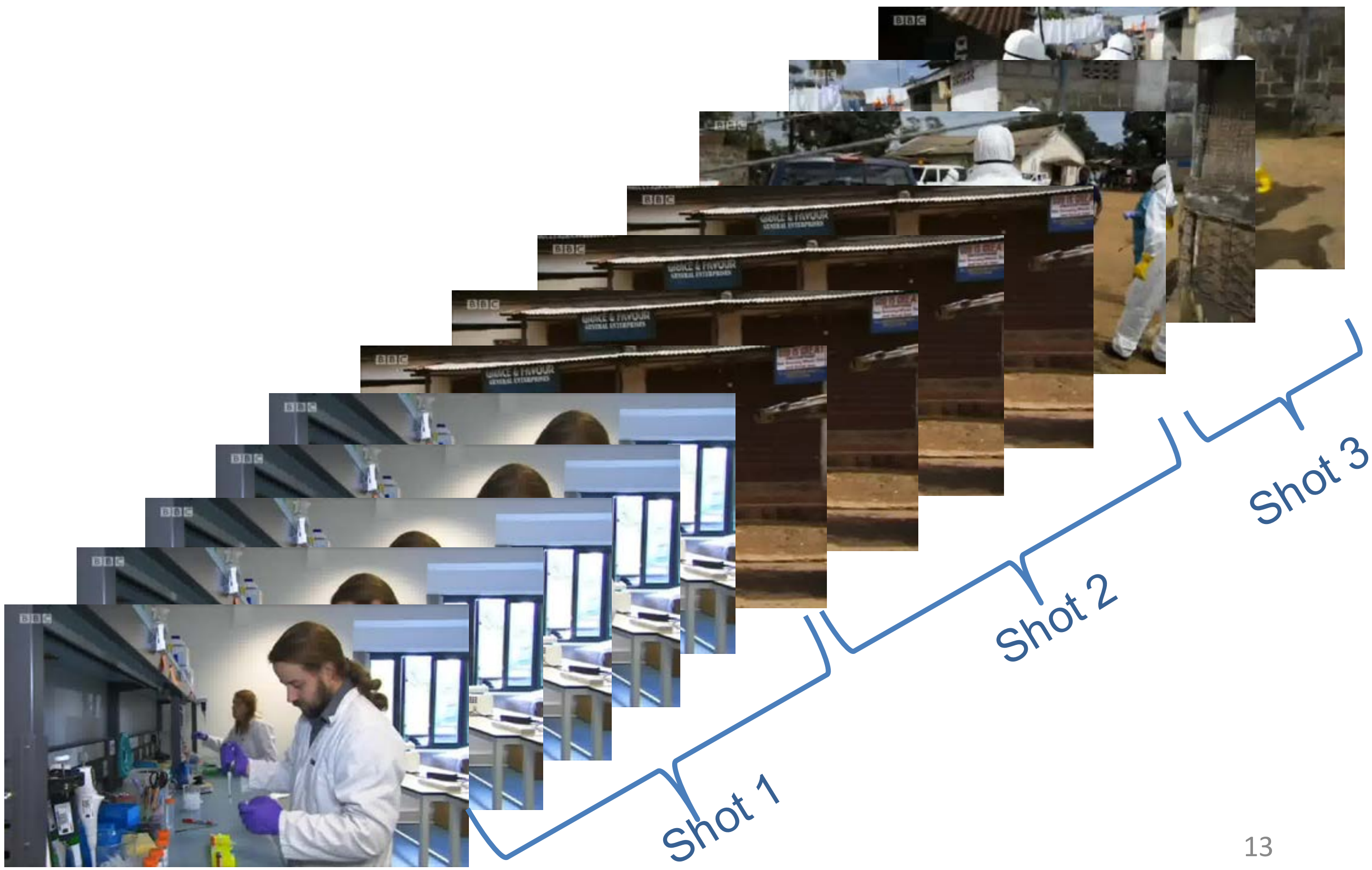
- Bridge the semantic gaps between multi-modal content.



Model

➢ Model overview

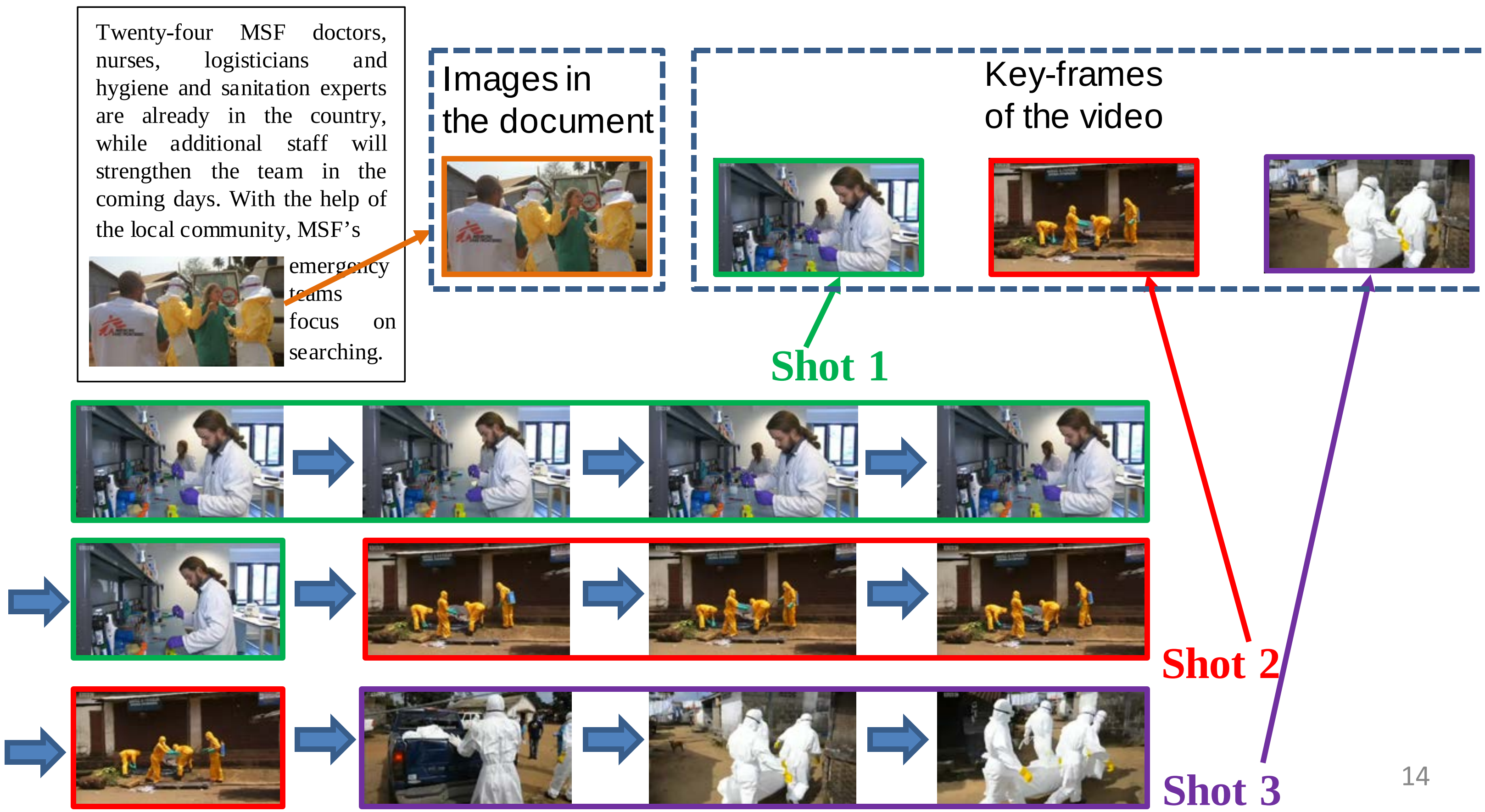
- Bridge the semantic gaps between multi-modal content.



Model

➢ Model overview

- Bridge the semantic gaps between multi-modal content.



Model

- Model overview
  - Bridge the semantic gaps between multi-modal content.

Emergency teams focus on searching .

Medicins Sans Frontieres (MSF) launched an emergency medical intervention in the West African.

Doctors and nurses fight against the deadly Ebola outbreak in Guinea .

New drug therapies are being evaluated.

There is no cure for the virus, and no vaccine which can protect against it.

Model

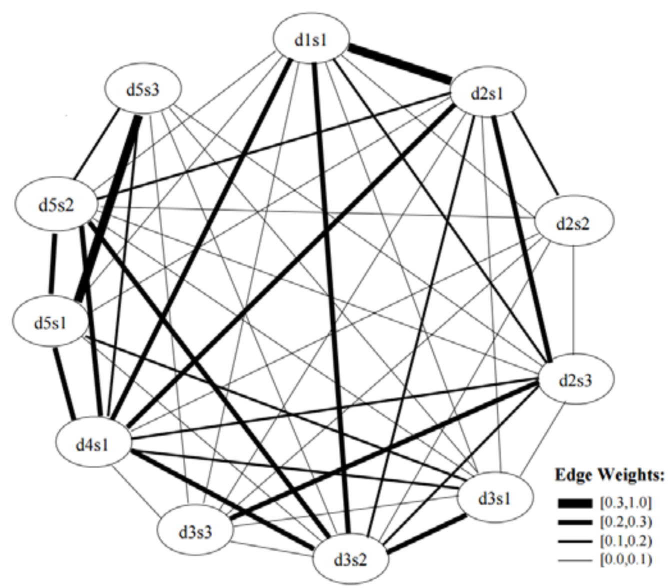
- Model overview
  - Document summarization: salience, non-redundancy
  - For our task: readability, coverage for the visual information
    - Readability: get rid of the errors introduced by ASR.
    - Visual information: indicator for event highlights

Model

- Salience for Text (Including document sentences and speech transcriptions)

  - $$Sa(t_i) = \mu \sum_j Sa(t_j) \cdot M_{ji} + \frac{1-\mu}{N}$$

    $$M_{ji} = sim(t_j, t_i)$$



LexRank [Erkan and Radev JAIR 2004]

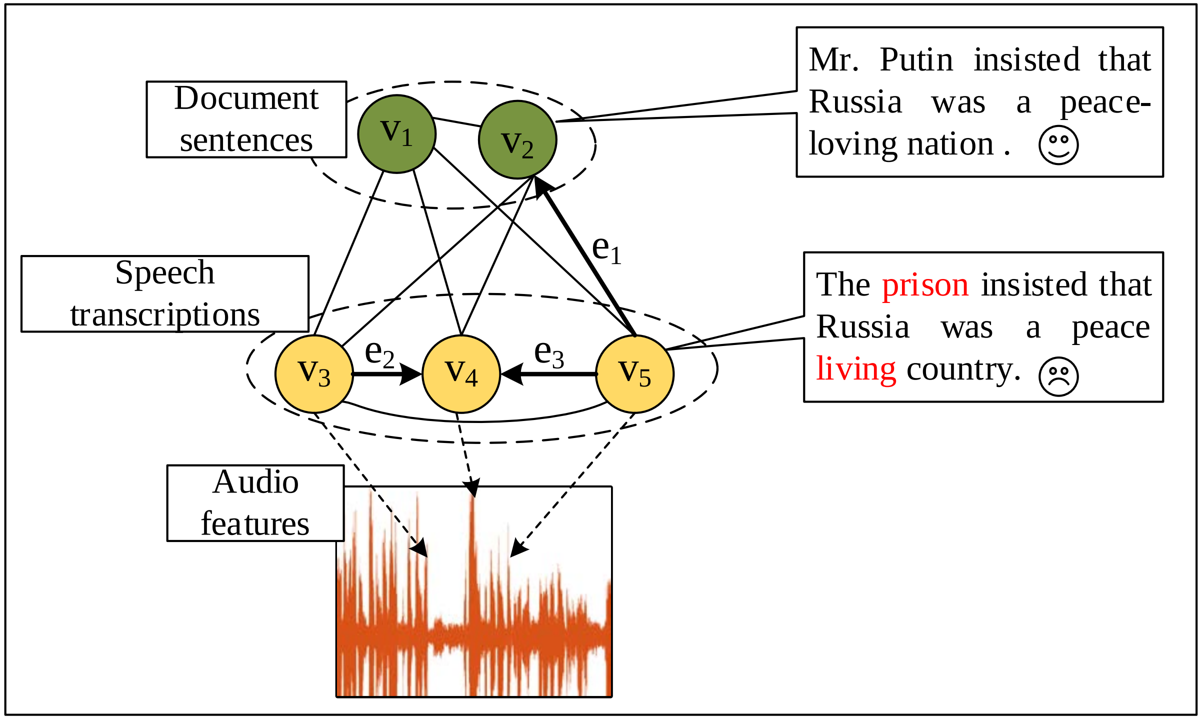
Model

- Salience for Text
  - LexRank algorithm with guidance strategies
    - Readability guidance strategy: speech transcriptions recommend the corresponding document sentences
    - Audio Guidance Strategies: Some audio features can indicate salience or readability, including audio power and audio magnitude and acoustic confidence

Model

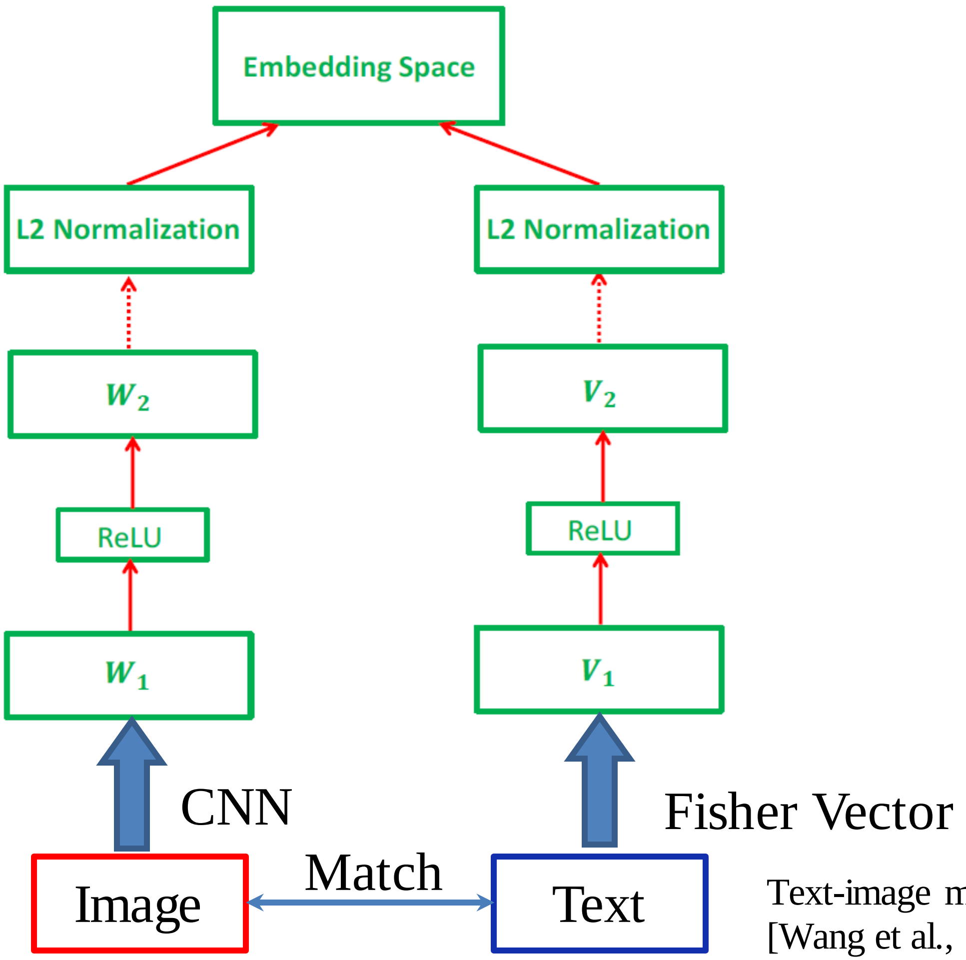
➢ Salience for Text

- LexRank algorithm with guidance strategies



Model

➢ Coverage for Visual



Text-image matching model
[Wang et al., CVPR 2016]

Model

➢ Coverage for Visual

>>A man in a tan jacket at the gas station pumping gas .
>>A man dressed in tan pumps gas .

Flickr30K and COCO Dataset

>>Whole streets and squares in the capital of more than 1 million people were covered in rubble .

Our Dataset

Model

Objective Functions

- Salience for Text

$$\mathcal{F}_s(S) = \sum_{t_i \in S} Sa(t_i) - \frac{\lambda_s}{|S|} \sum_{t_i, t_j \in S} sim(t_i, t_j)$$

- Coverage for Visual

$$\mathcal{F}_c(S) = \sum_{p_i \in I} \underline{Im(p_i)}\,\underline{b_i}$$

1: p_i is covered by the summary;
0: p_i is not covered

length ratio between the shot pi belonging to and the whole video.

- Considering all the modalities

$$\mathcal{F}_m(S) = \frac{1}{M_s} \sum_{t_i \in S} Sa(t_i) + \frac{1}{M_c} \sum_{p_i \in I} Im(p_i) b_i - \frac{\lambda_m}{|S|} \sum_{i,j \in S} sim(t_i, t_j)$$

Outline

1. Introduction
   - What is Multi-modal?
   - What is Asynchronous?
2. Model
   - Model overview
   - Salience for Text
   - Coverage for Visual
   - Objective Functions
3. Experiment
   - Dataset
   - Experimental Results
4. Conclusion and Future Works

Experiments

- Dataset
  - 50 news topics in the most recent five years, 25 in English and 25 in Chinese.
  - 20 topics for each language as a test set, 5 as a development set.
  - 20 documents and 5-10 videos for each topic.
  - 3 hand-annotated reference summaries for each topic.

	#Sentence	#Word	#Shot	Video Length
English	492.1	12,104.7	47.2	197s
Chinese	402.1	9,689.3	49.3	207s

Table 1: Corpus statistics.

Experiments

- Dataset

English	(1) Nepal earthquake (2) Terror attack in Paris (3) Train derailment in India (4) Germanwings crash (5) Refugee crisis in Europe
Chinese	(6) “东方之星”客船翻沉 (“Oriental Star”passenger ship sinking) (7) 银川公交大火 (The bus fire in Yinchuan) (8) 香港占中 (Occupy Central in HONG KONG) (9) 李娜澳网夺冠 (Li Na wins Australian Open) (10) 抗议“萨德”反导系统 (Protest against “THAAD”anti-missile system)

Table 2: Examples of news topics.

Experiments

- Comparative Methods
  - Text only
  - Text + audio
  - Text + audio + guide
  - Image caption
  - Image caption match
  - Image alignment
  - Image match

Experiments

➢ Experimental Results

Method	R-1	R-2	R-SU4
Text only	0.422	0.114	0.166
Text + audio	0.422	0.109	0.164
Text + audio + guide	0.440	0.117	0.171
Image caption	0.435	0.111	0.167
Image caption match	0.429	0.115	0.166
Image alignment	0.409	0.082	0.082
Image match	**0.442**	**0.133**	**0.187**

Table 3: Experimental results (F-score) for English.

Experiments

➢ Experimental Results

Method	R-1	R-2	R-SU4
Text only	0.409	0.113	0.167
Text + audio	0.407	0.111	0.166
Text + audio + guide	0.411	0.115	**0.173**
Image caption match	0.381	0.092	0.149
Image alignment	0.368	0.096	0.143
Image match	**0.414**	**0.125**	**0.173**

Table 4: Experimental results (F-score) for Chinese.

Experiments

➢ Experimental Results

	Method	Read	Inform
English	Text only	3.72	3.28
	Text + audio	3.08	3.44
	Text + audio + guide	3.68	3.64
	Image match	3.67	3.83
	Reference	4.52	4.36
Chinese	Text only	3.64	3.40
	Text + audio	3.16	3.48
	Text + audio + guide	3.60	3.72
	Image match	3.62	3.92
	Reference	4.88	4.84

Table 5: Manual summary quality evaluation. “Read” denotes “Readability” and “Inform” denotes “informativeness”.

Experiments

➢ Experimental Results

Ramchandra Tewari , a passenger who suffered a head injury , said he was asleep when he was suddenly flung to the floor of his coach . The impact of the derailment was so strong that one of the coaches landed on top of another , crushing the one below , said Brig. Anurag Chibber , who was heading the army 's rescue team . `` We fear there could be many more dead in the lower coach , " he said , adding that it was unclear how many people were in the coach . Kanpur is a major railway junction , and hundreds of trains pass through the city every day . `` I heard a loud noise , " passenger Satish Mishra said . Some railway officials told local media they suspected faulty tracks caused the derailment . Fourteen cars in the 23-car train derailed , Modak said . We do n't expect to find any more bodies , " said Zaki Ahmed , police inspector general in the northern city of Kanpur , about 65km from the site of the crash in Pukhrayan . When they tried to leave through one of the doors , they found the corridor littered with bodies , he said . The doors would n't open but we somehow managed to come out . But it has a poor safety record , with thousands of people dying in accidents every year , including in train derailments and collisions . By some analyst estimates , the railways need 20 trillion rupees ($ 293.34 billion) of investment by 2020 , and India is turning to partnerships with private companies and seeking loans from other countries to upgrade its network .

Figure 2: An example of generated summary for the news topic “India train derailment”.

Outline

1. Introduction
   - What is Multi-modal?
   - What is Asynchronous?
2. Model
   - Model overview
   - Salience for Text
   - Coverage for Visual
   - Objective Functions
3. Experiment
   - Dataset
   - Experimental Results
4. Conclusion and Future Works

Conclusion and Future Works

➢ Conclusion

- We addresses an asynchronous MMS task, namely, how to use related text, audio and video information to generate a textual summary.
- We design guidance strategies to selectively use the transcription of audio leading to more readable and informative summaries.
- We investigate various approaches to identify the relevance between the image and texts, and find that the image match model performs best.

Conclusion and Future Works

- Future Works
  - Make a distinction between document sentences and speech transcriptions.
  - Explore more correlations between text and vision.
  - Enlarge our dataset, specifically to collect more videos.

References

1. Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Transactions on Multimedia, 15(7):1553–1568.

2. Erkan, Radev, Dragomir R. LexRank: graph-based lexical centrality as salience in text summarization. Journal of Qiqihar Junior Teachers College, 2011, 22:2004.

3. Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016a. Learning deep structure-preserving image-text embeddings. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5005–5013.

Thank you!

Q&A