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

Haoran Li¹,², Junnan Zhu¹,², Cong Ma¹,², Jiajun Zhang¹,² and Chengqing Zong¹,²,³
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
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?

Eric Boehlert @EricBoehlert · 39
McVeigh got the death penalty. Hopefully the Virginia Nazi will, too

Ali Bundy @ThreeTouchDowes · 169?
Virginia Governor says armed militia had “better guns” than police officers. Let that sink in.

David Simon @RawDispairs · 169
Virginia law enforcement slow to interpose because Nazis were more heavily armed. Oh, Gotcha.

I’ll just leave this here.
Introduction

What is Asynchronous?

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

➢ What is Asynchronous?

- Asynchronous

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.

Ebola a serious disease that spreads rapidly through direct contact with infected people.
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 overview

Modalities

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.
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.
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.

How to get rid of ASR errors?
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.
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.
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.
Model

Model overview

- Bridge the semantic gaps between multi-modal content.

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

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

Emergency teams focus on searching.

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

New drug therapies are being evaluated.
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)

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

\[
M_{ji} = \text{sim}(t_j, t_i)
\]

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

Mr. Putin insisted that Russia was a peace-loving nation.

The prison insisted that Russia was a peace living country.
Model

- Coverage for Visual

![Diagram](attachment:image.png)

- 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**
    \[ F_s(S) = \sum_{t_i \in S} Sa(t_i) - \frac{\lambda_s}{|S|} \sum_{t_i, t_j \in S} \text{sim}(t_i, t_j) \]
  - **Coverage for Visual**
    \[ F_c(S) = \sum_{p_i \in I} \text{Im}(p_i)b_i \]
    - 1: \( p_i \) is covered by the summary; 0: \( p_i \) is not covered
    - length ratio between the shot \( p_i \) belonging to and the whole video.
  - **Considering all the modalities**
    \[ F_m(S) = \frac{1}{M_s} \sum_{t_i \in S} Sa(t_i) + \frac{1}{M_c} \sum_{p_i \in I} \text{Im}(p_i)b_i - \frac{\lambda_m}{|S|} \sum_{i,j \in S} \text{sim}(t_i, t_j) \]
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.

<table>
<thead>
<tr>
<th>Language</th>
<th>#Sentence</th>
<th>#Word</th>
<th>#Shot</th>
<th>Video Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>492.1</td>
<td>12,104.7</td>
<td>47.2</td>
<td>197s</td>
</tr>
<tr>
<td>Chinese</td>
<td>402.1</td>
<td>9,689.3</td>
<td>49.3</td>
<td>207s</td>
</tr>
</tbody>
</table>

Table 1: Corpus statistics.
## Experiments

### Dataset

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text only</td>
<td>0.422</td>
<td>0.114</td>
<td>0.166</td>
</tr>
<tr>
<td>Text + audio</td>
<td>0.422</td>
<td>0.109</td>
<td>0.164</td>
</tr>
<tr>
<td>Text + audio + guide</td>
<td>0.440</td>
<td>0.117</td>
<td>0.171</td>
</tr>
<tr>
<td>Image caption</td>
<td>0.435</td>
<td>0.111</td>
<td>0.167</td>
</tr>
<tr>
<td>Image caption match</td>
<td>0.429</td>
<td>0.115</td>
<td>0.166</td>
</tr>
<tr>
<td>Image alignment</td>
<td>0.409</td>
<td>0.082</td>
<td>0.082</td>
</tr>
<tr>
<td>Image match</td>
<td><strong>0.442</strong></td>
<td><strong>0.133</strong></td>
<td><strong>0.187</strong></td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text only</td>
<td>0.409</td>
<td>0.113</td>
<td>0.167</td>
</tr>
<tr>
<td>Text + audio</td>
<td>0.407</td>
<td>0.111</td>
<td>0.166</td>
</tr>
<tr>
<td>Text + audio + guide</td>
<td>0.411</td>
<td>0.115</td>
<td><strong>0.173</strong></td>
</tr>
<tr>
<td>Image caption match</td>
<td>0.381</td>
<td>0.092</td>
<td>0.149</td>
</tr>
<tr>
<td>Image alignment</td>
<td>0.368</td>
<td>0.096</td>
<td>0.143</td>
</tr>
<tr>
<td>Image match</td>
<td><strong>0.414</strong></td>
<td><strong>0.125</strong></td>
<td><strong>0.173</strong></td>
</tr>
</tbody>
</table>
## Experimental Results

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Read</th>
<th>Inform</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text only</td>
<td>3.72</td>
<td>3.28</td>
</tr>
<tr>
<td>Text + audio</td>
<td>3.08</td>
<td>3.44</td>
</tr>
<tr>
<td>Text + audio + guide</td>
<td>3.68</td>
<td>3.64</td>
</tr>
<tr>
<td>Image match</td>
<td>3.67</td>
<td>3.83</td>
</tr>
<tr>
<td>Reference</td>
<td>4.52</td>
<td>4.36</td>
</tr>
<tr>
<td><strong>Chinese</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text only</td>
<td>3.64</td>
<td>3.40</td>
</tr>
<tr>
<td>Text + audio</td>
<td>3.16</td>
<td>3.48</td>
</tr>
<tr>
<td>Text + audio + guide</td>
<td>3.60</td>
<td>3.72</td>
</tr>
<tr>
<td>Image match</td>
<td>3.62</td>
<td>3.92</td>
</tr>
<tr>
<td>Reference</td>
<td>4.88</td>
<td>4.84</td>
</tr>
</tbody>
</table>
Experiment Results

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

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.
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 address 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


Thank you!

Q&A