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

Abstract

The rapid increase in multimedia data transmission over the Internet necessitates the multi-modal summarization (MMS) from collections of text, image, audio and video. In this work, we propose an extractive multi-modal summarization method that can automatically generate a textual summary given a set of documents, images, audios and videos related to a specific topic. The key idea is to bridge the semantic gaps between multi-modal content. For audio information, we design an approach to selectively use its transcription. For visual information, we learn the joint representations of text and images using a neural network. Finally, all of the multi-modal aspects are considered to generate the textual summary by maximizing the salience, non-redundancy, readability and coverage through the budgeted optimization of submodular functions. We further introduce an MMS corpus in English and Chinese, which is released to the public$^1$. The experimental results obtained on this dataset demonstrate that our method outperforms other competitive baseline methods.

1 Introduction

Multimedia data (including text, image, audio and video) have increased dramatically recently, which makes it difficult for users to obtain important information efficiently. Multi-modal summarization (MMS) can provide users with textual summaries that can help acquire the gist of multimedia data in a short time, without reading documents or watching videos from beginning to end.

The existing applications related to MMS include meeting record summarization (Erol et al., 2003; Gross et al., 2000), sport video summarization (Tjondronegoro et al., 2011; Hasan et al., 2013), movie summarization (Evangelopoulos et al., 2013; Mademlis et al., 2016), pictorial storyline summarization (Wang et al., 2012), timeline summarization (Wang et al., 2016b) and social multimedia summarization (Del Fabro et al., 2012; Bian et al., 2013; Schinas et al., 2015; Bian et al., 2015; Shah et al., 2015, 2016). When summarizing meeting recordings, sport videos and movies, such videos consist of synchronized voice, visual and captions. For the summarization of pictorial storylines, the input is a set of images with text descriptions. None of these applications focus on summarizing multimedia data that contain asynchronous information about general topics.

In this paper, as shown in Figure 1, we propose an approach to a generate textual summary from a set of asynchronous documents, images, audios and videos on the same topic.

Since multimedia data are heterogeneous and contain more complex information than pure text does, MMS faces a great challenge in addressing the semantic gap between different modalities. The framework of our method is shown in Figure 1. For the audio information contained in videos, we obtain speech transcriptions through Automatic Speech Recognition (ASR) and design a method to use these transcriptions selectively. For visual information, including the key-frames extracted from videos and the images that appear in documents, we learn the joint representations of texts and images by using a neural network; we then can identify the text that is relevant to the image. In this way, audio and visual information can be integrated into a textual summary.

Traditional document summarization involves two essential aspects: (1) Salience: the summa-
ry should retain significant content of the input documents. (2) Non-redundancy: the summary should contain as little redundant content as possible. For MMS, we consider two additional aspects: (3) Readability: because speech transcriptions are occasionally ill-formed, we should try to get rid of the errors introduced by ASR. For example, when a transcription provides similar information to a sentence in documents, we should prefer the sentence to the transcription presented in the summary. (4) Coverage for the visual information: images that appear in documents and videos often capture event highlights that are usually very important. Thus, the summary should cover as much of the important visual information as possible. All of the aspects can be jointly optimized by the budgeted maximization of submodular functions (Khuller et al., 1999).

Our main contributions are as follows:

- We design an MMS method that can automatically generate a textual summary from a set of asynchronous documents, images, audios and videos related to a specific topic.

- To select the representative sentences, we consider four criteria that are jointly optimized by the budgeted maximization of submodular functions.

- We introduce an MMS corpus in English and Chinese. The experimental results on this dataset demonstrate that our system can take advantage of multi-modal information and outperforms other baseline methods.

2 Related Work

2.1 Multi-document Summarization

Multi-document summarization (MDS) attempts to extract important information for a set of documents related to a topic to generate a short summary. Graph based methods (Mihalcea and Tarau, 2004; Wan and Yang, 2006; Zhang et al., 2016) are commonly used. LexRank (Erkan and Radev, 2011) first builds a graph of the documents, in which each node represents a sentence and the edges represent the relationship between sentences. Then, the importance of each sentence is computed through an iterative random walk.

2.2 Multi-modal Summarization

In recent years, much work has been done to summarize meeting recordings, sport videos, movies, pictorial storylines and social multimedia.

Erol et al. (2003) aim to create important segments of a meeting recording based on audio, textual and visual activity analysis. Tjondronegoro et al. (2011) propose a way to summarize a sporting event by analyzing the textual information extracted from multiple resources and identifying the important content in a sport video. Evangelopoulos et al. (2013) use an attention mechanism to detect salient events in a movie. Wang et al. (2012) and Wang et al. (2016b) use image-text pairs to generate a pictorial storyline and timeline summarization. Li et al. (2016) develop an approach for multimedia news summarization for searching results on the Internet, in which the hLDA model is introduced to discover the topic structure of the news documents. Then, a news article and an image are chosen to represent each topic. For social media summarization, Fabro et al. (2012) and Schinas et al. (2015) propose to summarize the real-life events based on multimedia content such as photos from Flickr and videos from YouTube. Bion et al. (2013; 2015) propose a multimodal LDA to detect topics by capturing the correlations between textual and visual features of microblogs with embedded images. The output of their method is a set of representative images that describe the events. Shah et al. (2015; 2016) introduce EventBuilder
which produces text summaries for a social event leveraging Wikipedia and visualizes the event with social media activities.

Most of the above studies focus on synchronous multi-modal content, i.e., in which images are paired with text descriptions and videos are paired with subtitles. In contrast, we perform summarization from asynchronous (i.e., there is no given description for images and no subtitle for videos) multi-modal information about news topics, including multiple documents, images, and videos, to generate a fixed length textual summary. This task is both more general and more challenging.

3 Our Model

3.1 Problem Formulation

The input is a collection of multi-modal data \( \mathcal{M} = \{D_1, ..., D_{|D|}, V_1, ..., V_{|V|}\} \) related to a news topic \( T \), where each document \( D_i = \{T_i, I_i\} \) consists of text \( T_i \) and image \( I_i \) (there may be no image for some documents). \( V_i \) denotes video. \(|\cdot|\) denotes the cardinality of a set. The objective of our work is to automatically generate textual summary to represent the principle content of \( \mathcal{M} \).

3.2 Model Overview

There are many essential aspects in generating a good textual summary for multi-modal data. The salient content in documents should be retained, and the key facts in videos and images should be covered. Further, the summary should be readable and non-redundant and should follow the fixed length constraint. We propose an extraction-based method in which all these aspects can be jointly optimized by the budgeted maximization of submodular functions defined as follows:

\[
\max_{S \subseteq T} \{ \mathcal{F}(S) : \sum_{s \in S} l_s \leq L \} \quad (1)
\]

where \( T \) is the set of sentences, \( S \) is the summary, \( l_s \) is length (number of words) of sentence \( s \), \( L \) is budget, i.e., length constraint for the summary, and submodular function \( \mathcal{F}(S) \) is the summary score related to the above-mentioned aspects.

Text is the main modality of documents, and in some cases, images are embedded in documents. Videos consist of at least two types of modalities: audio and visual. Next, we give overall processing methods for different modalities.

Audio, i.e., speech, can be automatically transcribed into text by using an ASR system\(^2\). Then, we can leverage a graph-based method to calculate the salience score for all of the speech transcriptions and for the original sentences in documents. Note that speech transcriptions are often ill-formed; thus, to improve the readability, we should try to avoid the errors introduced by ASR. In addition, audio features including acoustic confidence (Valenza et al., 1999), audio power (Christel et al., 1998) and audio magnitude (Dagtas and Abdel-Mottaleb, 2001) have proved to be helpful for speech and video summarization which will benefit our method.

For visual, which is actually a sequence of images (frames), because most of the neighboring frames contain redundant information, we first extract the most meaningful frames, i.e., the key-frames, which can provide the key facts for the whole video. Then, it is necessary to perform semantic analysis between text and visual. To this end, we learn the joint representations for textual and visual modalities and can then identify the sentence that is relevant to the image. In this way, we can guarantee the coverage of generated summary for the visual information.

3.3 Salience for Text

We apply a graph-based LexRank algorithm (Erkan and Radev, 2011) to calculate salience score of the text unit, including the sentences in documents and the speech transcriptions from videos. LexRank first constructs a graph based on the text units and their relationship and then conducts an iteratively random walk to calculate the salience score of the text unit, \( sa(t_i) \), until convergence using the following equation:

\[
Sa(t_i) = \mu \sum_j Sa(t_j) \cdot M_{ji} + \frac{1 - \mu}{N} \quad (2)
\]

where \( \mu \) is the damping factor that is set to 0.85, \( N \) is the total number of the text units. \( M_{ji} \) is the relationship between text unit \( t_i \) and \( t_j \), which is computed as follows:

\[
M_{ji} = sim(t_j, t_i) \quad (3)
\]

The text unit \( t_i \) is represented by averaging the embeddings of the words (except stop-words) in \( t_i \). \( sim(\cdot, \cdot) \) denotes cosine similarity between two texts (negative similarities are replaced with 0).

\(^2\)We use IBM Watson Speech to Text service: www.ibm.com/watson/developercloud/speech-to-text.html
Figure 2: LexRank with guidance strategies. \( e_1 \) is guided because speech transcription \( v_3 \) is related to document sentence \( v_1 \); \( e_2 \) and \( e_3 \) are guided because of audio features. Other edges without arrow are bidirectional.

For MMS task, we propose two guidance strategies to amend the affinity matrix \( M \) and calculate salience score of the text as shown in Figure 2.

### 3.3.1 Readability Guidance Strategies

The random walk process can be understood as a recommendation: \( M_{ji} \) in Equation 2 denotes that \( t_j \) will recommend \( t_i \) to the degree of \( M_{ji} \). The affinity matrix \( M \) in the LexRank model is symmetric, which means \( M_{ij} = M_{ji} \). In contrast, for MMS, considering the unsatisfactory quality of speech recognition, symmetric affinity matrices are inappropriate. Specifically, to improve the readability, for a speech transcription, if there is a sentence in document that is related to this transcription, we would prefer to assign the text sentence a higher salience score than that assigned to the transcribed one. To this end, the process of a random walk should be guided to control the recommendation direction: when a document sentence is related to a speech transcription, the symmetric weighted edge between them should be transformed into a unidirectional edge, in which we invalidate the direction from document sentence to the transcribed one. In this way, speech transcriptions will not be recommended by the corresponding document sentences. Important speech transcriptions that cannot be covered by documents still have the chance to obtain high salience scores. For the pair of a sentence \( t_i \) and a speech transcription \( t_j \), \( M_{ij} \) is computed as follows:

\[
M_{ij} = \begin{cases} 
0, & \text{if } sim(t_i, t_j) > T_{text} \\
\text{sim}(t_i, t_j), & \text{otherwise}
\end{cases}
\]

(4)

where threshold \( T_{text} \) is used to determine whether a sentence is related to others. We obtain the proper semantic similarity threshold by testing on Microsoft Research Paraphrase (MSRParaphrase) dataset (Quirk et al., 2004). It is a publicly available paraphrase corpus that consists of 5801 pairs of sentences, of which 3900 pairs are semantically equivalent.

### 3.3.2 Audio Guidance Strategies

Some audio features can guide the summarization system to select more important and readable speech transcriptions. Valenza et al. (1999) use acoustic confidence to obtain accurate and readable summaries of broadcast news programs. Christel et al. (1998) and Dagtas and Abdel-Mottaleb (2001) apply audio power and audio magnitude to find significant audio events. In our work, we first balance these three feature scores for each speech transcription by dividing their respective maximum values among the whole amount of audio, and we then average these scores to obtain the final audio score for speech transcription. For each adjacent speech transcription pair \((t_k, t_{k'})\), if the audio score \( a(t_k) \) for \( t_k \) is smaller than a certain threshold while \( a(t_{k'}) \) is greater, which means that \( t_{k'} \) is more important and readable than \( t_k \), then \( t_k \) should recommend \( t_{k'} \), but \( t_{k'} \) should not recommend \( t_k \). We formulate it as follows:

\[
\begin{align*}
M_{kk'} &= \text{sim}(t_k, t_{k'}) \\
M_{k'k} &= 0
\end{align*}
\]

(5)

where the threshold \( T_{audio} \) is the average audio score for all the transcriptions in the audio.

Finally, affinity matrices are normalized so that each row adds up to 1.

### 3.4 Text-Image Matching

The key-frames contained in videos and the images embedded in documents often captures news highlights in which the important ones should be covered by the textual summary. Before measuring the coverage for images, we should train the model to bridge the gap between text and image, i.e., to match the text and image.

We start by extracting key-frames of videos based on shot boundary detection. A shot is defined as an unbroken sequence of frames. The abrupt transition of RGB histogram features often indicates shot boundaries (Zhuang et al., 1998). Specifically, when the transition of the RGB histogram feature for adjacent frames is greater than a certain ratio\(^3\) of the average transition for the whole video, we segment the shot. Then, the frames

---

\(^3\)The ratio is determined by testing on the
in the middle of each shot are extracted as keyframes. These key-frames and images in documents make up the image set that the summary should cover.

Next, it is necessary to perform a semantic analysis between the text and the image. To this end, we learn the joint representations for textual and visual modalities by using a model trained on the Flickr30K dataset (Young et al., 2014), which contains 31,783 photographs of everyday activities, events and scenes harvested from Flickr. Each photograph is manually labeled with 5 textual descriptions. We apply the framework of Wang et al. (2016a), which achieves state-of-the-art performance for text-image matching task on the Flickr30K dataset. The image is encoded by the VGG model (Simonyan and Zisserman, 2014) that has been trained on the ImageNet classification set containing 31,783 photographs of everyday activities, events and scenes harvested from Flickr. Each photograph is manually labeled with 5 textual descriptions.

We set the threshold as the average matching score for the positive text-image pair in Flickr30K, although the matching performance for our task could in principle be improved by adjusting this parameter.

After the text-image matching model is trained and the sentences are simplified, for each text-image pair (T_i, I_j) in our task, we can identify the matched pairs if the score s(T_i, I_j) is greater than a threshold T_{match}. We set the threshold as the average matching score for the positive text-image pair in Flickr30K, although the matching performance for our task could in principle be improved by adjusting this parameter.

3.3 Multi-modal Summarization

We model the salience of a summary S as the sum of salience scores S_a(t_i) of the sentence t_i in the summary, combining a λ-weighted redundancy penalty term:

\[ F_{r}(S) = \sum_{t_i \in S} S_a(t_i) - \frac{\lambda_1}{|S|} \sum_{t_i, t_j \in S} \text{sim}(t_i, t_j) \]  

(8)

We model the summary S coverage for the image set I as the weighted sum of image covered by the summary:

\[ F_{c}(S) = \sum_{p_i \in I} I_m(p_i) b_i \]  

(9)

where the weight I_m(p_i) for the image p_i is the length ratio between the shot p_i and the whole video. b_i is a binary variable to indicate

Note that the images in Flickr30K are similar to our task. However, the image descriptions are much simpler than the text in news, so the model trained on Flickr30K cannot be directly used for our task. For example, some of the information contained in the news, such as the time and location of events, cannot be directly reflected by images. To solve this problem, we simplify each sentence and speech transcription based on semantic role labelling (Gildea and Jurafsky, 2002), in which each predicate indicates an event and the arguments express the relevant information of this event. ARG0 denotes the agent of the event, and ARG1 denotes the action. The assumption is that the concepts including agent, predicate and action compose the body of the event, so we extract “ARG0+predicate+ARG1” as the simplified sentence that is used to match the images. It is worth noting that there may be multiple predicate-argument structures for one sentence and we extract all of them.

After the text-image matching model is trained and the sentences are simplified, for each text-image pair (T_i, I_j) in our task, we can identify the matched pairs if the score s(T_i, I_j) is greater than a threshold T_{match}. We set the threshold as the average matching score for the positive text-image pair in Flickr30K, although the matching performance for our task could in principle be improved by adjusting this parameter.

In the experiments, K = 50, m = 0.1 and λ_1 = 2. Wang et al. (2016a) also proved that structure-preserving constraints can make 1% Recall@1 improvement.

the length ratio between the shot p_i and the whole video. b_i is a binary variable to indicate

\[ F_{r}(S) = \sum_{t_i \in S} S_a(t_i) - \frac{\lambda_1}{|S|} \sum_{t_i, t_j \in S} \text{sim}(t_i, t_j) \]  

(8)

We model the summary S coverage for the image set I as the weighted sum of image covered by the summary:

\[ F_{c}(S) = \sum_{p_i \in I} I_m(p_i) b_i \]  

(9)

where the weight I_m(p_i) for the image p_i is the length ratio between the shot p_i and the whole video. b_i is a binary variable to indicate

Note that the images in Flickr30K are similar to our task. However, the image descriptions are much simpler than the text in news, so the model trained on Flickr30K cannot be directly used for our task. For example, some of the information contained in the news, such as the time and location of events, cannot be directly reflected by images. To solve this problem, we simplify each sentence and speech transcription based on semantic role labelling (Gildea and Jurafsky, 2002), in which each predicate indicates an event and the arguments express the relevant information of this event. ARG0 denotes the agent of the event, and ARG1 denotes the action. The assumption is that the concepts including agent, predicate and action compose the body of the event, so we extract “ARG0+predicate+ARG1” as the simplified sentence that is used to match the images. It is worth noting that there may be multiple predicate-argument structures for one sentence and we extract all of them.

After the text-image matching model is trained and the sentences are simplified, for each text-image pair (T_i, I_j) in our task, we can identify the matched pairs if the score s(T_i, I_j) is greater than a threshold T_{match}. We set the threshold as the average matching score for the positive text-image pair in Flickr30K, although the matching performance for our task could in principle be improved by adjusting this parameter.

3.3 Multi-modal Summarization

We model the salience of a summary S as the sum of salience scores S_a(t_i) of the sentence t_i in the summary, combining a λ-weighted redundancy penalty term:

\[ F_{r}(S) = \sum_{t_i \in S} S_a(t_i) - \frac{\lambda_1}{|S|} \sum_{t_i, t_j \in S} \text{sim}(t_i, t_j) \]  

(8)

We model the summary S coverage for the image set I as the weighted sum of image covered by the summary:

\[ F_{c}(S) = \sum_{p_i \in I} I_m(p_i) b_i \]  

(9)

where the weight I_m(p_i) for the image p_i is the length ratio between the shot p_i and the whole video. b_i is a binary variable to indicate
whether an image $p_i$ is covered by the summary, i.e., whether there is at least one sentence in the summary matching the image.

Finally, considering all the modalities, the objective function is defined as follows:

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

(10)

where $M_s$ is the summary score obtained by Equation 8 and $M_c$ is the summary score obtained by Equation 9. The aim of $M_s$ and $M_c$ is to balance the aspects of salience and coverage for images. $\lambda_s$ and $\lambda_m$ are determined by testing on development set. Note that to guaranteed monotonicity of $F$, $\lambda_s$, and $\lambda_m$ should be lower than the minimum salience score of sentences. To further improve non-redundancy, we make sure that similarity between any pair of sentences in the summary is lower than $T_{text}$.

Equations 8, 9 and 10 are all monotone submodular functions under the budget constraint. Thus, we apply the greedy algorithm (Lin and Bilmes, 2010) guaranteeing near-optimization to solve the problem.

## 4 Experiment

### 4.1 Dataset

There is no benchmark dataset for MMS. We construct a dataset as follows. We select 50 news topics in the most recent five years, 25 in English and 25 in Chinese. We set 5 topics for each language as a development set. For each topic, we collect 20 documents within the same period using Google News search\(^6\) and 5-10 videos in CCTV.com\(^7\) and Youtube\(^8\). More details of the corpus are illustrated in Table 1. Some examples of news topics are provided Table 2.

We employ 10 graduate students to write reference summaries after reading documents and watching videos on the same topic. We keep 3 reference summaries for each topic. The criteria for summarizing documents lie in: (1) retaining important content of the input documents and videos; (2) avoiding redundant information; (3) having good readability; (4) following the length limit.

We set the length constraint for each English and Chinese summary to 300 words and 500 characters, respectively.

<table>
<thead>
<tr>
<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>
</tr>
<tr>
<td>Chinese</td>
<td>402.1</td>
<td>9,689.3</td>
<td>49.3</td>
</tr>
</tbody>
</table>

Table 1: Corpus statistics.

## 4.2 Comparative Methods

Several models are compared in our experiments, including generating summaries with different modalities and different approaches to leverage images.

**Text only.** This model generates summaries only using the text in documents.

**Text + audio.** This model generates summaries using the text in documents and the speech transcriptions but without guidance strategies.

**Text + audio + guide.** This model generates summaries using the text in documents and the speech transcriptions with guidance strategies.

The following models generate summaries using both documents and videos but take advantage of images in different ways. The salience scores for text are obtained with guidance strategies.

**Image caption.** The image is first captioned using the model of Vinyals et al. (2016) which achieved first place in the 2015 MSCOCO Image Captioning Challenge. This model generates summaries using text in documents, speech transcription and image captions.

Note that the above-mentioned methods generate summaries by using Equation 8 and the follow-
ing methods using Equation 8,9 and 10.

**Image caption match.** This model uses generated image captions to match the text; i.e., if the similarity between a generated image caption and a sentence exceeds the threshold $T_{text}$, the image and the sentence match.

**Image alignment.** The images are aligned to the text in the following ways: The images in a document are aligned to all the sentences in this document and the key-frames in a shot are aligned to all the speech transcriptions in this shot.

**Image match.** The texts are matched with images using the approach introduced in Section 3.4.

### 4.3 Implementation Details

We perform sentence\(^9\) and word tokenization, and all the Chinese sentences are segmented by Stanford Chinese Word Segmenter (Tseng et al., 2005). We apply Stanford CoreNLP toolkit (Levy and D. Manning, 2003; Klein and D. Manning, 2003) to perform lexical parsing and use semantic role labelling approach proposed by Yang and Zong (2014). We use 300-dimension skip-gram English word embeddings which are publicly available\(^{10}\). Given that text-image matching model and image caption generation model are trained in English, to create summaries in Chinese, we first translate the Chinese text into English via Google Translation\(^{11}\) and then conduct text and image matching.

### 4.4 Multi-modal Summarization Evaluation

We use the ROUGE-1.5.5 toolkit (Lin and Hovy, 2003) to evaluate the output summaries. This evaluation metric measures the summary quality by matching n-grams between generated summary and reference summary. Table 3 and Table 4 show the averaged ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) F-scores regarding to the three reference summaries for each topic in English and Chinese.

For the results of the English MMS, from the first three lines in Table 3 we can see that when summarizing without visual information, the method with guidance strategies performs slightly better than do the first two methods. Because Rouge mainly measures word overlaps, manual evaluation is needed to confirm the impact of guidance strategies on improving readability. It is introduced in Section 4.5. The rating ranges from 1 (the poorest) to 5 (the best). When summarizing with textual and visual modalities, performances are not always improved, which indicates that the models of image caption, image caption match and image alignment are not suitable to MMS. The image match model has a significant advantage over other comparative methods, which illustrates that it can make use of multi-modal information.

Table 4 shows the Chinese MMS results, which are similar to the English results that the image match model achieves the best performance. We find that the performance enhancement for the image match model is smaller in Chinese than it is in English, which may be due to the errors introduced by machine translation.

We provides a generated summary in English using the image match model, which is shown in Figure 3.

### 4.5 Manual Summary Quality Evaluation

The readability and informativeness for summaries are difficult to evaluate formally. We ask five graduate students to measure the quality of summaries generated by different methods. We calculate the average score for all of the topics, and the results are displayed in Table 5. Overall, our method with guidance strategies achieves higher scores than do the other methods, but it is still obviously poorer than the reference sum-

\(^9\)We exclude sentences containing less than 5 words.

\(^{10}\)https://code.google.com/archive/p/word2vec/

\(^{11}\)https://translate.google.com

<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>0.442</td>
<td>0.133</td>
<td>0.187</td>
</tr>
</tbody>
</table>

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

<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>0.173</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>0.414</td>
<td>0.125</td>
<td>0.173</td>
</tr>
</tbody>
</table>

Table 4: Experimental results (F-score) for Chinese MMS.
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 Santhosh Murta said. Some railway officials told local media they suspected faulty tracks caused the derailment. Fourteen cars in the 23-car train detailed, Modak said. We do not 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 not 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 3: An example of generated summary for the news topic “India train derailment”. The sentences covering the images are labeled by the corresponding colors. The text can be partly related to the image because we use simplified sentence based on SRL to match the images. We can find some mismatched sentences, such as the sentence “Fourteen cars in the 23-car train derailed, Modak said.” where our text-image matching model may misunderstand the “car” as a “motor vehicle” but not a “coach”.

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>English</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>Chinese</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>

Table 6: Guidance examples. “CST” denotes manually modified correct ST. ASR errors are marked red and revisions are marked blue.

<table>
<thead>
<tr>
<th></th>
<th>DT</th>
<th>ST</th>
<th>CST</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>There were 12 bodies at least pulled from the rubble in the square.</td>
<td>Still being pulled from the rubble.</td>
<td>Many people are still being pulled from the rubble.</td>
</tr>
<tr>
<td>B</td>
<td>Conflict between police and protesters lit up on Tuesday.</td>
<td>Late-night tensions between police and protesters briefly lit up this Baltimore neighborhood Tuesday.</td>
<td>Late-night tensions between police and protesters briefly lit up in a Baltimore neighborhood Tuesday.</td>
</tr>
</tbody>
</table>

4.6 How Much is the Image Worth

Text-image matching is the toughest module for our framework. Although we use a state-of-the-art approach to match the text and images, the performance is far from satisfactory. To find a somewhat strong upper-bound of the task, we choose five topics for each language to manually label the text-image matching pairs. The MMS results on these topics are shown in Table 7 and Table 8. The experiments show that with the ground truth text-image matching result, the summary quality can be promoted to a considerable extent, which indicates visual information is crucial for MMS.

An image and the corresponding texts obtained using different methods are given in Figure 4 and Figure 5. We can conclude that the image caption
and the image caption match contain little of the image’s intrinsically intended information. The image alignment introduces more noise because it is possible that the whole text in documents or the speech transcriptions in shot are aligned to the document images or the key-frames, respectively. The image manually match can obtain similar results to the image manually match, which illustrates that the image match can make use of visual information to generate summaries.

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text + audio + guide</td>
<td>0.426</td>
<td>0.105</td>
<td>0.167</td>
</tr>
<tr>
<td>Image caption</td>
<td>0.423</td>
<td>0.106</td>
<td>0.167</td>
</tr>
<tr>
<td>Image caption match</td>
<td>0.400</td>
<td>0.086</td>
<td>0.167</td>
</tr>
<tr>
<td>Image alignment</td>
<td>0.399</td>
<td>0.069</td>
<td>0.136</td>
</tr>
<tr>
<td>Image match</td>
<td>0.436</td>
<td>0.126</td>
<td>0.177</td>
</tr>
<tr>
<td>Image manually match</td>
<td>0.446</td>
<td>0.150</td>
<td>0.207</td>
</tr>
</tbody>
</table>

Table 7: Experimental results (F-score) for English MMS on five topics with manually labeled text-image pairs.

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text + audio + guide</td>
<td>0.417</td>
<td>0.115</td>
<td>0.171</td>
</tr>
<tr>
<td>Image caption match</td>
<td>0.396</td>
<td>0.095</td>
<td>0.152</td>
</tr>
<tr>
<td>Image alignment</td>
<td>0.306</td>
<td>0.072</td>
<td>0.111</td>
</tr>
<tr>
<td>Image match</td>
<td>0.401</td>
<td>0.127</td>
<td>0.179</td>
</tr>
<tr>
<td>Image manually match</td>
<td>0.419</td>
<td>0.162</td>
<td>0.208</td>
</tr>
</tbody>
</table>

Table 8: Experimental results (F-score) for Chinese MMS on five topics with manually labeled text-image pairs.

5 Conclusion

This paper addresses an asynchronous MMS task, namely, how to use related text, audio and video information to generate a textual summary. We formulate the MMS task as an optimization problem with a budgeted maximization of submodular functions. To selectively use the transcription of audio, guidance strategies are designed using the graph model to effectively calculate the salience score for each text unit, 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. The final experimental results obtained using our MMS corpus in both English and Chinese demonstrate that our system can benefit from multi-modal information.

Adding audio and video does not seem to improve dramatically over text only model, which indicates that better models are needed to capture the interactions between text and other modalities, especially for visual. We also plan to enlarge our MMS dataset, specifically to collect more videos.

Acknowledgments

The research work has been supported by the Natural Science Foundation of China under Grant No. 61333018 and No. 61403379.
References


