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.

[1]http://www.nlpr.ia.ac.cn/cip/jjzhang.htm

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-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1092–1102
Copenhagen, Denmark, September 7–11, 2017.

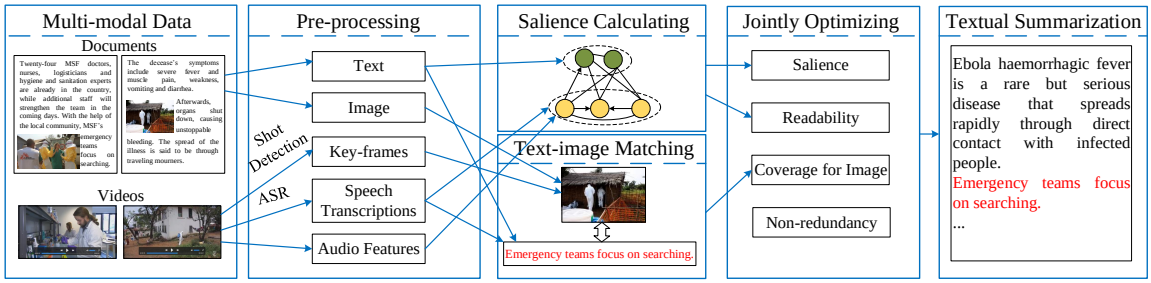

Figure 1: The framework of our MMS model.

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, text 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. Bian 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 $\mathcal{T}$, 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:

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

where T is the set of sentences, S is the summary, l_s is length (number of words) of sentence s, $\mathcal{L}$ 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 keyframes, 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 μ 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. $sim(\cdot)$ 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

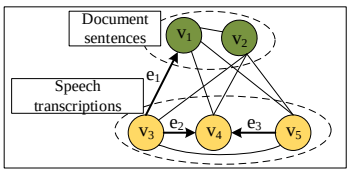

Figure 2: LexRank with guidance strategies. e_1 is guided because speech transcription v_3 is related to document sentence v_1; e_2 and e_3 are guided because of audio features. Other edges without arrow are bidirectional.

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:

$$M_{ij} = \begin{cases} 0, & \text{if } sim(t_i, t_j) > T_{text} \\ sim(t_i, t_j), & \text{otherwise} \end{cases} \quad (4)$$

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:

$$\begin{cases} M_{kk'} = sim(t_k, t_{k'}) \\ M_{k'k} = 0 \end{cases} \quad (5)$$
$$\text{if } a(t_k) < T_{audio} \text{ and } a(t_{k'}) > T_{audio}$$

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 key-frames. 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 task following the standard procedure (Wang et al., 2016a). The 4096-dimensional feature from the pre-softmax layer is used to represent the image. The text is first encoded by the Hybrid Gaussian-Laplacian mixture model (HGLMM) using the method of Klein et al. (2014). Then, the HGLMM vectors are reduced to 6000 dimensions through PCA. Next, the sentence vector v_s and image vector v_i are mapped to a joint space by a two-branch neural network as follows:

$$\begin{cases} x = W_2 \cdot f(W_1 \cdot v_s + b_s) \\ y = V_2 \cdot f(V_1 \cdot v_i + b_i) \end{cases} \quad (6)$$

where $W_1 \in \mathbb{R}^{2048\times6000}$, $b_s \in \mathbb{R}^{2048}$, $W_2 \in \mathbb{R}^{512\times2048}$, $V_1 \in \mathbb{R}^{2048\times4096}$, $b_i \in \mathbb{R}^{2048}$, $V_2 \in \mathbb{R}^{512\times2048}$, f is Rectified Linear Unit (ReLU).

The max-margin learning framework is applied to optimize the neural network as follows:

$$\begin{aligned} L = &\sum_{i,k} \max[0, m + s(x_i, y_i) - s(x_i, y_k)] \\ &+ \lambda_1 \sum_{i,k} \max[0, m + s(x_i, y_i) - s(x_k, y_i)] \end{aligned} \quad (7)$$

where for positive text-image pair (x_i, y_i), the top K most violated negative pairs (x_i, y_k) and (x_k, y_i) in each mini-batch are sampled. The objective function L favors higher matching score $s(x_i, y_i)$ (cosine similarity) for positive text-image pairs than for negative pairs[4].

shot detection dataset of TRECVID. http://www-nlpir.nist.gov/projects/trecvid/

[4]In the experiments, $K = 50$, $m = 0.1$ and $\lambda_1 = 2$. Wang et al. (2016a) also proved that structure-preserving constraints can make 1% Recall@1 improvement.

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.

3.5 Multi-modal Summarization

We model the salience of a summary S as the sum of salience scores $Sa(t_i)$[5] of the sentence t_i in the summary, combining a λ-weighted redundancy penalty term:

$$\mathcal{F}_s(S) = \sum_{t_i \in S} Sa(t_i) - \frac{\lambda_s}{|S|} \sum_{t_i, t_j \in S} sim(t_i, t_j) \quad (8)$$

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

$$\mathcal{F}_c(S) = \sum_{p_i \in I} Im(p_i) b_i \quad (9)$$

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

[5]Normalized by the maximum value among all the sentences.

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:

$$\mathcal{F}_m(S) = \frac{1}{M_s}\sum_{t_i \in S} Sa(t_i) + \frac{1}{M_c}\sum_{p_i \in I} Im(p_i)b_i - \frac{\lambda_m}{|S|}\sum_{i,j \in S} sim(t_i, t_j) \quad (10)$$

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. λ_s, and λ_m are determined by testing on development set. Note that to guaranteed monotone of $\mathcal{F}$, λ_s, and λ_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 a 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.

	#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.

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.

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-

[6]http://news.google.com/
[7]http://www.cctv.com/
[8]https://www.youtube.com/

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 a 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.

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 MMS.

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 MMS.

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

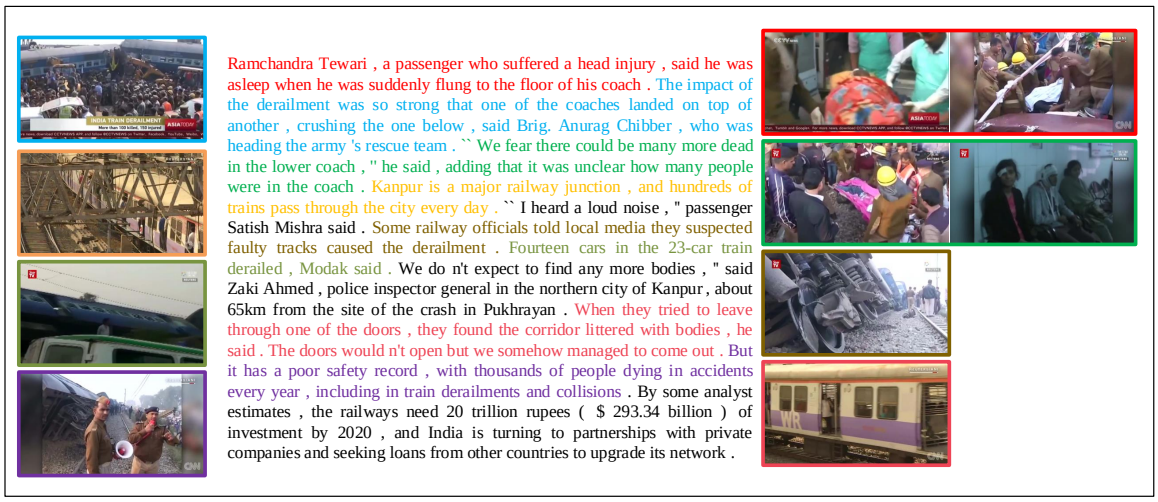

Figure 3: An example of generated summary for the news topic "India train derailment". The sentences covering the images are labeled by the corresponding colors. The text can be partly related to the image because we use simplified sentence based on SRL to match the images. We can find some mismatched 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".

maries. Specifically, when speech transcriptions are not considered, the informativeness of the summary is the worst. However, adding speech transcriptions without guidance strategies decreases readability to a large extent, which indicates that guidance strategies are necessary for MMS. The **image match** model achieves higher informativeness scores than do the other methods without using images.

We give two instances of readability guidance that arise between document text (DT) and speech transcriptions (ST) in Table 6. The errors introduced by ASR include segmentation (instance A) and recognition (instance B) mistakes.

	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".

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

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

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 an d Figure 5. We can conclude that the **image caption**

Image caption:
A group of people standing on top of a lush green field.
Image caption match:
We could barely stay standing.
Image hard alignment:
The need for doctors would grow as more survivors were pulled from the rubble.
Image match:
The search, involving US, Indian and Nepali military choppers and a battalion of 400 Nepali soldiers, has been joined by two MV-22B Osprey.
Image manually match:
The military helicopter was on an aid mission in Dolakha district near Tibet.

Figure 4: An example image with corresponding English texts that different methods obtain.

Image caption match:
就 星 州 民众 举行 抗议 集会， 文尚 均 表示， 国防部 愿意 与 当地 居民 沟通 。
(On behalf of the protest rally of people in Seongju, Moon Sang-gyun said that the Ministry of National Defense is willing to communicate with local residents.)
Image hard alignment:
朴槿惠 在 国家 安全 保障 会议 上 呼吁 民众 支持 “ 萨德 ” 部署 。
(Park Geun-hye called on people to support the "THAAD" deployment in the National Security Council.)
Image match:
从 7月 12日 开始， 当地 民众 连续 数日 在 星 州郡 厅 门口 请愿 。
(The local people petitioned in front of the Seongju County Office for days from July 12.)
Image manually match:
当天， 星 州郡 数千 民众 集会， 抗议 在 当地 部署 “ 萨德 ”
(On that day, thousands of people gathered in Seongju to protest the local deployment of "THAAD".)

Figure 5: An example image with corresponding Chinese texts that different methods obtain.

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 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.

Method	R-1	R-2	R-SU4
Text + audio + guide	0.426	0.105	0.167
Image caption	0.423	0.106	0.167
Image caption match	0.400	0.086	0.149
Image alignment	0.399	0.069	0.136
Image match	0.436	0.126	0.177
Image manually match	**0.446**	**0.150**	**0.207**

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

Method	R-1	R-2	R-SU4
Text + audio + guide	0.417	0.115	0.171
Image caption match	0.396	0.095	0.152
Image alignment	0.306	0.072	0.111
Image match	0.401	0.127	0.179
Image manually match	**0.419**	**0.162**	**0.208**

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

Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2013. Multimedia summarization for trending topics in microblogs. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1807–1812. ACM.

Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. 2015. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, 17(2):216–228.

Michael G Christel, Michael A Smith, C Roy Taylor, and David B Winkler. 1998. Evolving video skims into useful multimedia abstractions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 171–178. ACM Press/Addison-Wesley Publishing Co.

Serhan Dagtas and Mohamed Abdel-Mottaleb. 2001. Extraction of tv highlights using multimedia features. In *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, pages 91–96. IEEE.

Manfred Del Fabro, Anita Sobe, and Laszlo Böszörmenyi. 2012. Summarization of real-life events based on community-contributed content. In *The Fourth International Conferences on Advances in Multimedia*, pages 119–126.

Gunes Erkan and Dragomir R. Radev. 2011. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Qiqihar Junior Teachers College*, 22:2004.

Berna Erol, D-S Lee, and Jonathan Hull. 2003. Multimodal summarization of meeting recordings. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 3, pages III–25. IEEE.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics, Volume 28, Number 3, September 2002*.

Ralph Gross, Michael Bett, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. 2000. Towards a multimodal meeting record. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1593–1596. IEEE.

Taufiq Hasan, Hynek Bořil, Abhijeet Sangwan, and John HL Hansen. 2013. Multi-modal highlight generation for sports videos using an information-theoretic excitability measure. *EURASIP Journal on Advances in Signal Processing*, 2013(1):173.

Samir Khuller, Anna Moss, and Joseph Seffi Naor. 1999. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Zechao Li, Jinhui Tang, Xueming Wang, Jing Liu, and Hanqing Lu. 2016. Multimedia news summarization in search. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):33.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920. Association for Computational Linguistics.

Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, and Ioannis Pitas. 2016. Multimodal stereoscopic movie summarization conforming to narrative characteristics. *IEEE Transactions on Image Processing*, 25(12):5828–5840.

Rada Mihalcea and Paul Tarau. 2004. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, chapter TextRank: Bringing Order into Text.

Chris Quirk, Chris Brockett, and William Dolan. 2004. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, chapter Monolingual Machine Translation for Paraphrase Generation.

Manos Schinas, Symeon Papadopoulos, Georgios Petkos, Yiannis Kompatsiaris, and Pericles A Mitkas. 2015. Multimodal graph-based event detection and summarization in social media streams. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 189–192. ACM.

Rajiv Ratn Shah, Anwar Dilawar Shaikh, Yi Yu, Wenjing Geng, Roger Zimmermann, and Gangshan Wu. 2015. Eventbuilder: Real-time multimedia event

summarization by visualizing social media. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 185–188. ACM.

Rajiv Ratn Shah, Yi Yu, Akshay Verma, Suhua Tang, Anwar Dilawar Shaikh, and Roger Zimmermann. 2016. Leveraging multimodal information for event summarization and concept-level sentiment analysis. *Knowledge-Based Systems*, 108:102–109.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. 2011. Multi-modal summarization of key events and top players in sports tournament videos. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 471–478. IEEE.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter.

Robin Valenza, Tony Robinson, Marianne Hickey, and Roger Tucker. 1999. Summarisation of spoken audio through information extraction. In *ESCA Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*.

Xiaojun Wan and Jianwu Yang. 2006. Improved affinity graph based multi-document summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*.

Dingding Wang, Tao Li, and Mitsunori Ogihara. 2012. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In *AAAI*.

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.

Yang William Wang, Yashar Mehdad, R. Dragomir Radev, and Amanda Stent. 2016b. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 58–68. Association for Computational Linguistics.

Haitong Yang and Chengqing Zong. 2014. Multi-predicate semantic role labeling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 363–373. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations. *Transactions of the Association of Computational Linguistics*, 2:67–78.

Jiajun Zhang, Yu Zhou, Chengqing Zong, Jiajun Zhang, Yu Zhou, Chengqing Zong, Jiajun Zhang, Chengqing Zong, and Yu Zhou. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(10):1842–1853.

Yueting Zhuang, Yong Rui, Thomas S Huang, and Sharad Mehrotra. 1998. Adaptive key frame extraction using unsupervised clustering. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, volume 1, pages 866–870. IEEE.