Read, Watch, Listen and Summarize: Multi-modal Summarization for Asynchronous Text, Image, Audio and Video

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang* and Chengqing Zong

Abstract—Automatic text summarization is a fundamental natural language processing (NLP) application that aims to condense a source text into a shorter version. The rapid increase in multimedia data transmission over the Internet necessitates multi-modal summarization (MMS) from asynchronous collections of text, image, audio and video. In this work, we propose an extractive MMS method that unites the techniques of NLP, speech processing and computer vision to explore the rich information contained in multi-modal data and to improve the quality of multimedia news summarization. The key idea is to bridge the semantic gaps between multi-modal content. Audio and visual are main modalities in the video. For audio information, we design an approach to selectively use its transcription and to infer the salience of the transcription with audio signals. For visual information, we learn the joint representations of text and images using a neural network. Then, we capture the coverage of the generated summary for important visual information through text-image matching or multi-modal topic modeling. Finally, all the multi-modal aspects are considered to generate a textual summary by maximizing the salience, non-redundancy, readability and coverage through the budgeted optimization of submodular functions. We further introduce a publicly available MMS corpus in English and Chinese. The experimental results obtained on our dataset demonstrate that our methods based on image matching and image topic framework outperform other competitive baseline methods.

Index Terms—Summarization, Multimedia, Multi-modal, Cross-modal, Natural language processing, Computer vision

1 Introduction

Text summarization plays a vital role in our daily life and has been studied for several decades. From information retrieval to text mining, we are frequently exposed to text summarization. With the coming of the information age and the emergence of multimedia technology, multimedia data (including text, image, audio and video) have increased dramatically. Multimedia data have greatly changed the way people live and make it difficult for users to obtain important information efficiently. Intuitively, readers can grasp the gist of the event more easily by scanning the image or the video than by only reading news document, and thus we believe that the multi-modal data will also reduce the difficulty for machine to understand a news event. While most summarization systems focus on only natural language processing (NLP), the opportunity to jointly optimize the quality of the summary with the aid of automatic speech recognition (ASR) and computer vision (CV) processing systems is widely ignored. On the other hand, given a news event (i.e., news topic), multimedia data are generally asynchronous in real life, which means there is no given explicit description for images and no subtitles for videos. Thus, multi-modal summarization (MMS) [1] faces a major challenge in understanding the semantics of visual information. In this work, we present an MMS system that can provide users with textual summaries to help to acquire the gist of asynchronous multimedia data in a short time without reading documents or watching videos from beginning to end. The purpose of this work is to unite the NLP, ASR and CV techniques to explore a new framework for mining the rich information contained in multi-modal data to improve the quality of multimedia news summarization.

The existing applications related to MMS include meeting record summarization [2], [3], sport video summarization [4], [5], movie summarization [6], [7], pictorial storyline summarization [8], timeline summarization [9] and social multimedia summarization [10], [11], [12], [13], [14], [15]. Previous studies on these topics mainly focus on summarizing synchronous multi-modal content. Meeting recordings, sport videos and movies consist of synchronized voice, visual and captions, and pictorial storylines consist of a set of images with textual descriptions. None of these applications focus on summarizing multimedia data that contain asynchronous information about general events. In this paper, we propose an approach to generate a textual summary from a set of asynchronous documents, images, audios and videos on the same news event, as shown in Fig. 1. Because multimedia data are heterogeneous and contain more complex information than that contained in pure text, MMS faces a major challenge in addressing the semantic gap between different modalities. The framework

- *H. Li, J. Zhu, C. Ma, and J. Zhang* are with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences. *Corresponding author.*
  E-mail: {haoran.li, junnan.zhu, cong.ma, jjzhang}@nlpr.ia.ac.cn
- *C. Zong is with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, University of Chinese Academy of Sciences, Chinese Academy of Sciences, and CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences.*
  E-mail: cqzong@nlpr.ia.ac.cn

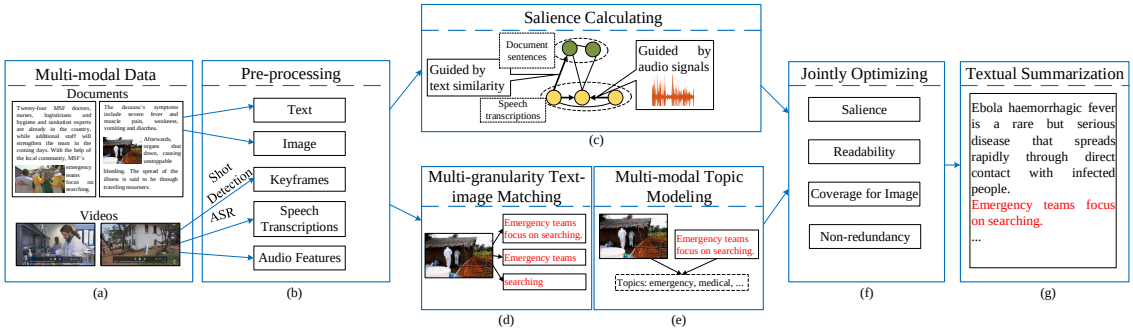


Fig. 1. The framework of our MMS model.

of our method is shown in Fig. 1. For the audio information contained in videos, we obtain speech transcriptions through ASR and design a method to selectively use these transcriptions (Fig. 1 (c)). For visual information, including the keyframes extracted from videos and the images that appear in documents, we learn the joint representations of text and images with a neural network; we can then identify the text that is relevant to the image based on text-image matching (Fig. 1 (d)) or multi-modal topic modeling (Fig. 1 (e)). In this way, audio and visual information can be integrated into a textual summary by joint optimization (Fig. 1 (f)).

Traditional document summarization considers two essential aspects: (1) Salience: the summary should retain the 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 try to get rid of the errors introduced by ASR. For example, when a transcription provides similar information to a sentence in documents, we 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 these aspects can be jointly optimized through the budgeted maximization of submodular functions [16].

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 event.
- We consider four criteria that are jointly optimized by the budgeted maximization of submodular functions to select the representative sentences.
- We perform semantic analysis between the textual and visual data. We identify text in multiple granularities that is relevant to image and measure the similarity between an image and text based on multi-modal topic modeling. In this way, we can guarantee the coverage of summary for the visual information.
- 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 outperform other baseline methods.

2 Related Work

Our work is inspired by two lines of research: multi-document summarization and multi-modal summarization.

2.1 Multi-document Summarization

Multi-document summarization (MDS) attempts to extract important information from a set of documents related to an event to generate a summary of much smaller size. MDS can be abstractive or extractive. Extractive-based models use various linguistic features, such as sentence position [17], [18] and tf*idf [19], to identify the most salient sentences in a set of documents. Graph-based methods [20], [21], [22], [23], [24], [25], [26] are commonly used extractive-based MDS models based on the hypothesis that sentences that are similar to many other sentences in a document set are more important. LexRank [22] first builds a graph from the document in which each node represents a sentence and the edges represent the relationship between sentences. Then, the importance of each sentence is computed through iterative random walk, which is applied in PageRank [27]. Finally, the top-ranked sentences are selected to build summaries. In this paper, we calculate the salience score of the sentences in documents and speech transcriptions from videos based on a LexRank algorithm with guidance strategies.

2.2 Multi-modal Summarization

In recent years, much work has been performed to summarize meeting recordings, sport videos, movies, pictorial storylines and social multimedia. Erol et al. [2] aim to create important segments of a meeting recording based on an analysis of audio, text and visual activity. Tjondronegoro et al. [4] propose a method to summarize a sporting event by analyzing the textual information extracted from multiple resources and identifying the important content. Li et al. [28] summarize news images by text and visualize text by images. Evangelopoulos et al. [6] use an attention mechanism in which visual, audio and textual features are extracted by applying multi-modal analysis to detect salient events in a movie. Mademlis et al. [7] design a multi-modal video summarization algorithm for stereoscopic movies, utilizing visual, texture, illumination, audio and semantic movie characteristics. Wang et al. [8] and Wang et al. [9] use

image-text pairs to generate a pictorial storyline and timeline summarization. Li et al. [29] develop an approach for multimedia news summarization for search results on the Internet in which the hierarchical Latent Dirichlet Allocation (hLDA) model is introduced to discover the topic structure of news documents. Then, a news article and an image are chosen to represent each topic. For social media summarization, Fabro et al. [10] and Schinas et al. [12] propose to summarize the real-life events based on multimedia content, such as photos from Flickr and videos from YouTube. Bian et al. [11], [13] propose a multi-modal LDA to detect topics by capturing the correlations between the 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. [14], [15] introduce EventBuilder, which produces textual summaries for a social event by leveraging Wikipedia and visualizes the event with social media activities.

Most of the above studies focus on synchronous multi-modal content, in which images are paired with textual descriptions and videos are paired with subtitles. In contrast, we perform summarization from asynchronous (i.e., there is no description given for images and no subtitles for videos) multi-modal information about news events, including multiple documents, images and videos, to generate a fixed length textual summary. This task is both general and 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 event $\mathcal{T}$, where each news 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 a news video and $|\cdot|$ denotes the cardinality of a set. The objective of our work is to automatically generate a fixed-length textual summary to represent the principle content of the multi-modal data $\mathcal{M}$.

3.2 Model Overview

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

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

where T is a set of sentences, S is a summary, l_s is the 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 and the specific submodular functions will be introduced in Section 3.8.

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

For text, we calculate the salience score of sentences using a graph-based LexRank algorithm [22] (Section 3.3).

For audio, i.e., speech, on the one hand, it can be automatically transcribed into text by using an ASR system[1]. Then, we leverage a graph-based method to calculate the salience score for all the speech transcriptions and the original sentences in news documents. Note that speech transcriptions are often ill-formed; thus, to improve the readability, we try to avoid the errors introduced by ASR (Section 3.3.1). On the other hand, audio features, including acoustic confidence [30], audio power [31] and audio magnitude [32], can indicate the relative importance of different parts, and have been proved to be helpful for speech and video summarization (Section 3.3.2).

For visual data, we consider images in news documents and news videos. For the visual information in a video, which is 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 provide the highlights for the whole video. Then, semantic analysis is performed between the textual and visual data (including the extracted keyframes and the original images embedded in the documents). To this end, we first learn the joint representations for textual and visual modalities using a text-image matching model (Section 3.4). Then, we measure the coverage of the visual information from two aspects: (1) images directly covered by the summary; (2) topics related to the images covered by the summary. For the first point, we identify the text in multiple granularities (Section 3.5) that is relevant to the image using the trained text-image matching model. For the second point, we explore the topics of both images and sentences through multi-modal topic modeling (Section 3.6), and we measure the similarity between an image and text based on the topic distribution (Section 3.7). We expect the topics related to the images to be covered by the sentences in the generated summary. In this way, we can guarantee the coverage of the generated summary for the visual information.

3.3 Salience for Text

We apply a graph-based LexRank algorithm [33] to calculate the salience score $Sa(s_i)$ of text, including the sentences in documents and the speech transcriptions from videos.

$$Sa(s_i) = \mu \sum_j Sa(s_j) \cdot M_{ji} + \frac{1-\mu}{N} \tag{2}$$

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 s_i and s_j, which is computed as follows:

$$M_{ji} = \frac{sim(s_j, s_i)}{\sum_k sim(s_j, s_k)} \tag{3}$$

The text unit s_i is represented by averaging the embeddings of the words (except stop-words) in s_i. $sim(\cdot)$ denotes the cosine similarity between two text.

1. www.ibm.com/watson/developercloud/speech-to-text.html.

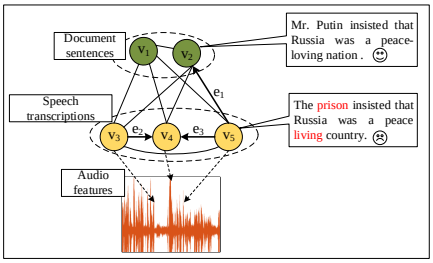


Fig. 2. LexRank with guidance strategies. e_1 is guided because speech transcription v_5 is related to document sentence v_2 and the speech recognition errors are marked red; e_2 and e_3 are guided because of audio features. Other edges without arrow are bidirectional.

We propose two guidance strategies for MMS to amend the affinity matrix M and calculate the salience score of the text as shown in Fig. 2.

3.3.1 Readability Guidance Strategies

The random walk process can be understood as a recommendation: M_{ji} in Equation 2 denotes that s_j recommends s_i to the degree of M_{ji}. The affinity matrix M in the LexRank model is symmetric, which means $M_{ij} = M_{ji}$. While advancements in ASR have been achieved, there is no guarantee against well-edited written text for the ASR output. Treating speech transcriptions and document sentences uniformly in the process of calculating text salience may result in a suboptimal readability for the generated summary. Thus, symmetric affinity matrices are inappropriate for MMS due to the unsatisfactory quality of speech recognition. There is room for potential readability improvement by leveraging the interrelation between speech transcriptions and document sentences. For example, as shown in Fig. 2, speech transcription v_5 "*The prison insisted that Russia was a peace living country.*" intends to express the same meaning as sentence v_2 "*Mr. Putin insisted that Russia was a peace loving nation.*", but there are several ASR errors in v_5, so that the readability of v_5 is poor. To improve the readability, we prefer that v_2 rather than v_5 appears in the summary. Thus, for a speech transcription, if a sentence in the document is related to the transcription, we assign the document sentence a higher salience score than that assigned to the transcribed sentence. To this end, the random walk is guided to control the direction of recommendation: when a document sentence is related to a speech transcription, the symmetric weighted edge between them is transformed into a unidirectional edge in which we invalidate the direction from the document sentence to the transcribed sentence. In this way, speech transcriptions will not be recommended by the related document sentences, and these document sentences will be encouraged. Note that important speech transcriptions that are not covered by documents still have a chance to receive high salience scores. For the pair of sentence s_i and speech transcription s_j, M_{ij} is computed as follows:

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

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 the Microsoft Research Paraphrase (MSRParaphrase) dataset [34], which 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. [30] use acoustic confidence to obtain accurate and readable summaries of broadcast news programs. Christel et al. [31] and Dagtas and Abdel-Mottaleb [32] apply audio power and audio magnitude to identify significant audio events. In our work, we first normalize these three feature scores for each speech transcription by dividing by their respective maximum values among the entire audio recording; we then average these scores to obtain the final audio score for the speech transcription. For each adjacent speech transcription pair $(s_k, s_{k'})$, if the audio score $a(s_k)$ for s_k is smaller than a certain threshold while $a(s_{k'})$ is greater, which means that $s_{k'}$ is more important and readable than s_k, then s_k should recommend $s_{k'}$, while $s_{k'}$ should not recommend s_k. We formulate this process as follows:

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

where the threshold T_{audio} is the average audio score for all the transcriptions of the audio. As shown in Fig. 2, v_4 has higher priority to appear in the summary than v_2 and v_5 due to the audio features.

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

3.4 Text-Image Matching Model

The keyframes in the videos and the images embedded in the documents often capture news highlights that represent the important information that the summary should cover. Before measuring the coverage for the images, we need a model to bridge the gap between text and image. We can solve this problem by cross-modal analysis [35], [36]. Cross-modal semantic matching can be better explored when multi-modal data is projected into the joint subspace [37]. Thus, we need to learn joint representations of text and images [38], [39], and then match the text with the image.

We start by extracting the keyframes of videos based on shot boundary detection. A shot is defined as an unbroken sequence of frames. An abrupt transition of RGB histogram features often indicates a shot boundary [40]. Specifically, when the transition of the RGB histogram features for adjacent frames is greater than a certain ratio[2] of the average transition for the whole video, we segment the shot. Then, the frames in the middle of each shot are extracted as keyframes. These keyframes and images in documents constitute the image set that the summary should cover.

Next, semantic analysis between the text and the image is necessary. We learn the joint representations for textual

2. The ratio is determined by testing on the shot-detection dataset of TRECVID: http://www-nlpir.nist.gov/projects/trecvid/

and visual modalities by using a model trained on the Flickr30K [41] and MSCOCO [42] datasets. Flickr30K contains 31,783 photographs of everyday activities, events and scenes harvested from Flickr. Each photograph is manually labeled with 5 textual descriptions. The larger MSCOCO dataset consists of 123,000 images, each with five image descriptions. We apply the framework of Wang et al. [43]. The image is encoded by the VGG model [44] that has been trained on the ImageNet classification task following the standard procedure [43]. The 4096-dimensional feature from the pre-softmax layer is used to represent the image. As such, this process can be viewed as an instance of transfer learning; that is, a representation trained on an image classification task is used for a text-image matching task. The text is represented by the mean of the GloVe [45] vectors of its content words. 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:

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

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

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

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

where for the 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 scores $m(x_i, y_i)$ (cosine similarity) for positive text-image pairs than for negative pairs[3].

After the text-image matching model is trained, for each text-image pair (s_i, p_j) in our task, we can calculate the matching score $m(s_i, p_j)$. 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.

Note that the images in Flickr30K are similar to our task; however, as shown in Fig. 3 (a), the image descriptions in Flickr30K are much simpler than the text in news. Flickr30K often contains short descriptions, i.e., single-sentence captions, for images, while in our dataset, some of the information contained in the news, such as the time and location of events, is not directly reflected by images. Therefore, we believe that the text-image matching model trained on Flickr30K cannot be be directly used for the text in our dataset at the whole sentence level, i.e., we need the ability to match images with text units at a lower level of granularity. To solve this problem, we conduct multi-granularity text-image matching beyond the sentence level. Specifically, before we match textual and visual information using the trained text-image matching model, we split the integrated sentence into coherent substructures to make text-image matching more tractable. We investigate this idea experimentally using frame-semantic parsing, chunking and word tokenization.

3. In the experiments, $K = 50$, $\Delta = 0.1$ and $\lambda_1 = 2$. Wang et al. [43] also proved that structure-preserving constraints can make 1% Recall@1 improvement.

3.5 Multi-granularity Text-Image Matching

In this section, we introduce text-image matching in multiple granularities. We first extract sentence segments at the semantic frame level (Section 3.5.1), chunking level (Section 3.5.2) and token level (Section 3.5.3). Then, for each sentence-image or segment-image pair (s_i, p_j), we calculate the matching score using the trained text-image matching model (Section 3.4). The reason we match the image with text units at smaller granularity rather than the complete sentence is simple: smaller granularities lead to easier text-image matching and smaller granularity text units are more similar to the image descriptions in Flickr30K. These characteristics can be seen in Fig. 3. For example, for the image in Fig. 3 (b), compared with the complete sentence "*On May 9 local time, Russia held grand celebrations marking the 70th anniversary of Russia 's victory in the great patriotic war .*", it is more realistic to match the text units "*Russia held grand celebrations*", "*grand celebrations*" and "*celebrations*".

3.5.1 Semantic Frame Level Text-Image Matching

Frame-semantic parsing [46] plays an important role in semantic analysis. Given a sentence, a frame-semantic parser maps words to the frames they evoke and then, for each frame, labels arguments with frame-specific roles [47].

The basic idea is that each verb in a sentence is labeled with its propositional arguments, and the labeling for each particular verb is called a "frame". Each frame represents an event, and the arguments express the relevant information about this event. There is a set of arguments indicating the semantic role of each term in a frame. For example, ARG0 denotes the agent of the event, and ARG1 denotes the action. We intend to simplify each sentence and speech transcription based on frame-semantic parsing. 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 (hereafter, we call this type of simplified sentence a frame) that is used to match the images. It is worth noting that the semantic role labeler we use in this work is based on PropBank semantic annotation [48]. There may be multiple predicate-argument structures for each sentence, and we extract all of them. An example of frame-semantic parsing is shown in Fig. 4. The original sentence "*President Bush authorized federal disaster assistance for the affected areas and made plans for an inspection tour of the state*" is transformed into two simplified sentences "*President Bush authorized federal disaster assistance*" and "*President Bush made plans for an inspection tour of the state*". The simplified sentences have less diversity in meaning, which benefits the text-image matching.

3.5.2 Chunking Level Text-Image Matching

Chunking, which is the process of splitting a long sentence into grammatical non-overlapping constituents, i.e., phrases, is a common technique in NLP. The advantage of chunking is that the extracted phrases are grammatical and

(a)

The players of the baseball team are standing on the field , with many people watching from the stands.
Baseball players are playing on a field in a stadium filled with people .
A team plays baseball at a large crowded stadium .
A crowd cheers on a baseball team .
Game is playing in the stadium

(b)

Original sentence: On May 9 local time, Russia held grand celebrations marking the 70th anniversary of Russia 's victory in the great patriotic war .
Frames: Russia held grand celebrations || Russia marking the 70th anniversary of Russia 's victory ...
Chunks: grand celebrations || the 70th anniversary || the great patriotic war ...
Words: celebrations || marking || anniversary ...

Fig. 3. (a) An instance in Flickr30K. (b) An image and the potential matched text in our dataset. For text units based on frame, chunk and word, we divide different units by "||".

Original sentence: [President Bush]$_{\text{ARG0}}$ [authorized]$_{\text{predicate}}$ [federal disaster assistance]$_{\text{ARG1}}$ [for the affected areas]$_{\text{ARG2}}$ and [made]$_{\text{predicate}}$ [plans for an inspection tour of the state]$_{\text{ARG1}}$.
Simplified sentence1: [President Bush]$_{\text{ARG0}}$ [authorized]$_{\text{predicate}}$ [federal disaster assistance]$_{\text{ARG1}}$
Simplified sentence2: [President Bush]$_{\text{ARG0}}$ [made]$_{\text{predicate}}$ [plans for an inspection tour of the state]$_{\text{ARG1}}$

Fig. 4. An example for simplified sentence based on frame-semantic parsing.

meaningful. Chunks, especially noun phrases (NPs) and verb phrase (VPs), provide a suitable level of abstraction over natural language [49]. An NP is a constituent that can be the subject or object of an event. A VP is a constituent that contains a verb and other associated components, including modal, auxiliary, and modifier. Fig. 3 shows that NPs and VPs provide a simple way to match text and image. In this work, we extract NPs and VPs as meaningful chunks of each sentence and speech transcription to match images.

3.5.3 Token Level Text-Image Matching

A token can be a word (except a stop-word) or a symbol. As the smallest unit that expresses semantics, a token is a straightforward granularity for text-image matching.

3.6 Multi-modal Topic Modeling

After the text-image matching model is trained, we obtain the joint representation of text and images. Next, we identify the topics of text and images. The motivation behind this process is that textual descriptions of images often provide important information about semantic aspects (topics), and image features are often correlated with semantic topics [50]. Wang et al. [51] generate a timeline summarization for Tweet streams by detecting topic evolution. For our task, the multi-modal topic model can reveal various aspects of text and images; then we can explore a representative set of text covering the aspects of the images. Topic models, such as LDA [52], can jointly learn latent topics and topic allocations of documents. To reveal the semantic aspects, we build the multi-modal topic model based on a neural topic model (NTM) [53]. In an NTM, the topic–word distribution is modeled as a look-up layer of words, and the topic–document distribution is modeled as a look-up layer of documents. The output layer of the neural network is given by the dot product of these two distributions. There are two variants of NTM and we adopt the unsupervised variant in our experiments. In contrast to standard topic models such as LDA, NTM can model document topics beyond word unigrams. Specifically, it can address n-grams that are represented with embeddings. We apply NTM to multi-modal topic modeling. For a document or a video d across the entire multi-modal corpus, w is a word (for video, the word is extracted from the speech transcription) or an image in d. The multi-modal topic model calculates the conditional probability $p(w|d)$ using the distribution of the word (or image)-topic $p(w|t)$ and topic-document (or video) $p(t|d)$:

$$p(w|d) = \sum_{i=1}^{T} p(w|t_i)p(t_i|d) \tag{8}$$

where t_i is a latent topic and T is the total number of topics.

Let $D(d) \in \mathbb{R}^{1\times T}$ denotes the distribution of d over all the topics, i.e., $D(d) = [p(t_1|d), ..., p(t_T|d)]$ and $D(w) \in \mathbb{R}^{1\times T}$ denotes the distribution of w over topics, i.e., $D(w) = [p(w|t_1), ..., p(w|t_T)]$. Note that $D(w)$ is shared among the multi-modal corpus. Equation 8 can be represented as follows:

$$D(w) = sigmoid(W(w) \cdot W_2) \tag{9}$$

where $W(w) \in \mathbb{R}^{300}$ denotes the representation of w, and $W_2 \in \mathbb{R}^{300\times T}$. The dot product of $D(w)$ and $D(d)$ represents the conditional probability $p(w|d)$ in Equation 8.

Follow Cao et al. [53], we use pairwise ranking approach to update $D(d)$ and W_2. The loss function is:

$$L_t = \sum_{w} \max[0, \Delta_t + D(w) \cdot D(d)^T - D(w) \cdot D(d^{'})^T] \tag{10}$$

where $d^{'}$ is a document or a video which does not contain w and the margin Δ_t is set to 0.5. The objective of L_t is that the matching score of $D(w)$ and $D(d)$ should be higher for the cases that w appears in d than other document or video which does not contain w.

After the multi-modal topic model is trained, we obtain T topics and the topic distribution for all the words and images in our dataset.

3.7 Measuring the Similarity between Image and Text Based on the Topic Distribution

We first need to obtain the topic-based representation of an image and text to measure the similarity between the image and text based on the topic distribution. For an image or a word w, we can compute its topic distribution $D(w)$ using Equation 9. We represent the text by calculating the average of the topic distribution of all words in the text. The similarity between a sentence and an image can

be expressed by the commonly used Kullback-Leibler (KL) divergence [54], which has been used to identify the latent semantic structure [55], [56]. In this work, we apply the transformed information radius (IR) instead of KL. Given a topic z, the topic distribution of sentence s and image p are denoted as $p_s^{(z)}$ and $p_p^{(z)}$, respectively. The IR divergence between $p_s^{(z)}$ and $p_p^{(z)}$ is calculated based on KL divergence:

$$IR(p_s^{(z)}, p_p^{(z)}) = KL(p_s^{(z)}||\frac{p_s^{(z)} + p_p^{(z)}}{2}) + KL(p_p^{(z)}||\frac{p_s^{(z)} + p_p^{(z)}}{2}) \tag{11}$$

where, $KL(s||p) = \sum_i s_i \log \frac{s_i}{p_i}$. The divergence is transformed into similarity measure [57]:

$$sim_{topic}(s, p) = \frac{1}{N} \sum_{k=1}^{N} 10^{-IR(p_s^{(z=k)}, p_p^{(z=k)})} \tag{12}$$

where N is the total number of topics. The reason we choose IR is that there is no issue with infinite values because $p_s + p_p \neq 0$ if either $p_s \neq 0$ or $p_p \neq 0$[4] and it is also symmetric, i.e., $IR(s, p) = IR(p, s)$.

The similarity score $sim_{topic}(s, p)$ represents how well sentence s covers the topics of image p, and we expect the summary to cover as many important topics related to the images as possible.

3.8 Multi-modal Summarization

We propose an extractive MMS method through which text salience, image coverage and non-redundancy can be jointly optimized.

3.8.1 Text Salience of Summarization

Inspired by the work of Lin and Bilmes [58], we model the text salience of summary S by a diversity-aware objective:

$$\mathcal{F}_s(S) = \sum_{i=1}^{K} \sqrt{\sum_{s_j \in P_i \cap S} Sa(s_j)} \tag{13}$$

where P_i, $i = 1, ..., K$ is a disjoint partition of the set of all the sentences and speech transcriptions V into separate clusters, and the salience scores $s(t_j)$ are normalized to [0,1] by dividing by the maximum value among all the sentences. This objective function rewards salience and diversity because it is more beneficial to choose a sentence from a cluster that does not yet have any of its elements in the summary. If a sentence is chosen from a cluster, other sentences from this cluster have a diminishing gain due to the square root function[5].

To generate P_i, we apply CLUTO [59][6] to cluster the sentences, where the IDF-weighted vector is used as the feature vector, and a K-mean clustering algorithm is applied. In our experiment, we set $K = 0.2|V|$, as do Lin and Bilmes [58].

4. Else, we set $sim_{topic}(s, p) = 0$.

5. In fact, it can be replaced with any other non-decreasing concave functions, such as the logarithmic function.

6. http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

3.8.2 Matching-based Image Coverage of Summarization

We model the summary S coverage for the image set I as follows:

$$\begin{aligned} \mathcal{F}_m(S) &= \sum_{p_i \in I} Im(p_i) \cdot m(p_i, c_j) \\ c_j &= \arg\max_{c_k \in S} m(p_i, c_k) \end{aligned} \tag{14}$$

where the $Im(p_i)$ is the weight for image p_i. For keyframe p_i, $Im(p_i)$ is the average salience score of the speech transcriptions within the shot to which p_i belongs. For document image p_i, $Im(p_i)$ is the average salience score of the sentences in the document in which p_i is embedded. c_j is a sentence or a sentence segment obtained based on semantic framing, chunking or word tokenizing.

This objective function aims to maximize the weighted coverage of the selected images. For an image p_i, only a c_j with the maximum matching score with p_i contributes to the coverage for p_i. That is, when p_i is covered by c_j, no other c_k can further improve the coverage of p_i. The intuition behind this is straightforward: the generated summary is limited in length, and we intend to cover as many important images as possible. According to Equation 14, we consider not only the count of the images covered by the summary but also the overall importance score of the covered images. As a result, we select sentences that are relevant to more important images or that contain more segments relevant to more important images.

3.8.3 Topic-based Image Coverage of Summarization

We model the topic coverage of summary S for image set I as follows:

$$\begin{aligned} \mathcal{F}_t(S) &= \sum_{p_i \in I} Im(p_i) \cdot sim_{topic}(p_i, s_j) \\ s_j &= \arg\max_{s_k \in S} sim_{topic}(p_i, s_k) \end{aligned} \tag{15}$$

This objective function aims to maximize the weighted topic coverage of the selected images. For a specific p_i, similar to matching-based image coverage, only a sentence or speech transcription s_j with the maximum topic-based similarity score with p_i contributes to the coverage for p_i, and no other s_k can further improve the coverage of p_i. Therefore, the generated summary covers as many important topics of the images as possible.

3.8.4 Multi-modal Summarization Objective Function

Finally, considering all the modalities, we design matching-based and topic-based objective functions for image coverage.

Matching-based objective function is defined as follows:

$$\begin{aligned} \mathcal{F}_M(S) = \frac{1}{M_s} \sum_{i=1}^{K} \sqrt{\sum_{s_j \in P_i \cap S} Sa(s_j)} \\ + \frac{1}{M_c} \sum_{p_i \in I} Im(p_i) \cdot m(p_i, c_j) \\ c_j = \arg\max_{c_k \in S} m(p_i, c_k) \end{aligned} \tag{16}$$

where M_s is the summary score obtained by Equation 13 and M_c is the summary score obtained by Equation 14. The

aim of M_s and M_c is to balance the aspects of salience and coverage for images.

Topic-based objective function is defined as follows:

$$\mathcal{F}_T(S) = \frac{1}{M_s}\sum_{i=1}^{K}\sqrt{\sum_{s_j \in P_i \cap S} Sa(s_j)} + \frac{1}{M_t}\sum_{p_i \in I} Im(p_i)\cdot sim_{topic}(p_i, s_j) \quad (17)$$
$$s_j = \arg\max_{s_k \in S} sim_{topic}(p_i, s_k)$$

where M_t is the summary score obtained by Equation 15.

The Equations 13, 14, 15, 16 and 17 are all monotone submodular functions (see proofs in Section 3.8.5) under the budget constraint. Thus, we apply a greedy algorithm [60] guaranteeing near-optimization to solve the problem. Note that an integer linear program (ILP) solver, as used by Lin and Bilmes [60], can also be used to maximize the objective functions, but solving the ILP with approximately 400 sentences takes more than 20 hours, whereas the greedy algorithm requires only 1 second.

3.8.5 Proof of the Monotone Submodularity of Our Objective Functions

Submodularity and monotonicity are two necessary ingredients to guarantee that the greedy algorithm gives near-optimal solutions. We formally define submodularity as follows.

Definition 1. *Submodularity. Given a set X, a set function $f: 2^X \rightarrow \mathbb{R}$ is called submodular if for any two sets A and B such that $A \in B \in X$ and element $x \in X \backslash B$,*
$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$.

Theorem 1. *Given monotone submodular set functions $f_1: 2^X \rightarrow \mathbb{R}$ and $f_2: 2^X \rightarrow \mathbb{R}$, the function $F = \lambda_1 \cdot f_1 + \lambda_2 \cdot f_2: 2^X \rightarrow \mathbb{R}$ is monotone submodular, if $\lambda_1 > 0$ and $\lambda_2 > 0$.*

This property will be useful when combining two monotone submodular functions such as Equation 16 and 17.

Theorem 2. *$\mathcal{F}_s(S)$ is monotone submodular.*

Proof. See the work of Lin and Bilmes [60] □

Theorem 3. *$\mathcal{F}_m(S)$ and $\mathcal{F}_t(S)$ are monotone submodular.*

In this paper, we only give the proof for $\mathcal{F}_m(S)$, and $\mathcal{F}_t(S)$ can be proved in the same way.

Proof. (Monotonicity). Inspired by Kobayashi et al. [61], first, we prove the monotonicity. Next, we prove the submodularity. For simplicity, we use the following abbreviation: $c_A = \arg\max_{c_k \in A} m(p_i, c_k)$. Let the ground set V represent all the sentences and speech transcriptions related to a specific news event. For a sentence set A and a sentence $s \in V \backslash A$, $\mathcal{F}_m(A \cup \{s\}) - \mathcal{F}_m(A) = \sum_{p_i \in I} Im(p_i) \cdot (m(p_i, c_{A\cup\{s\}}) - m(p_i, c_A))$. Since $m(p_i, c_{A\cup\{s\}}) > m(p_i, c_A)$, $\mathcal{F}_m(A \cup \{s\}) - \mathcal{F}_m(A) > 0$ holds.

(Submodularity). For any two sentence sets A and B that $A \in B \in V$, and sentence $s \in V \backslash B$, we have $\mathcal{F}_m(A \cup \{s\}) - \mathcal{F}_m(A) - \mathcal{F}_m(B \cup \{s\}) - \mathcal{F}_m(B) = \sum_{p_i \in I} Im(p_i) \cdot (m(p_i, c_{A\cup\{s\}}) - m(p_i, c_A) - m(p_i, c_{B\cup\{s\}}) + m(p_i, c_B))$. Let $\delta := (m(p, c_{A\cup\{s\}}) - m(p, c_A) - m(p, c_{B\cup\{s\}}) + m(p, c_B))$. Next, we need to prove that $\delta \geq 0$.

There are three cases as follows:

If $c_{B\cup\{s\}} \in s$, then $c_{A\cup\{s\}} \in s$ holds, since $A \cup \{s\} \in B \cup \{s\}$. This means that $m(p, c_{A\cup\{s\}}) = m(p, c_{B\cup\{s\}}) = m(p, c_{\{s\}})$. Obviously, $m(p, c_B) > m(p, c_A)$, since $A \in B$. Therefore, $\delta \geq 0$.

If $c_{B\cup\{s\}} \notin s$ and $c_{A\cup\{s\}} \notin s$, we have $m(p, c_{A\cup\{s\}}) = m(p, c_A)$ and $m(p, c_{B\cup\{s\}}) = m(p, c_B)$. Therefore, $\delta = 0$.

If $c_{B\cup\{s\}} \notin s$ and $c_{A\cup\{s\}} \in s$, we have $m(p, c_{A\cup\{s\}}) \geq m(p, c_A)$ and $m(p, c_{B\cup\{s\}}) = m(p, c_B)$. Therefore, $\delta \geq 0$.

□

Theorem 4. *$\mathcal{F}_M(S)$ and $\mathcal{F}_T(S)$ are monotone submodular.*

Proof. It can be proved by the rule of Theorem 1. □

4 Experiment

4.1 Data Collection and Annotation

There is no benchmark dataset for MMS. We construct a dataset as follows. We select 50 news events from the most recent five years: 25 in English and 25 in Chinese. We set 5 events in each language as the development set. For each event, we collect 20 documents within the same period using Google News search and 5-10 videos from CCTV.com and YouTube. More details about the dataset are given in our previous work [1].

We employ 10 graduate students to write reference summaries after reading documents and watching videos on the same event. We keep 3 reference summaries for each event. The criteria for summarizing documents are (1) retain the important content of the input documents and videos; (2) avoid redundant information; (3) have a good readability; (4) satisfy the length limit. We set the length constraint for the English and Chinese summaries to 300 words and 500 characters, respectively.

4.2 Comparative Methods

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

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

Audio only. This model generates summaries using only the speech transcriptions from videos.

Text + audio. This model generates summaries using the text in documents and the speech transcriptions from videos 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. [62], which won the 2015 MSCOCO Image Captioning Challenge. This model generates summaries using document text, speech transcription and image captions.

Note that the abovementioned methods generate summaries using Equation 13; the following methods use Equations 13 ,14 and 16.

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 matches the sentence.

Image alignment. The images are hard-aligned to the text in the following ways: The images in a document are aligned to all the sentences in that document, and the keyframes in a shot are aligned to all the speech transcriptions in that shot.

The following models match text with image using the approach introduced in Section 3.4 with multi-granular sentence segments as the text units.

Image match sent. The text units are sentences.

Image match frame. The text units are frames, i.e., simplified sentences based on frame-semantic parsing.

Image match chunk. The text units are chunks.

Image match word. The text units are words.

Image match frame+chunk. The text units are frames and chunks.

Image match frame+chunk+sent. The text units are frames, chunks and sentences.

Image match frame+chunk+sent+word. The text units are frames, chunks, sentences and words.

The following two models match text with images based on the multi-modal topic model introduced in Section 3.6 and measure the similarity between an image and text based on the topic distribution with different metrics, i.e., KL and IR. Note that the default topic number we adopt is 200 and we assess different topic numbers in Section 4.7.

Image topic KL. The metric is based on KL.

Image topic IR. The metric is based on IR.

4.3 Implementation Details

We apply Stanford CoreNLP toolkit [63] to perform lexical parsing and use semantic role labelling approach proposed by Yang and Zong [64]. We use the publicly available 300-dimensional skip-gram English word embeddings[7]. We train the Chinese word embeddings using Word2Vec with a corpus containing approximately 600 million words. Given that the 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 and then conduct text and image matching. To apply the **image caption** model to Chinese, we translate the generated English captions for the images into Chinese via Google Translation.

4.4 Multi-modal Summarization Evaluation

We use the ROUGE-1.5.5 toolkit [65] with the parameters [8] suggested by Owczarzak et al. [66], which yield high correlation with human judgments, to evaluate the summaries against the reference summaries. All our ROUGE scores have a 95% confidence interval of at most ± 0.25, as reported by the official ROUGE script. ROUGE scores measure the summary quality by matching n-grams between a generated summary and the reference summary. TABLE 1 shows the average ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) F-scores for the three reference summaries for each event in English and Chinese.

For English MMS, the first four lines in TABLE 1 show that when summarizing without speech transcriptions, the **text+audio+guide** model performs better than the **audio only** model and the **text+audio** model, but there is no obvious advantage over the **text only** model. On one hand, we can conclude that guidance strategies are necessary when generating summaries with audio information; one the other hand, because ROUGE mainly measures word overlap, manual evaluation is needed to confirm the real impact of guidance strategies on readability. This topic is discussed in Section 4.5. The performance is not always improved when summarizing textual and visual modalities, which indicates that the **image caption**, **image caption match**, **image alignment** and **image match word** models are not suitable for MMS. When performing text-image matching with a single granular text unit, the **image match frame** and the **image match chunk** models perform better than the other methods, and combining these two models leads to further improvement (+2.067% R-1, +2.959% R-2, +2.958% R-SU4 over the **text-only** model). The experimental results also illustrate that the **image topic** models can make use of multi-modal information, and the **image topic IR** model achieves a comparable ROUGE score to that of the **image match frame+chunk** model (+2.806% R-1, +2.920% R-2, +3.092% R-SU4 for **image topic IR** model, over the **text-only** model).

The **image caption** and the **image caption match** models heavily depend on the quality of the generated image descriptions; however, the image captioning model mainly focuses on the surface meaning of the image, which is not sufficient for MMS. The **image alignment** model can be regarded as a rough text-image matching model. Images in a document cannot always be aligned to the sentences in that document; similarly, the keyframes in a shot can not always be aligned to the speech transcription of that shot. For the models based on text-image matching, there are generally many possible choices of linguistic components to use as the text unit for text-image matching, such as a sentence, frame, chunk, or word. However, as shown in TABLE 1, a word may be too general to accurately represent the meaning of an image. On the other hand, a sentence may be too specific to accurately capture the general meaning of an image, which could increase the difficulty of text-image matching. Between these two granularities, a simplified sentence based on a frame or a chunk is both general and concise. For the **image topic** models, it is broad enough to capture some latent topic information shared by text and images. We will give more discussions in Section 4.7 and 4.6.

The Chinese MMS results are similar to the English results. Specifically, we find that the **image match frame** model achieves the best ROUGE-2 score, possibly due to properties of neural machine translation, whose output vocabulary size is limited[9], which makes text-image matching

7. https://code.google.com/archive/p/word2vec/

8. -x -u -n 2 -m -2 4 -c 95 -r 1000 -f A -p 0.5 -t 0 -a

9. For example, 80k target words as the target vocabulary for the Google neural machine translation system [67].

TABLE 1
Experimental results (F-score) for English and Chinese MMS task. For Chinese task, we adopt word level evaluation.

Model	English			Chinese		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
Text only	0.41928	0.11342	0.16399	0.40052	0.10955	0.16233
Audio only	0.41283	0.08450	0.14332	0.25539	0.05016	0.08034
Text+audio	0.41773	0.11064	0.16217	0.39877	0.10686	0.15906
Text+audio+guide	0.41746	0.11349	0.16380	0.40095	0.10978	0.16253
Image caption	0.41470	0.10329	0.15946	0.36837	0.08259	0.13851
Image caption match	0.41496	0.10518	0.15893	0.37356	0.08327	0.13920
Image alignment	0.40729	0.08262	0.14072	0.28871	0.06870	0.10115
Image match sent	0.42234	0.11576	0.16706	0.39261	0.11187	0.15939
Image match frame	0.42594	0.13241	0.18088	0.39672	**0.13376**	0.17847
Image match chunk	0.43836	0.12627	0.18063	0.38625	0.12692	0.17367
Image match word	0.41863	0.10640	0.16249	0.37696	0.11470	0.16068
Image match frame+chunk	0.43995	**0.14301**	0.19357	0.38624	0.13011	0.17650
Image match frame+chunk+sent	0.43874	0.13909	0.18980	0.37485	0.11952	0.16437
Image match frame+chunk+sent+word	0.44073	0.13782	0.19085	0.38277	0.11835	0.16525
Image topic KL	0.43813	0.13387	0.18594	0.41066	0.12192	0.17294
Image topic IR	**0.44734**	0.14262	**0.19491**	0.**42540**	0.13232	**0.18513**

easier.

We present an example of the generated summaries in Fig. 5.

4.5 Manual Summary Quality Evaluation

The readability (how easy of understanding) and informativeness (how much important information contained) of summaries are difficult to evaluate formally and they are both highly subjective criteria, whose assessment varies from person to person. We ask five graduate students to measure the quality of summaries generated by different methods. The average scores for the events are shown in TABLE 2. Overall, our method with guidance strategies achieves higher scores than those of the other methods, but the results are clearly worse than the reference summaries. Specifically, the informativeness of the summary is the worst when speech transcriptions are not considered. However, adding speech transcription without guidance strategies decreases the readability to a large extent, which indicates that guidance strategies are necessary for MMS. The **image match frame** model achieves higher informativeness scores than do the other methods without using images.

We present two instances of readability guidance that arise between the document text (DT) and speech transcription (ST) in TABLE 3. The errors introduced by ASR include segmentation (instance A) and recognition (instance B) mistakes.

4.6 How Much is an Image Worth

Text-image matching is the most challenging module of our framework. Although we use a state-of-the-art approach to match text and images, the performance is far from satisfactory. To determine a somewhat strong upper bound for the task, we choose five events for each language and manually label the text-image matching pairs. The MMS results on these events are shown in TABLE 4. The experiments show that with the ground truth text-image matching result, the quality of the English summary is greatly improved, which indicates visual information is crucial for MMS. Note that

TABLE 2
Manual summary quality evaluation.

	Method	Readability	Informativeness
English	Text only	3.72	3.28
	Text + audio	3.08	3.44
	Text + audio + guide	3.68	3.64
	Image match frame	3.67	3.83
	Image topic IR	3.80	4.10
	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 frame	3.62	3.92
	Image topic IR	3.73	4.00
	Reference	**4.88**	**4.84**

TABLE 3
Guidance examples. "CST" denotes manually modified correct ST. ASR errors are marked red and revisions are marked blue.

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.

for Chinese, the **image match chunk** model achieves comparable performance to the **image manual** model, which again demonstrates that neural machine translation decoding with limited vocabulary size will improve text-image matching.

Two images and the corresponding text obtained using different methods are shown in Fig. 6. We can conclude that the **image caption** and the **image caption match** models convey little of the image's intrinsic information. The **image alignment** model introduces more noise because it is

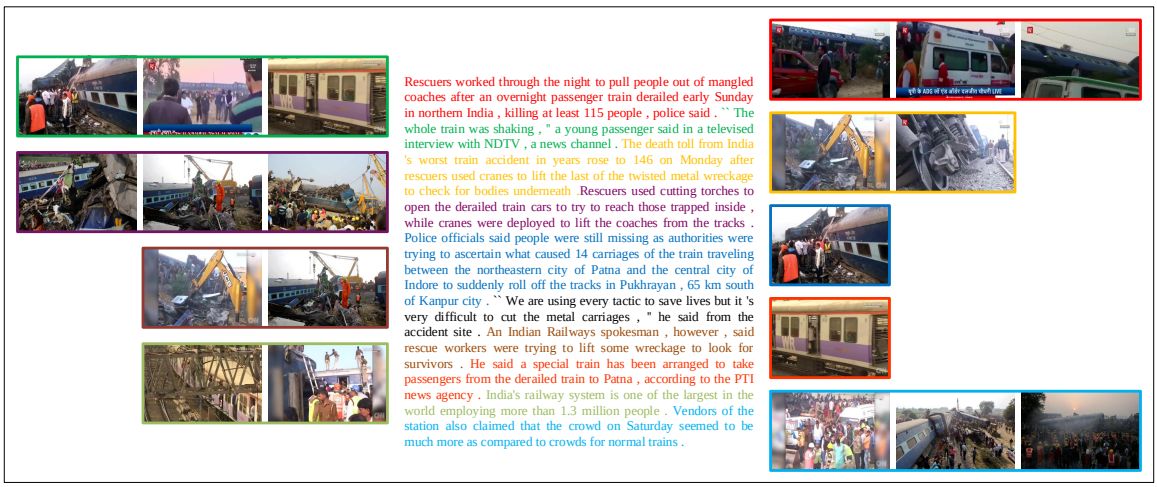


Fig. 5. An example of generated summary with the **image match frame+chunk** model for the news event "India train derailment". The sentences covering the images are labeled by the corresponding colors. The text can be partly related to the image because we use sentence segments to match the images.

TABLE 4
Experimental results (F-score) for English and Chinese MMS task on five events with manually labeled text-image pairs.

Model	English			Chinese		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
Text+audio+guide	0.39619	0.08724	0.14552	0.39719	0.11150	0.16318
Image caption	0.38798	0.08006	0.13550	0.39299	0.10018	0.15824
Image caption match	0.36093	0.06042	0.12015	0.36665	0.08026	0.13431
Image alignment	0.40011	0.06240	0.12813	0.26186	0.03914	0.07565
Image match sent	0.41768	0.11588	0.16907	0.38668	0.11633	0.16404
Image match frame	0.41959	0.13096	0.18330	0.39129	0.12449	0.17410
Image match chunk	0.42741	0.12516	0.18330	0.39128	**0.13895**	**0.18322**
Image match word	0.41381	0.10578	0.16230	0.35340	0.07719	0.12840
Image match frame+chunk	0.42886	0.13853	0.18943	0.37297	0.11233	0.16358
Image topic KL	0.43022	0.12846	0.18200	0.41952	0.12552	0.18231
Image topic IR	0.44001	0.12805	0.18818	0.40239	0.11633	0.17119
Image manual	**0.44521**	**0.16236**	**0.20702**	**0.40473**	0.13796	0.18085

possible for the entire text in a document or the speech transcription in a shot are aligned to the document images or the keyframes, respectively. The **image match** and **image topic** models produce results similar to those of the **image manually match**, which illustrates that these models can make use of visual information to generate summaries.

However, which modality is more important, textual or visual? To answer this question, we define two α-weighted objective functions to give different weights to textual and visual information.

We define an α-weighted matching-based objective function as follows:

$$\begin{aligned} \mathcal{F}_M(S) &= \frac{\alpha}{M_s}\mathcal{F}_s(S) + \frac{1-\alpha}{M_c}\mathcal{F}_m(S) \\ c_j &= \arg\max_{c_k \in S} m(p_i, c_k) \end{aligned} \tag{18}$$

where $\mathcal{F}_s(S)$ and $\mathcal{F}_s(S)$ are defined in Equation 13 and Equation 14, respectively.

We define an α-weighted topic-based objective function as follows:

$$\begin{aligned} \mathcal{F}_T(S) &= \frac{\alpha}{M_s}\mathcal{F}_s(S) + \frac{1-\alpha}{M_t}\mathcal{F}_t(S) \\ s_j &= \arg\max_{s_k \in S} sim_{topic}(p_i, s_k) \end{aligned} \tag{19}$$

where $\mathcal{F}_t(S)$ is defined in Equation 15.

The experimental results are shown in Figs. 8 and 9. The best performance is achieved when α is approximately 0.5, which suggests that textual and visual information are almost equally important to MMS. More precisely, the models with α greater than 0.5 perform slightly better than the models with α less than 0.5. Considering that visual contents, including videos and images in documents, tend to be the highlights of the events, which can be regard as "summarized information" to some degree, we should pursue more effective ways to use the visual information in the future, such as filtering the noise in the visual information.

4.7 Discussion on Multi-modal Topic Model

To better understand the nature of the multi-modal topic model, we present two images with the top 5 ranked topics generated by the multi-modal topic model shown in Fig. 7.

The image on the left is from a news event about the "Sewol Ferry Disaster", and the multi-modal topic model recognizes topics related to water (**Topic 2**) and rescue (**Topic 4**). The image on the right is from a news event about the "Savar building collapse", and the multi-modal topic model recognizes topics related to shop (**Topic 1 and 2**) and warehouse (**Topic 3 and 4**).

Fig. 7. Example images with topics generated by the multi-modal topic model. For each image, we present the top five ranked topics, which are represented with top three ranked words.

Image caption:
A group of men standing next to each other.
Image caption match:
A tall, adult man stood by a young man who was confronting riot police.
Image alignment:
The police made 235 arrests .
Image match sent:
About 10 fire crews battled a three-alarm fire at a large senior center under construction at Chester and gay streets , as police officers stood guard with long guns .
Image match frame:
Even as 2,000 National Guard troops and police from multiple states came in to secure the streets, more protests were planned.
Image match chunk:
There are still 3,000 Maryland national guard members , standing in tan camouflage uniforms and armed with assault rifles , spread across Baltimore 's streets .
Image match word:
Thousands of police officers and national guardsmen have been patrolling the city .
Image topic IR: Late night tensions between police and protesters briefly lit up this Baltimore neighborhood Tuesday.
Image manually match:
The governor sent 500 state troopers to Baltimore and requested as many as 5,000 officers from neighboring states.

Fig. 6. An example image with corresponding English text that different methods obtain. The matched text segments are marked red.

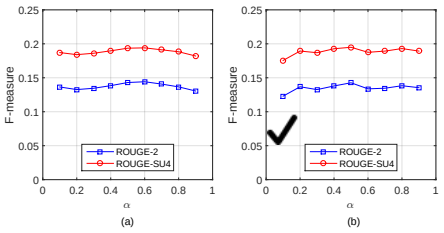


Fig. 8. Performance (F-score) comparison on English MMS task with various α. (a) The **image match arg+chunk** model. (b) The **image topic IR** model.

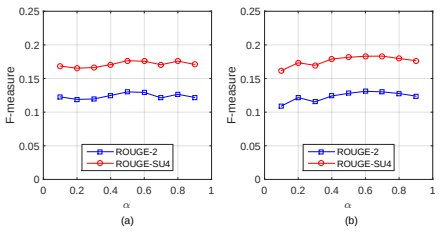


Fig. 9. Performance (F-score) comparison on Chinese MMS task with various α. (a) The **image match arg+chunk** model. (b) The **image topic IR** model.

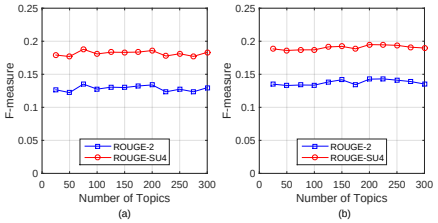


Fig. 10. Performance (F-score) comparison of different numbers of topics for English MMS based on the multi-modal topic model with different metrics to measure the similarity: (a) KL metric, (b) IR metric.

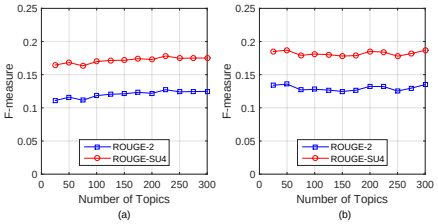


Fig. 11. Performance (F-score) comparison of different numbers of topics for Chinese MMS based on the multi-modal topic model with different metrics to measure the similarity: (a) KL metric, (b) IR metric.

Furthermore, Figs. 10 and 11 plot the performance of different numbers of topics ranging from 25 to 300. The robust performance on both English and Chinese MMS tasks suggests that our model is almost insensitive to the number of topics, and 200 is a reasonable choice.

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. We address readability by selectively using the transcription of audio through guidance strategies. More specifically, we design a novel graph-based 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 text, and find that the **image match** model and the **image topic** model perform well for the MMS task. The final experimental results obtained using our MMS corpus in both English and Chinese demonstrate that our system benefits from multi-modal information.

Adding audio and video does not appear to dramatically improve the performance with respect to the text only model, which indicates that better models are needed to capture the interactions between text and other modalities, especially for visual information. We also plan to expand 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. 61673380.

REFERENCES

[1] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal summarization for asynchronous collection of text, image, audio and video." in *EMNLP*, 2017, pp. 1092–1102.

[2] B. Erol, D.-S. Lee, and J. Hull, "Multimodal summarization of meeting recordings," in *ICME*, vol. 3. IEEE, 2003, pp. III–25.

[3] R. Gross, M. Bett, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel, "Towards a multimodal meeting record," in *ICME*, vol. 3. IEEE, 2000, pp. 1593–1596.

[4] D. Tjondronegoro, X. Tao, J. Sasongko, and C. H. Lau, "Multi-modal summarization of key events and top players in sports tournament videos," in *WACV*. IEEE, 2011, pp. 471–478.

[5] T. Hasan, H. Bořil, A. Sangwan, and J. H. Hansen, "Multi-modal highlight generation for sports videos using an information-theoretic excitability measure," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, p. 173, 2013.

[6] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[7] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multimodal stereoscopic movie summarization conforming to narrative characteristics," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5828–5840, 2016.

[8] D. Wang, T. Li, and M. Ogihara, "Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs." in *AAAI*, 2012.

[9] W. Y. Wang, Y. Mehdad, D. R. Radev, and A. Stent, "A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization," in *NAACL-HLT*, 2016, pp. 58–68.

[10] M. Del Fabro, A. Sobe, and L. Böszörmenyi, "Summarization of real-life events based on community-contributed content," in *The Fourth International Conferences on Advances in Multimedia*, 2012, pp. 119–126.

[11] J. Bian, Y. Yang, and T.-S. Chua, "Multimedia summarization for trending topics in microblogs," in *CIKM*. ACM, 2013, pp. 1807–1812.

[12] M. Schinas, S. Papadopoulos, G. Petkos, Y. Kompatsiaris, and P. A. Mitkas, "Multimodal graph-based event detection and summarization in social media streams," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 189–192.

[13] J. Bian, Y. Yang, H. Zhang, and T.-S. Chua, "Multimedia summarization for social events in microblog stream," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 216–228, 2015.

[14] R. R. Shah, A. D. Shaikh, Y. Yu, W. Geng, R. Zimmermann, and G. Wu, "Eventbuilder: Real-time multimedia event summarization by visualizing social media," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 185–188.

[15] R. R. Shah, Y. Yu, A. Verma, S. Tang, A. D. Shaikh, and R. Zimmermann, "Leveraging multimodal information for event summarization and concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 108, pp. 102–109, 2016.

[16] S. Khuller, A. Moss, and J. S. Naor, "The budgeted maximum coverage problem," *Information Processing Letters*, vol. 70, no. 1, pp. 39–45, 1999.

[17] V. Varma, V. Varma, and V. Varma, "Sentence position revisited: a robust light-weight update summarization 'baseline' algorithm," in *International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, 2009, pp. 46–52.

[18] Y. Ouyang, W. Li, Q. Lu, and R. Zhang, "A study on position information in document summarization," in *COLING*, 2010, pp. 919–927.

[19] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing Management*, vol. 40, no. 6, pp. 919–938, 2004.

[20] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *ACL*, 2004.

[21] X. Wan and J. Yang, "Improved affinity graph based multi-document summarization," in *NAACL*, 2006, pp. 181–184.

[22] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Qiqihar Junior Teachers College*, vol. 22, p. 2004, 2011.

[23] X. Zhou, X. Wan, and J. Xiao, "Cminer: Opinion extraction and summarization for chinese microblogs," *IEEE Transactions on Knowledge & Data Engineering*, vol. 28, no. 7, pp. 1650–1663, 2016.

[24] X. Li, L. Du, and Y. D. Shen, "Update summarization via graph-based sentence ranking," *IEEE Transactions on Knowledge & Data Engineering*, vol. 25, no. 5, pp. 1162–1174, 2013.

[25] P. Goyal, L. Behera, and T. M. Mcginnity, "A context-based word indexing model for document summarization," *IEEE Transactions on Knowledge & Data Engineering*, vol. 25, no. 8, pp. 1693–1705, 2013.

[26] H. Li, J. Zhang, Y. Zhou, and C. Zong, "Guiderank: A guided ranking graph model for multilingual multi-document summarization," in *NLPCC-ICCPOL*. Springer, 2016, pp. 608–620.

[27] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998.

[28] P. Li, J. Ma, and S. Gao, "Learning to summarize web image and text mutually," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. ACM, 2012, p. 28.

[29] Z. Li, J. Tang, X. Wang, J. Liu, and H. Lu, "Multimedia news summarization in search," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 3, p. 33, 2016.

[30] R. Valenza, T. Robinson, M. Hickey, and R. Tucker, "Summarisation of spoken audio through information extraction," in *ESCA Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio*, 1999.

[31] M. G. Christel, M. A. Smith, C. R. Taylor, and D. B. Winkler, "Evolving video skims into useful multimedia abstractions," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 1998, pp. 171–178.

[32] S. Dagtas and M. Abdel-Mottaleb, "Extraction of tv highlights using multimedia features," in *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*. IEEE, 2001, pp. 91–96.

[33] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004.

[34] C. Quirk, C. Brockett, and W. B. Dolan, "Monolingual machine translation for paraphrase generation." in *EMNLP*, 2004, pp. 142–149.

[35] Y. Yang, F. Shen, Z. Huang, H. T. Shen, and X. Li, "Discrete nonnegative spectral clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1834–1845, 2017.

[36] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 154–162.

[37] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE transactions on cybernetics*, vol. 45, no. 5, pp. 1083–1094, 2015.

[38] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 10, pp. 2085–2098, 2015.

[39] Z. Li and J. Tang, "Weakly supervised deep matrix factorization for social image understanding," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 276–288, 2017.

[40] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Image Processing, 1998. ICIP 98.*, vol. 1. IEEE, 1998, pp. 866–870.

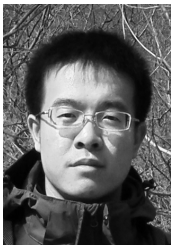
[41] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[42] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *ECCV*, 2014.

[43] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *CVPR*, 2016, pp. 5005–5013.

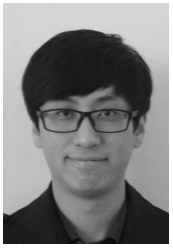
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[45] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.
[46] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Computational linguistics*, vol. 28, no. 3, pp. 245–288, 2002.
[47] M. Kshirsagar, S. Thomson, N. Schneider, J. Carbonell, A. N. Smith, and C. Dyer, "Frame-semantic role labeling with heterogeneous annotations," in *ACL*, 2015, pp. 218–224.
[48] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational linguistics*, vol. 31, no. 1, pp. 71–106, 2005.
[49] C. Arora, M. Sabetzadeh, L. Briand, F. Zimmer, and R. Gnaga, "Automatic checking of conformance to requirement boilerplates via text chunking: An industrial case study," in *Empirical Software Engineering and Measurement*. IEEE, 2013, pp. 35–44.
[50] D. Mahajan, S. Sellamanickam, S. Sanyal, and A. Madaan, "A classification based framework for concept summarization," in *ICDM*. IEEE, 2012, pp. 1008–1013.
[51] Z. Wang, L. Shou, K. Chen, G. Chen, and S. Mehrotra, "On summarization and timeline generation for evolutionary tweet streams," *IEEE Transactions on Knowledge & Data Engineering*, vol. 27, no. 5, pp. 1301–1315, 2015.
[52] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, 2003.
[53] Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji, "A novel neural topic model and its supervised extension." in *AAAI*, 2015.
[54] S. Kullback, "Letter to the editor: The kullback-leibler distance," 1987.
[55] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440.
[56] I. Vulić, W. De Smet, and M.-F. Moens, "Identifying word translations from comparable corpora using latent topic models," in *ACL*, 2011, pp. 479–484.
[57] C. D. Manning, H. Schütze *et al.*, *Foundations of statistical natural language processing*. MIT Press, 1999, vol. 999.
[58] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *ACL*, 2011, pp. 510–520.
[59] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, no. 1. Boston, 2000, pp. 525–526.
[60] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in *NAACL*, 2010, pp. 912–920.
[61] H. Kobayashi, M. Noguchi, and T. Yatsuka, "Summarization based on embedding distributions," in *EMNLP*, 2015, pp. 1984–1989.
[62] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, 2016.
[63] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *ACL*, 2003.
[64] H. Yang and C. Zong, "Multi-predicate semantic role labeling," in *EMNLP*, 2014, pp. 363–373.
[65] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *NAACL*, 2003.
[66] K. Owczarzak, M. J. Conroy, T. H. Dang, and A. Nenkova, *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, 2012, ch. An Assessment of the Accuracy of Automatic Evaluation in Summarization, pp. 1–9.
[67] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.



Haoran Li received the B.S. degree from Shandong University, Jinan, China, in 2013, and is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include summarization, multimedia, and natural language processing.

Junnan Zhu received the B.S. degree from Central South University, Changsha, China, in 2015, and is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include summarization and natural language processing.



Cong Ma received the B.S. degree from University of Science and Technology Beijing , Beijing, China, in 2017, and is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include summarization, sentiment analysis and natural language processing.

Jiajun Zhang received the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, in 2011, Beijing, China. He is currently an Associate Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include machine translation, multilingual natural language processing and deep learning. He served area chair for COLING-2018 and SPC for IJCAI-2017, IJCAI-2018 and AAAI-2019. He was selected in Young Elite Scientists Sponsorship Program by CAST in 2015.

Chengqing Zong received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1998. He is a Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include natural language processing, machine translation, and sentiment analysis. He is a Member of International Committee on Computational Linguistics. He is an Associate Editor of the ACM Transactions on Asian and Low-Resource Language Information Processing and an Editorial Board Member of IEEE Intelligent Systems, the journal Machine Translation, and the Journal of Computer Science and Technology. He served ACL-IJCNLP2015 as PC co-Chair, IJCAI'2017, IJCAI-ECAI'2018 and AAAI'2019 as area chair.