A Picture Paints a Thousand Words:
a Method of Generating Image-text Timelines

Shize Xu  
xsz@pku.edu.cn

Liang Kong  
kongliang@pku.edu.cn

Yan Zhang†  
zhy@cis.pku.edu.cn

Department of Machine Intelligence, Peking University, Beijing, China  
Key Laboratory on Machine Perception, Ministry of Education, Beijing, China

ABSTRACT

Manual timelines have greatly helped us to keep pace with the big world. In this paper, we introduce a novel solution which generates image-text timelines for news events based on Evolutionary Image-Text Summarization, which is an important and challenging problem. We first extract image’s semantic information under translation model, and then fuse the high quality images with text timeline under an image assignment algorithm which can optimize the global coordination of the final timeline. The experimental results show that news readers can receive more satisfaction from the image-text timelines we generate.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Image-text, Cross-modality, Summarization, Timeline

1. INTRODUCTION

With the rapid spreading of Web2.0, the news reports consist of various types of information, especially text information and image information. Since the social rhythm is quickening, news consumers need a more concise and convenient way to perceive directly. But given such large texts and images collection related to a specified news subject (i.e., Occupy Wall Street), readers usually want to simply find out the current progresses, causes and effects.

Benefitting from the researches on automatic multi-document summarization, we can generate a description for a specific event. However, almost all of the existing work focuses on one specific modality, text or image. Analysis and consolidation of inter-modality are academically novel. Illustrated summary can give a more concise presentation for an event.

Timeline could be a good display form with abbreviated, informative reorganization for faster and better browsing. The websites usually spend a lot of human cost on providing image-text timelines. Learning from manual timelines, we can draw two requisite features for a text-image timeline. First, from inner-modality view, both texts or images should present valuable information in one day and keep a coherent temporal sequence among different dates. Second, from inter-modality view, texts and images should talk under the same event and informatively replenish each other. Studying on how to combine these two kinds of resources is of practical significance.

In this paper, we introduce a novel solution for the automatic image-text evolutionary summarization problem to meet the challenges mentioned above. For the challenge to inner-modality, we generate text timeline under a trans-temporal framework while keeping a good data individuality and correlativeness. For the challenge to inter-modality, we propose a model employing GLOBALOPTIMIZEDIMAGE-ASSIGNMENT method to enrich the text timeline with the high quality images. The experimental results conducted on the real data from 6 famous news websites show that our approach can outperform the rivals effectively.

2. RELATED WORK

To our best knowledge, this paper is one of the few innovative researches that steps into cross-modality summarization field, though there are kinks of research focusing on summarization of single domain. Great efforts have been made to generate text timeline but they ignore the important role of images[3]. Contrariwise, some focus on visual modality[11].

To mine the latent knowledge of image, visual features are described as visual words in [6] and a probabilistic model is proposed based on the assumption that images and their co-occurring textual data are generated by mixtures of latent topics. Keyphrases usually denote the semantic features of image, Feng et al. operate over the output of a probabilistic image annotation model that preprocesses the pictures and suggests keywords to describe their content[5]. Image retrieval has also been used to enrich text content, i.e., Agrawal et al. propose techniques for finding images from the web and attach to a text section for augmenting[1].

Unlike previous studies, our method can combine modalities of text and image in a harmony.
3. METHODOLOGY

3.1 Preprocessing

To reveal the perceptive relevance between images and texts, we first model the semantic features of images.

Images will be preprocessed so that they can be represented by word-like units. Local image descriptors are computed using the SIFT algorithm[7] and subsequently quantized into a discrete set of visual words via a clustering algorithm such as K-means. Formally, each image is expressed in a bag-of-words format vector \([wv_1, wv_2, \ldots, wv_n]\), where \(wv_i = n\) only if the image has \(n\) regions denoted by \(v_i\). Then we manage to translate visual words into textual words through statistical translation model. In our dataset, most news images have news texts around them. On an intuitive level, the text contexts can tell what these images are talking about, and thus well define the semantic content of the images. Given a set of news webpages’ html docs, VIPS[2] and WebKit\(^1\) rendering tool help us to segment pages and pick those text blocks whose coordinates are neighboring to each specified image. We treat these text contexts as image’s semantic annotations. The model now works with a bag-of-words representation and treats each image-annotation pair as a single document consisting of textual and visual words. Given a visual word \(vw\) and a textual annotation word \(w\), based on their co-occurrences we estimate \(P(w|vw)\) using maximum likelihood estimation with Laplace smoothing[4]:

\[
P(w|vw) = \frac{\sum_i n(wv_i, w_i) + \lambda}{\sum_i n(vw_i) + \lambda V}
\]

where \(n(wv_i, w_i)\) denotes the frequency of \(v_w\) and \(w\) is the \(i\)th image-annotation pair, \(n\) denotes the number of pairs in training set and \(V\) denotes the visual vocabulary size. \(\lambda\) is the smooth parameter which makes our model insensitive to data noise and is fixed to be 1. Then given an unknown image we can translate it with its visual words into a set of semantic features, which in practical form is an integral probability distribution of textual words.

3.2 Image mining based on text timeline

Intuitively, a required summary needs to be informational rich with both global biased and local biased view. That means the target summary should be correlative with sentences from today as well as neighboring dates, which will lead to significant temporal continuity. ETTS[12] offers a good process and we extend its work to generate text timelines. We conduct a temporal proximity projection procedure to achieve global biased characteristic.

To generate the component summary on date \(t\), besides considering the sentences with timestamp \(t\), we also emphatically regard the sentences with different timestamps. All sentences in the collection are projected onto the time horizon of \(t\) to construct a global affinity graph with Gaussian kernels \(\Gamma(\Delta t)[8]\), where \(\Delta t = |t - t'|\) is the distance between the pending date \(t\) and neighboring date \(t'\):

\[
\Gamma(\Delta t) = \exp\left[\frac{-\Delta t^2}{2\sigma^2}\right]
\]

Parameter \(\sigma\) will be tuned in later experiments, which controls the spreading of kernel curves. After DivRank[9] is applied, we can estimate the local rank as well as the global rank of each sentence and then calculate the merged rank with linear weighting, which can be proved to be optimal. Thus the text timeline is generated.

We will assign images to the given text timeline. Formally, given the image and sentence collection \(I, S\) partitioned by the timestamp set \(T, I_i = \{I_i|1 \leq i \leq |I_i|, \}\), and \(S_t = \{S_i|1 \leq i \leq |S_i|\}\), where \(S_i\) and \(I_i\) is a sentence and an image with the timestamp \(t \in T\). For a specific date \(t\), we compute the target affinity matrices \(M = (m_{ij})_{|I_i|\times|S_i|}\) to measure the semantic relevance between image and sentence. Since our translation model can offer a semantic feature for every image, presented as \(I_i \rightarrow \{p(w_1), p(w_2), \ldots, p(w_{|W|})\}\), where \(p(w_i)\) denotes the semantic transition probability of image \(I_i\) to word \(w\) and \(W\) is the vocabulary shared by both modalities. We can calculate the semantic similarity \(sim(I_i, S_j)\) between image \(I_i\) and sentence \(S_j\):

\[
sim(I_i, S_j) = \sum \frac{p(w_i) \cdot \pi_j(w)}{\sqrt{\sum \frac{p(w_i) \cdot isf(w)}{\sum \frac{p(w_i) \cdot isf(w)}}}}
\]

\(\pi_j(w)\) is the translation probability from image \(I_i\) to word \(w\) with \(isf\) weighting. \(\pi(w)\) denotes the \(tf \cdot isf\) weighting of word \(w\), and \(tf(w)\) is the term frequency of word \(w\) in \(S_i\). \(isf\) is the inverse-sentence-frequency to diminish the weight of common words and increase the discrimination:

\[
isf(w) = (1 + \log \frac{|S|}{N_w})
\]

where \(N_w\) is the number of sentences or images containing word \(w\). \(m_{ij}\) denotes that the relevance between image \(I_i\) and sentence \(S_j\) can be derived as follow. And like the handling of sentences, temporal weighting is added to the similarity.

\[
m_{ij} = \Gamma(|t - t_i|) \cdot \Gamma(|t - t_j|) \cdot \frac{\Gamma(\Delta t)}{\Gamma(|\Delta t|)}
\]

Note that normalization is employed on the matrix to make the sum of each row equal to 1. In general, the optimal setting of \(\sigma\) may vary according to the dataset because sentences and images would have wider semantic scope on news timeline, so that a higher value of \(\sigma\) is required. And vice versa.

Each image \(I_i\) calculates its worthiness score \(\mathcal{R}(I_i)\) based on the ordered sentences. Specifically, given text summary of date \(t\) which has size of \(n_t\), image \(I_i\’s\) score \(\mathcal{R}(I_i|t)\) can be estimated as follow:

\[
\mathcal{R}(I_i|t) = \sum_{j=1}^{n_t} \frac{m_{ij}}{\log_2(1 + j)}
\]

3.3 Image-text Timeline Generation

Now we conduct a optimal image assignment method as the final step of our summarization work. On an intuitive level, this problem admits a natural greedy algorithm. For every date \(t\), all images (include today’s and neighboring days’) are sort by their scores \(\mathcal{R}(I_i|t)\), then more valuable images are picked out to make up summary for date \(t\) and no longer selectable for the other dates. But consider the following situation. On two specific dates \(t_1\) and

\(^1\)http://www.webkit.org
4. EXPERIMENTS

4.1 Datasets

We randomly choose 8 news subjects with special coverage and handcrafted timelines by editors from 6 selected news websites: New York Times, BBC, CNN, China Daily, Reuters and Yahoo! News. The manually-edited timelines are employed as our golden reference standards to evaluate the proposed systems empirically. Table 1 shows the details.

<table>
<thead>
<tr>
<th>Event (Query)</th>
<th>Article</th>
<th>Sentence</th>
<th>Image</th>
<th>Time Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>James and Bebe</td>
<td>708</td>
<td>38970</td>
<td>583</td>
<td>Mar 9-May 16, 2011</td>
</tr>
<tr>
<td>Occupy Wall Street</td>
<td>598</td>
<td>32472</td>
<td>469</td>
<td>Sept 17-Nov 4, 2011</td>
</tr>
<tr>
<td>MH370 Search</td>
<td>122</td>
<td>30655</td>
<td>351</td>
<td>Jul 1-Nov 27, 2015</td>
</tr>
<tr>
<td>Murdoch’s Leaked Text</td>
<td>219</td>
<td>15148</td>
<td>175</td>
<td>June 21-Aug 31, 2011</td>
</tr>
<tr>
<td>Euro 2012 Qualifiers</td>
<td>262</td>
<td>13044</td>
<td>256</td>
<td>Jun 1-Nov 16, 2012</td>
</tr>
</tbody>
</table>

* We use their abbreviations (JE, SJH, KJI, OWS, NBA, ME, EQ and RE) to denote above datasets for a more concise show.

4.2 Evaluation Metrics

There are no acknowledged metrics to judge image-text timeline’s quality, so we build an evaluation system which includes two golden indexes, QDCG and SDCG. We employ five volunteers as the news readers to rate the quality of images showed to them in timelines with three levels, judging if the images fit the text well. The levels range from 1 to 3 where level 3 denotes the best. To evaluate the whole timeline with the index, we estimate average quality discounted cumulative gain (QDCG) scores as follows:

\[
QDCG = \frac{1}{T} \sum_{t \in T} \sum_{i=1}^{K_t} \frac{q_{rate}^{Q}(i|t) - 1}{\log_2(1 + i)}
\]

where \(q_{rate}^{Q}(i|t)\) denotes the rating level of the ith image in date t’s timeline. The main idea about the usage of logarithm is that more important sentences pick more valuable images. What’s more, in this way images and sentences can be well organized to help reading. Besides judging the image quality, we are concerned with the attitude of readers to the inserted images. So the judges quantize how the images influence their reading experience with three grades at the same time. An image may get high grade if it does well in complementing the information as well as increasing reading enjoyment. We also estimate average satisfaction discounted cumulative gain (SDCG) as follows:

\[
SDCG = \frac{1}{T} \sum_{t \in T} \sum_{i=1}^{K_t} \frac{2^{q_{rate}^{S}(i|t)} - 1}{\log_2(1 + i)}
\]

4.3 System Setup and Baselines

The unit of timestamp is one day. The upper limit \(K_t\) of the number of picked images are fixed to 8 for every \(t \in T\) after considering datasets’ size and browsing experience. We tentatively set \(\sigma = 10\), and tune them later. Since the image-text summarization work is novel among current researches, it becomes a challenge to highlight our method’s effectiveness. We focus on the following image-text summarization algorithms as the baselines.

Randomly assignment. The method randomly picks and assigns the images with the specified timestamp.

Image retrieval assignment. The method runs text summarization first. Based on the sentences in the summarization, we use COMITY \([1]\) to mine images from web after forming key phrases for each day. Then the most relevant images are assigned to the text timeline.

Greedy assignment. This approach shares the same \(\mathcal{R}(I|t)\) for each image as our method, but the images with high scores are selected by the first date and removed directly from the candidate list for the other dates. The same strategy is applied to the following dates.

4.4 Results

Figure 1 shows that our ranking framework can outperform the others. Randomly assignment has the worst performance as expected. Retrieval-based gets a poor score of SDCG due to that it picks many discordant images when taking no textual context influence into account. It’s worth noting that Greedy assignment gets a much better result than the previous two and we can verify the superiority of translation model and worthiness score estimating.

Further more, we can see that our assignment approach defeats Greedy by a significant upgrade. This result verifies
the effectiveness of our GlobalOptimizedImageAssignment algorithm. Note that our Optimized method serves the last five datasets more better. Other methods play especially poorly on those datasets which contain less burstiness, as Optimized is still maintaining high quality. Now we can draw an inspiring conclusion. With modeling the temporal continuity, cross-modality and maximize assigning score at the same time, our method can get the most satisfying timelines of summaries over the other methods.

4.5 Parameter Tuning and Sample Output

The key parameter \( \sigma \) measures the temporal projection influence from global candidates to local candidates and hence the size of neighbors. Figure 2 shows that the events with high burstiness prefer a smaller value to dominate the influence from neighbors. In the opposite we get converse results. We can see that the summarization method to some extent is sensitive to the data. Developing a parameter adaptive framework can be one of our future work.

At the end, we analyze our generated timelines. We notice that for all datasets we can highlight the material facts of events and keep information diversity at the same time. Table 2 shows two-days timeline (discontinuously) of “Occupy Wall Street” as the sample output.

5. CONCLUSIONS

We study the feasibility of automatically generating timelines of image-text summaries and present a solution for this problem. Our method exploits the internal relationships between image and text under our translation model. We generate text timeline with time individuality and correlative-

Table 2: Part timeline of “Occupy Wall Street”

<table>
<thead>
<tr>
<th>Date</th>
<th>Image</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct 27, 2011</td>
<td><img src="image2.png" alt="Image" /></td>
<td>2. More than 700 protesters were arrested on Saturday on Brooklyn Bridge.</td>
</tr>
<tr>
<td>Oct 28, 2011</td>
<td><img src="image3.png" alt="Image" /></td>
<td>3. Protesters dressed as “corporate zombies”, in full zombie regalia and clutching fake cash, parade down Wall Street.</td>
</tr>
<tr>
<td>Oct 29, 2011</td>
<td><img src="image4.png" alt="Image" /></td>
<td>4. Billionaire investor George Soros says he can sympathize with the ongoing protests on Wall Street, which have spread to other US cities.</td>
</tr>
<tr>
<td>Oct 30, 2011</td>
<td><img src="image5.png" alt="Image" /></td>
<td>5. A large rally is planned for Wednesday in New York.</td>
</tr>
</tbody>
</table>

6. REFERENCES