Reader-Aware Multi-Document Summarization: An Enhanced Model and The First Dataset

Abstract

We investigate the problem of reader-aware multi-document summarization (RA-MDS) and introduce a new dataset for this problem. To tackle RA-MDS, we extend a variational auto-encodes (VAEs) based MDS framework by jointly considering news documents and reader comments. To conduct evaluation for summarization performance, we prepare a new dataset. We describe the methods for data collection, aspect annotation, and summary writing as well as scrutinizing by experts. Experimental results show that reader comments can improve the summarization performance, which also demonstrates the usefulness of the proposed dataset. The annotated dataset for RA-MDS is available online.

Introduction

The goal of multi-document summarization (MDS) is to automatically generate a brief, well-organized summary for a topic which describes an event with a set of documents from different sources. BIBREF0 , BIBREF1 , BIBREF2 , BIBREF3 , BIBREF4 , BIBREF5 , BIBREF6 . In the typical setting of MDS, the input is a set of news documents about the same topic. The output summary is a piece of short text document containing several sentences, generated only based on the input original documents.

With the development of social media and mobile equipments, more and more user generated content is available. Figure FIGREF2 is a snapshot of reader comments under the news report “The most important announcements from Google's big developers' conference”. The content of the original news report talks about some new products based on AI techniques. The news report generally conveys an enthusiastic tone. However, while some readers share similar enthusiasms, some others express their worries about new products and technologies and these comments can also reflect their interests which may not be very salient in the original news reports. Unfortunately, existing MDS approaches cannot handle this issue. We investigate this problem known as reader-aware multi-document summarization (RA-MDS). Under the RA-MDS setting, one should jointly consider news documents and reader comments when generating the summaries.

One challenge of the RA-MDS problem is how to conduct salience estimation by jointly considering the focus of news reports and the reader interests revealed by comments. Meanwhile, the model should be insensitive to the availability of diverse aspects of reader comments. Another challenge is that reader comments are very noisy, not fully grammatical and often expressed in informal expressions. Some previous works explore the effect of comments or social contexts in single document summarization such as blog summarization BIBREF7 , BIBREF8 . However, the problem setting of RA-MDS is more challenging because the considered comments are about an event which is described by multiple documents spanning a time period. Another challenge is that reader comments are very diverse and noisy. Recently, BIBREF9 employed a sparse coding based framework for RA-MDS jointly considering news documents and reader comments via an unsupervised data reconstruction strategy. However, they only used the bag-of-words method to represent texts, which cannot capture the complex relationship between documents and comments.

Recently, BIBREF6 proposed a sentence salience estimation framework known as VAESum based on a neural generative model called Variational Auto-Encoders (VAEs) BIBREF10 , BIBREF11 . During our investigation, we find that the Gaussian based VAEs have a strong ability to capture the salience information and filter the noise from texts. Intuitively, if we feed both the news sentences and the comment sentences into the VAEs, commonly existed latent aspect information from both of them will be enhanced and become salient. Inspired by this consideration, to address the sentence salience estimation problem for RA-MDS by jointly considering news documents and reader comments, we extend the VAESum framework by training the news sentence latent model and the comment sentence latent model simultaneously by sharing the neural parameters. After estimating the sentence salience, we employ a phrase based compressive unified optimization framework to generate a final summary.

There is a lack of high-quality dataset suitable for RA-MDS. Existing datasets from DUC and TAC are not appropriate. Therefore, we introduce a new dataset for RA-MDS. We employed some experts to conduct the tasks of data collection, aspect annotation, and summary writing as well as scrutinizing. To our best knowledge, this is the first dataset for RA-MDS.

Our contributions are as follows: (1) We investigate the RA-MDS problem and introduce a new dataset for the problem of RA-MDS. To our best knowledge, it is the first dataset for RA-MDS. (2) To tackle the RA-MDS, we extend a VAEs-based MDS framework by jointly considering news documents and reader comments. (3) Experimental results show that reader comments can improve the summarization performance, which also demonstrates the usefulness of the dataset.

Overview

As shown in Figure FIGREF7 , our reader-aware news sentence salience framework has three main components: (1) latent semantic modeling; (2) comment weight estimation; (3) joint reconstruction. Consider a dataset INLINEFORM0 and INLINEFORM1 consisting of INLINEFORM2 news sentences and INLINEFORM3 comment sentences respectively from all the documents in a topic (event), represented by bag-of-words vectors. Our proposed news sentence salience estimation framework is extended from VAESum BIBREF6 , which can jointly consider news documents and reader comments. One extension is that, in order to absorb more useful information and filter the noisy data from comments, we design a weight estimation mechanism which can assign a real value INLINEFORM4 for a comment sentence INLINEFORM5 . The comment weight INLINEFORM6 is integrated into the VAEs based sentence modeling and data reconstruction component to handle comments.

Reader-Aware Salience Estimation

Variational Autoencoders (VAEs) BIBREF10 , BIBREF11 is a generative model based on neural networks which can be used to conduct latent semantic modeling. BIBREF6 employ VAEs to map the news sentences into a latent semantic space, which is helpful in improving the MDS performance. Similarly, we also employ VAEs to conduct the semantic modeling for news sentences and comment sentences. Assume that both the prior and posterior of the latent variables are Gaussian, i.e., INLINEFORM0 and INLINEFORM1 , where INLINEFORM2 and INLINEFORM3 denote the variational mean and standard deviation respectively, which can be calculated with a multilayer perceptron (MLP). VAEs can be divided into two phases, namely, encoding (inference), and decoding (generation). All the operations are depicted as follows: DISPLAYFORM0

Based on the reparameterization trick in Equation EQREF9 , we can get the analytical representation of the variational lower bound INLINEFORM0 : DISPLAYFORM0

where INLINEFORM0 denotes a general sentence, and it can be a news sentence INLINEFORM1 or a comment sentnece INLINEFORM2 .

By feeding both the news documents and the reader comments into VAEs, we equip the model a ability of capturing the information from them jointly. However, there is a large amount of noisy information hidden in the comments. Hence we design a weighted combination mechanism for fusing news and comments in the VAEs. Precisely, we split the variational lower bound INLINEFORM0 into two parts and fuse them using the comment weight INLINEFORM1 : DISPLAYFORM0

The calculation of INLINEFORM0 will be discussed later.

The news sentence salience estimation is conducted by an unsupervised data reconstruction framework. Assume that INLINEFORM0 are INLINEFORM1 latent aspect vectors used for reconstructing all the latent semantic vectors INLINEFORM2 . Thereafter, the variational-decoding progress of VAEs can map the latent aspect vector INLINEFORM3 to INLINEFORM4 , and then produce INLINEFORM5 new aspect term vectors INLINEFORM6 : DISPLAYFORM0

VAESum BIBREF6 employs an alignment mechanism BIBREF12 , BIBREF13 to recall the lost detailed information from the input sentence. Inspired this idea, we design a jointly weighted alignment mechanism by considering the news sentence and the comment sentence simultaneously. For each decoder hidden state INLINEFORM0 , we align it with each news encoder hidden state INLINEFORM1 by an alignment vector INLINEFORM2 . We also align it with each comments encoder hidden state INLINEFORM3 by an alignment vector INLINEFORM4 . In order to filter the noisy information from the comments, we again employ the comment weight INLINEFORM5 to adjust the alignment vector of comments: DISPLAYFORM0

The news-based context vector INLINEFORM0 and the comment-based context vector INLINEFORM1 can be obtained by linearly blending the input hidden states respectively. Then the output hidden state can be updated based on the context vectors: DISPLAYFORM0

Then we can generate the updated output aspect vectors based on INLINEFORM0 . We add a similar alignment mechanism into the output layer.

INLINEFORM0 , INLINEFORM1 , and INLINEFORM2 can be used to reconstruct the space to which they belong respectively. In order to capture the information from comments, we design a joint reconstruction approach here. Let INLINEFORM3 be the reconstruction coefficient matrix for news sentences, and INLINEFORM4 be the reconstruction coefficient matrix for comment sentences. The optimization objective contains three reconstruction terms, jointly considering the latent semantic reconstruction and the term vector space reconstruction for news and comments respectively: DISPLAYFORM0

This objective is integrated with the variational lower bound of VAEs INLINEFORM0 and optimized in a multi-task learning fashion. Then the new optimization objective is: DISPLAYFORM0

where INLINEFORM0 is a set of all the parameters related to this task. We define the magnitude of each row of INLINEFORM1 as the salience scores for the corresponding news sentences.

We should note that the most important variable in our framework is the comment weight vector INLINEFORM0 , which appears in all the three components of our framework. The basic idea for calculating INLINEFORM1 is that if the comment sentence is more similar to the news content, then it contains less noisy information. For all the news sentences INLINEFORM2 and all the comment sentences INLINEFORM3 , calculate the relation matrix INLINEFORM4 by: DISPLAYFORM0

Then we add an average pooling layer to get the coefficient value for each comment sentence: DISPLAYFORM0

Finally, we add a sigmoid function to adjust the coefficient value to INLINEFORM0 : DISPLAYFORM0

Because we have different representations from different vector space for the sentences, therefore we can calculate the comment weight in different semantic vector space. Here we use two spaces, namely, latent semantic space obtained by VAEs, and the original bag-of-words vector space. Then we can merge the weights by a parameter INLINEFORM0 : DISPLAYFORM0 where INLINEFORM0 and INLINEFORM1 are the comment weight calculated from latent semantic space and term vector space. Actually, we can regard INLINEFORM2 as some gates to control the proportion of each comment sentence absorbed by the framework.

Summary Construction

In order to produce reader-aware summaries, inspired by the phrase-based model in BIBREF5 and BIBREF9 , we refine this model to consider the news sentences salience information obtained by our framework. Based on the parsed constituency tree for each input sentence, we extract the noun-phrases (NPs) and verb-phrases (VPs). The overall objective function of this optimization formulation for selecting salient NPs and VPs is formulated as an integer linear programming (ILP) problem: DISPLAYFORM0

where INLINEFORM0 is the selection indicator for the phrase INLINEFORM1 , INLINEFORM2 is the salience scores of INLINEFORM3 , INLINEFORM4 and INLINEFORM5 is co-occurrence indicator and the similarity a pair of phrases (INLINEFORM6 , INLINEFORM7) respectively. The similarity is calculated with the Jaccard Index based method. In order to obtain coherent summaries with good readability, we add some constraints into the ILP framework. For details, please refer to BIBREF14 , BIBREF5 , and BIBREF9 . The objective function and constraints are linear. Therefore the optimization can be solved by existing ILP solvers such as simplex algorithms BIBREF15 . In the implementation, we use a package called lp_solve.

Data Description

In this section, we describe the preparation process of the dataset. Then we provide some properties and statistics.

Background

The definition of the terminology related to the dataset is given as follows.

Topic: A topic refers to an event and it is composed of a set of news documents from different sources.

Document: A news article describing some aspects of the topic. The set of documents in the same topic typically span a period, say a few days.

Category: Each topic belongs to a category. There are 6 predefined categories: (1) Accidents and Natural Disasters, (2) Attacks (Criminal/Terrorist), (3) New Technology, (4) Health and Safety, (5) Endangered Resources, and (6) Investigations and Trials (Criminal/Legal/Other).

Aspect: Each category has a set of predefined aspects. Each aspect describes one important element of an event. For example, for the category “Accidents and Natural Disasters”, the aspects are “WHAT”, “WHEN”, “WHERE”, “WHY”, “WHO_AFFECTED”, “DAMAGES”, and “COUNTERMEASURES”.

Aspect facet: An aspect facet refers to the actual content of a particular aspect for a particular topic. Take the topic “Malaysia Airlines Disappearance” as an example, facets for the aspect “WHAT” include “missing Malaysia Airlines Flight 370”, “two passengers used passports stolen in Thailand from an Austrian and an Italian.” etc. Facets for the aspect “WHEN” are “ Saturday morning”, “about an hour into its flight from Kuala Lumpur”, etc.

Comment: A piece of text written by a reader conveying his or her altitude, emotion, or any thought on a particular news document.

Data Collection

The first step is to select topics. The selected topics should be in one of the above categories. We make use of several ways to find topics. The first way is to search the category name using Google News. The second way is to follow the related tags on Twitter. One more useful method is to scan the list of event archives on the Web, such as earthquakes happened in 2017 .

For some news websites, in addition to provide news articles, they offer a platform to allow readers to enter comments. Regarding the collection of news documents, for a particular topic, one consideration is that reader comments can be easily found. Another consideration is that all the news documents under a topic must be collected from different websites as far as possible. Similar to the methods used in DUC and TAC, we also capture and store the content using XML format.

Each topic is assigned to 4 experts, who are major in journalism, to conduct the summary writing. The task of summary writing is divided into two phases, namely, aspect facet identification, and summary generation. For the aspect facet identification, the experts read and digested all the news documents and reader comments under the topic. Then for each aspect, the experts extracted the related facets from the news document. The summaries were generated based on the annotated aspect facets. When selecting facets, one consideration is those facets that are popular in both news documents and reader comments have higher priority. Next, the facets that are popular in news documents have the next priority. The generated summary should cover as many aspects as possible, and should be well-organized using complete sentences with a length restriction of 100 words.

After finishing the summary writing procedure, we employed another expert for scrutinizing the summaries. Each summary is checked from five linguistic quality perspectives: grammaticality, non-redundancy, referential clarity, focus, and coherence. Finally, all the model summaries are stored in

XML files.

Data Properties

The dataset contains 45 topics from those 6 predefined categories. Some examples of topics are “Malaysia Airlines Disappearance”, “Flappy Bird”, “Bitcoin Mt. Gox”, etc. All the topics and categories are listed in Appendix SECREF7 . Each topic contains 10 news documents and 4 model summaries. The length limit of the model summary is 100 words (slitted by space). On average, each topic contains 215 pieces of comments and 940 comment sentences. Each news document contains an average of 27 sentences, and each sentence contains an average of 25 words. 85% of non-stop model summary terms (entities, unigrams, bigrams) appeared in the news documents, and 51% of that appeared in the reader comments. The dataset contains 19k annotated aspect facets.

Dataset and Metrics

The properties of our own dataset are depicted in Section SECREF28 . We use ROUGE score as our evaluation metric BIBREF16 with standard options. F-measures of ROUGE-1, ROUGE-2 and ROUGE-SU4 are reported.

Comparative Methods

To evaluate the performance of our dataset and the proposed framework RAVAESum for RA-MDS, we compare our model with the following methods:

RA-Sparse BIBREF9 : It is a framework to tackle the RA-MDS problem. A sparse-coding-based method is used to calculate the salience of the news sentences by jointly considering news documents and reader comments.

Lead BIBREF17 : It ranks the news sentences chronologically and extracts the leading sentences one by one until the length limit.

Centroid BIBREF18 : It summarizes clusters of news articles automatically grouped by a topic detection system, and then it uses information from the centroids of the clusters to select sentences.

LexRank BIBREF1 and TextRank BIBREF19 : Both methods are graph-based unsupervised framework for sentence salience estimation based on PageRank algorithm.

Concept BIBREF5 : It generates abstractive summaries using phrase-based optimization framework with concept weight as salience estimation. The concept set contains unigrams, bigrams, and entities. The weighted term-frequency is used as the concept weight.

We can see that only the method RA-Sparse can handle RA-MDS. All the other methods are only for traditional MDS without comments.

Experimental Settings

The input news sentences and comment sentences are represented as BoWs vectors with dimension INLINEFORM0 . The dictionary INLINEFORM1 is created using unigrams, bigrams and named entity terms. INLINEFORM2 and INLINEFORM3 are the number of news sentences and comment sentences respectively. For the number of latent aspects used in data reconstruction, we let INLINEFORM4 . For the neural network framework, we set the hidden size INLINEFORM5 and the latent size INLINEFORM6 . For the parameter INLINEFORM7 used in comment weight, we let INLINEFORM8 . Adam BIBREF20 is used for gradient based optimization with a learning rate 0.001. Our neural network based framework is implemented using Theano BIBREF21 on a single GPU.

Results on Our Dataset

The results of our framework as well as the baseline methods are depicted in Table TABREF40 . It is obvious that our framework RAVAESum is the best among all the comparison methods. Specifically, it is better than RA-Sparse significantly (INLINEFORM0), which demonstrates that VAEs based latent semantic modeling and joint semantic space reconstruction can improve the MDS performance considerably. Both RAVAESum and RA-Sparse are better than the methods without considering reader comments.

Further Investigation of Our Framework

To further investigate the effectiveness of our proposed RAVAESum framework, we adjust our framework by removing the comments related components. Then the model settings of RAVAESum-noC are similar to VAESum BIBREF6 . The evaluation results are shown in Table TABREF42 , which illustrate that our framework with reader comments RAVAESum is better than RAVAESum-noC significantly(INLINEFORM0).

Moreover, as mentioned in VAESum BIBREF6 , the output aspect vectors contain the word salience information. Then we select the top-10 terms for event “Sony Virtual Reality PS4”, and “`Bitcoin Mt. Gox Offlile”' for model RAVAESum (+C) and RAVAESum-noC (-C) respectively, and the results are shown in Table TABREF43 . It is obvious that the rank of the top salience terms are different. We check from the news documents and reader comments and find that some terms are enhanced by the reader comments successfully. For example, for the topic “Sony Virtual Reality PS4”, many readers talked about the product of “Oculus”, hence the word “oculus” is assigned a high salience by our model.

Case Study

Based on the news and comments of the topic “Sony Virtual Reality PS4”, we generate two summaries with our model considering comments (RAVAESum) and ignoring comments (RAVAESum-noC) respectively. The summaries and ROUGE evaluation are given in Table TABREF45 . All the ROUGE values of our model considering comments are better than those ignoring comments with large gaps. The sentences in italic bold of the two summaries are different. By reviewing the comments of this topic, we find that many readers talked about “Oculus”, the other product with virtual reality techniques. This issue is well identified by our model and select the sentence “Mr. Yoshida said that Sony was inspired and encouraged to do its own virtual reality project after the enthusiastic response to the efforts of Oculus VR and Valve, another game company working on the technology.”.

Conclusions

We investigate the problem of reader-aware multi-document summarization (RA-MDS) and introduce a new dataset. To tackle the RA-MDS, we extend a variational auto-encodes (VAEs) based MDS framework by jointly considering news documents and reader comments. The methods for data collection, aspect annotation, and summary writing and scrutinizing by experts are described. Experimental results show that reader comments can improve the summarization performance, which demonstrate the usefulness of the proposed dataset.