ICTNET at Temporal Summarization Track TREC 2013

Qian Liu[1,2], Yue Liu[1], Dayong Wu[1], Xueqi Cheng[1]

1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190

2. University of Chinese Academy of Sciences, Beijing, 100190

liuqian@software.ict.ac.cn

{liuyue,cxq}@ict.ac.cn

Abstract

This paper describes our participation in temporal summarization track of TREC2013. All runs are submitted for both two tasks, namely sequential update summarization task and value tracking task. A real-time framework was proposed based on a trigger model. This model consists of two parts. One is selecting the relevant documents by searching on the document titles. The other is obtaining import sentences to an event. Using the KBA 2013 English-and-unknown-language stream corpus, the experimental results show the effectiveness of our approach.

1 Introduction

This is the first year of temporal summarization track. We were provided with a large collection of timestamped documents from a variety of news and social media sources covering the time period October 2011 through January 2013. The goal of the first task is to generate sequential update summarization, which are useful, new and timely sentence-length updates about a developing event[1]. The event refers to a temporally acute topic, and each topic contains the start time and end time. There are five types of events: accident, bombing, earthquake, shooting and storm. The goal of the second task is to track the value of important event-related attributes. If the estimate changes, the system should return the new value as well as the identifier of the supporting sentence. The attributes include deaths, injuries, displaced, financial impact and locations. Formally, given the time-ordered corpus, the keyword query and the relevant time range, our system outputs a list of sentence identifiers.

2 Our Approach

2.1 Data preprocessing

The standard dataset is TREC KBA 2013 Stream Corpus[2], however, we use a cleansed and smaller version of corpus: the KBA 2013 'english-and-unknown-language' stream corpus[3]. Since the total size of this dataset is still very large, we need to preprocess it for timeliness. We decipher the corpus using a standard gpg and XZ decompression. The corpus has been split into ten components such as news, weblog and review. We only focus on documents labeled with news and discard other types of documents. This is reasonable because we find that the given topics are all about the significant news events.

2.2 sequential update summarization task

Although there is related work based on Wikipedia[4], we propose a real-time framework to guarantee the low latency. That is iterating over the corpus only once in temporal order, and outputting the results of all topics simultaneously. Our processes can be spited into two steps: For each document, we first determine whether it is relevant to any query. Second, the sentences containing trigger words, such as kill, die and injure, are selected from matched documents. In the first step, we conduct some

experiments and find that the search performance using document titles is much better than that using full documents. Therefore, we only check the titles of news. Specifically, if a title covers all words of a query, we think this document is matched with the topic. Otherwise, the relevance grade is zero and skipping it.In the second step, a set of trigger words are learned from training data. After stemming, we select nouns and verbs as trigger words. In addition to these words, the synonymies can be extended using WordNet. The real-time framework is shown as follows:

Algorithm 1

Input: stream of documents
Input: topic queries
Input: onset time of each event
Output: list of sentence identifiers

```
Initialize: RESULT={}
for each document d in corpus do
  if d.timestamp < event end time and d.timestamp > event start time then
      for each query q do
        if d.title and q are matched then
          Select sentences using trigger words and add them to T
          Discard sentences from T which include more than 50 words
          Compute the similarity between sentences in T and existing sentences in
           RESULT and remove the redundancy sentences from T
          Add T to RESULT
        end if
      end for
  end if
end for
return RESULT
```

The simHash algorithm is used to compute the similarity between a new sentence and sentences in result list. Note that, we increase the weights of numerics, because they are very important in a catastrophic event. For example, the number of victims in earthquake becomes higher and higher as time goes on, but the sentences are written in the same expression patterns.

2.3 value tracking task

The algorithm of value tracking task is similar to that shown above. The trigger words are the names of five types of attributes. Each document is processed in sequence. For a sentences match a certain attribute name, then we check to see if the surrounding words of attribute contain numeric value. The size of window is set to five words. After that, we extract the number as the value of this attribute. Lastly, if the value is no change in the existing values, skipping it and continue processing sentences. Otherwise, adding the new attribute value to existing value list.

3 Results

We submitted total three runs for two tasks: ICTNET_run1, ICTNET_run2 and ICTNET_ValueTask. The temporal summarization track is the first year, thus there are no previous work we can compare with. For run1, we remove duplicated sentences based on literal similarity. And we compute semantic similarity using WordNet in run2. The results are as follows.

QID	Query	Category	#Updates	#Nuggets
1	2012 Buenos Aires Rail Disaster	Human accident	23	56
2	2012 Pakistan garment factory fires	Human accident	59	89
3	2012 Aurora shooting	School shooting	42	139
4	Wisconsin Sikh temple shooting	School shooting	51	97
5	Hurricane Isaac (2012)	Weather	77	108
6	Hurricane Sandy	Weather	92	419
8	Typhoon Bopha	Weather	79	88
9	2012 Guatemala earthquake	Earthquake	62	45
10	2012 Tel Aviv bus bombing	Terrorist	15	38

Figure 1. The comparison of ICTNET_run1 updates and gold standard updates (nuggets).

QID	Query	Category	#Updates	#Nuggets
1	2012 Buenos Aires Rail Disaster	Human accident	23	56
2	2012 Pakistan garment factory fires	Human accident	57	89
3	2012 Aurora shooting	School shooting	41	139
4	Wisconsin Sikh temple shooting	School shooting	48	97
5	Hurricane Isaac (2012)	Weather	75	108
6	Hurricane Sandy	Weather	91	419
8	Typhoon Bopha	Weather	78	88
9	2012 Guatemala earthquake	Earthquake	58	45
10	2012 Tel Aviv bus bombing	Terrorist	15	38

Figure 2. The comparison of ICTNET_run2 updates and gold standard updates (nuggets).

Run ID		Expected Latency Gain Latency	Comprehensiveness
ICTNET_run1	AVG	0.1249	0.2525
	STD	0.0755	0.1688
	MIN	0.0100	0.0234
	MAX	0.2777	0.5372
ICTNET_run2	AVG	0.1270	0.2512
	STD	0.0752	0.1693
	MIN	0.0100	0.0227
	MAX	0.2777	0.5372

Table 1. The comparison of ICTNET_run1 and ICTNET_run2.

4 Conclusion

This paper reports a trigger-based real-time framework and technical scheme for two tasks in TREC 2013 Temporal Summarization Track. Most methods are straightforward, and the most indicative finding is that filtering out irrelevant documents may boost the performance in the next procedure of sentence selection. In this year, we only search on the field of document title for matching the relevant events as early as possible. In the future, we will consider more information on the level of sentence.

5 Acknowledgements

We would like to thank all organizers and assessors of TREC and NIST. This work is sponsored by NSF of China Grants No. 61100083, and by 242 Program of China Grants No. 2013G130, and by the National Key Technology R&D Program (2012BAH39B04).

6 Reference

[1] J. Aslam, F. Diaz, M. Ekstrand-Abueg, V. Pavlu, and T. Sakai. *Temporal Summarization*. Available: http://www.trec-ts.org/trec_ts2013_planning.pdf

[2] *TREC KBA streamcorpus*. Available: s3://aws-publicdatasets/trec/kba/kba-streamcorpus-2013-v0_2_0/

[3] *KBA english-and-unknown-language corpus*. Available: s3://aws-publicdatasets/trec/kba/kba-streamcorpus-2013-v0_2_0-english-and-unknown-language/

[4] M. Georgescu, D. D. Pham, N. Kanhabua, S. Zerr, S. Siersdorfer, and W. Nejdl, "Temporal summarization of event-related updates in wikipedia," in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 281-284.