BJUT at TREC 2013 Temporal Summarization Track

Zhen YANG, Fei YAO, Huayang SUN, Yun ZHAO, Yingxu LAI, Kefeng FAN
{yangzhen, laiyingxu}@bjut.edu.cn, {S201207027, sunhuayang, sanzhiyu}@emails.bjut.edu.cn, kefengfan@163.com
College of Computer Science, Beijing University of Technology, Beijing 100124, China

***Abstract*—This paper describes the first participation of BJUT in the TREC Temporal Summarization Track 2013. Since this is the first track which is held on temporal summarization, the traditional text retrieval framework is introduced to solve the newly emerging temporal summarization problem at first, and the conventional approach is found that it doesn't work without any consideration on extra expansion information to lose the retrieval limits. Therefore, the baseline is improved by considering the expansion information over the summarization, which includes the use of query expansion based on time/similarity factors, summarization based on information clusters and so on. We do not intend to identify specific methods for solutions. Rather a list of method is presented in capabilities where it is anticipated the methods are likely to adapt over time. Surprisingly, we find the traditional text retrieval methods with default parameters, such as tf-idf model, BM25 model, perform very well and can be used in many areas. Meanwhile some expansion information methods, such as k-means, show complex performance and their parameters need to be chosen carefully to achieve better performance.**

***Index Terms*—Temporal summarization, Information retrieval, Query expansion**

I. Introduction

The TREC Temporal Summarization Track run for the first time in this year, and its goal is to develop systems that allow users to efficiently monitor the information associated with an event over time. The TREC KBA 2013 Stream Corpus is used as evaluation data in this track. This corpus consists of a set of time stamped documents from various news and social media sources covering the time period October 2011 through January 2013.There are two tasks in TREC 2013 Temporal Summarization track:

- TASK1: Temporal Summarization. In this task, a system should emit relevant and novel sentences to an event (exact metrics will be released in a separate document).
- TASK2: Value Tracking. In this task, a system should emit accurate attribute value estimates for an event.

In TASK1, the effectiveness of retrieval is viewed as the basic standard [1]. The TASK2 is based on the TASK1 to feedback more accurate information to the users. By comparing TASK1 with the effectiveness of TASK2, we can evaluate whether the retrieval system can use previous queries and user interactions to improve the search performance.

II. Summarization Based on Query Expansion and Information Clustering

A. The Framework of Temporal Summarization Track

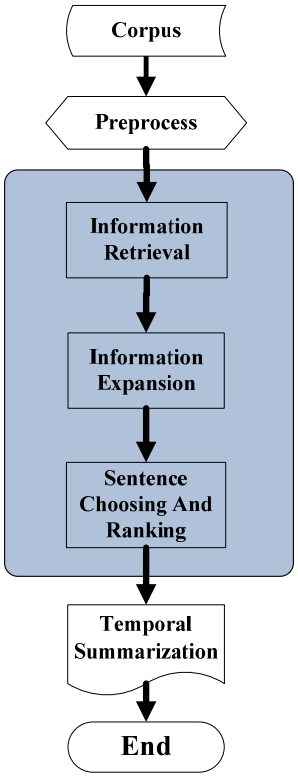

FIG. 1 The framework of the temporal summarization track

The experiments in this paper are carried out at TREC 2013 temporal summarization track. Traditional information retrieval method is proposed to solve the temporal summarization track by following reasons:

- The mission of the temporal summarization can be described as a retrieval task.
- The retrieval model is able to find information which is relevant to event efficiently.

However, only retrieval is not enough, because there isn't much extra information to lose the retrieval limit. So an information expansion method is necessary. We choose clustering as the information expansion method.

As shown in FIG. 1, the framework of our temporal summarization system can be described as follows, which includes preprocess and index module, information retrieval module, information expansion module and sentence choosing and ranking module.

- Preprocess and index module. The GPG file format is converted to TXT file format at first, and then the index is built in these files.
- Retrieval module. Lemur search [2] is the search engine for the search process in our experiment. Lemur search service enables the user to submit the queries and obtain top documents returned by Lemur search engine. Query expansion and term weighting can be applied in Indri search. This is the reason why Lemur search is used to be retrieval module.
- Information expansion module. K-means [3] clustering is a method which is popular for cluster analysis in data mining. It is used to be an information expansion method in extending the retrieval result.
- Sentence choosing and ranking module. After the topic clustering, the centers of the different clustering are chosen to build the summarizations. Then the summarizations are ranked by time factor and similarity factor.

The frame with solid line in FIG. 1 is the main method that we used for the temporal summarization track. The details of our work will be introduced in next section. We will mainly describe two key parts: information retrieval module and information expansion module.

B. Preprocessing and Index Building

The KBA 2013 'English-and-unknown-language' corpus is approximately 4.5TB. For post usage, the original KBA corpus is needed to be preprocessed. The overall general process is described as follows:

- Decrypt File. First step is to decrypt the files using the authorized key from authority. This step converts the GPG file format to SC file format.
- Parse File. We use streamcorpus toolbox to parse these SC files to TXT files. The streamcorpus toolbox is given by TREC and provides a common data interchange format for document processing pipelines that apply language processing tools to large streams of text.
- Build Index. The last step is to build index by lemur for the information retrieval module.

C. Retrieval Module

We use the Lemur Project as a tool for information retrieval, which enables users to build language model and retrieve information.

The following two methods are mainly used in our retrieval module.

TFIDF Method

In the tf-idf method, the value increases proportionally to the number of times of which a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control the fact that some words are generally more common than others. The tf-idf method is used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

BM25 Method

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. One of the most prominent instantiations of the function is as follows.

Given a query Q, containing keywords q_1, q_2,…,q_n, the BM25 score of a document D is:

$$score(D,Q)=\sum_{i=1}^{n} IDF(q_i)\cdot\frac{f(q_i,D)(k_1+1)}{f(q_i,D)+k_1(1-b+b\frac{|D|}{avgdl})} \quad (1)$$

Where $f(q_i,D)$ is q_i's term frequency in the document D, $|D|$ is the length of the document D in words, and *avgdl* is the average document length in the text collection from which documents are drawn. k_1 and b are free parameters, usually chosen, in absence of an advanced optimization, as $k_1\in[1.2, 2.0]$ and b=0.75. $IDF(q_i)$ is the *IDF* weight of the query term q_i.

D. Information Expansion Module

Considering only traditional retrieval method may not perform well, we add clustering based on topic as an information expansion method to build the summarization.

As for clustering, K-means clustering is chosen after many experiments. K-means clustering is a method which is popular for cluster analysis in data mining [4]. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

By using clustering, we can get the clusters based on topics between different events for information expansion. We choose the centers of the clusters and the top sentences as the summarization. Finally each event we totally choose about 50 sentences from the thousands of the results which be tracked by Indri.

The last step is to rank these central sentences. Time and similarity are the two factors which are used to rank the summarizations. After this step, the final temporal summarization can be obtained.

III. Experimental Results

There are two parts of the results in our temporal summarization works: sequential update summarization result and value tracking result.

TABLE 1 Value Tracking Results

QueryID	TeamID	Attribute	Mean Error	Mean Average Error	Min Average Error	Max Average Error
1	BJUT	deaths	**2.847222**	50117.68881	2.84722	116382.853703
2	BJUT	deaths	53.70388	3111.513596	51.3073	10117.172959
3	BJUT	deaths	**0.812780**	87.620126	0.81278	332.904932
4	BJUT	deaths	**0.118716**	19335.04489	0.118716	36168.070876
5	BJUT	deaths	24.51977	332.458379	22.0181	737.899523
6	BJUT	deaths	**133.2798**	66477.22422	133.280	154793.943626
8	BJUT	deaths	974.866	141499.5733	949.822	477973.053624
9	BJUT	deaths	**4.999738**	2635.058405	4.99974	7027.642681
10	BJUT	deaths	47.74999	112.353104	1.97465	238.993469

A. Sequential Update Summarization

For each SUS task submitted to the track, the evaluation file contains the per-topic evaluation scores as well as the overall mean. In addition, the file contains the min/max/mean scores for each topic and overall as computed across the 26 SUS runs submitted to the track.

The script and data used to compute the scores are posted in the temporal summarization track's section of the Tracks page in the active participants' part of the TREC web site. The data files include:

- Nuggets - The initial extracted relevant text
- Pooled Updates - A listing of which updates were pooled
- Matches - The matches between nuggets and updates

However now there are no overlaps between our stream ids and the sampled. But we have sent the sentence text to make sure if they can match for information between ours and the gold information set.

B. Value Tracking

Value tracking result is based on sequential update summarization result. It is retrieved from the summarization. The evaluation results are shown in Table 1.

The data in Table 1 are the results of the value tracking mission. In the table are 7 values including QueryID, TeamID, Attribute, Mean Error, Mean Average Error, Min Average Error and Max Average Error. The results which are marked with bold size show that our system performs comparable or better to the best automatic runs submitted to TREC Temporal Summarization Track 2013. Among the nine topics, five topics obtain the minimum average error.

IV. DISCUSSION

In this notebook, information retrieval and clustering technique were used to solve the temporal summarization problem. As the value tracking result was shown, surprisingly, we find the traditional text retrieval methods with default parameters, such as tf-idf model, BM25 model, perform very well. Considering retrieval and clustering results are both based on similarity measure, we can only find the relevant information around the central sentences or topic. But we think that there are other relevant sentences except these centers. To extracting these summarization sentences need more comprehension from semantic and structural analysis.

REFERENCES

[1] K.S.Jones and E-N Brigitte. Introduction:Automatic Sum marizing. Information Processing & Management, 1995, 3 1(5): 625−630.

[2] http://www.lemurproject.org

[3] http://en.wikipedia.org/wiki/K-means_clustering

[4] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys, 1999, 31(3): 264−323.