The Information Extraction systems of BUPT_PRIS at TREC2014 Temporal Summarization Track

Yuanyuan Qi , Qinlong Wang, Chuchu Huang, Bo Tang,
Weiran Xu,Guang Chen,Jun Guo
School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications
Beijing, P .R. China, 100876
qiyuanyuan@bupt.edu.cn

Abstract:

This paper describes the information extraction systems of BUPT_PRIS at Temporal Summarization Track, Which includes data obtaining and preprocessing, index building and query expansion, sentences scoring module. This year only keep one task: sequential update summarization, the task: value tracking is cancelled. For the sequential update summarization we focus attention on queries expansion and sentence scoring. There are three methods of query expansion introduced in this report: WordNets, Word representation, spatial analysis method. We also show the evaluation results for our team and the comparison with the best and median evaluations

1. Introduction

The Temporal Summarization track was introduced last year, this year TREC Organizer only keeps one task: sequential update summarization. The goal of the Temporal Summarization track is to develop systems that allow users to efficiently monitor the information associated with an event over time. We build an information extraction system to complete the task.

2. Dataset and preprocessing

1) The dataset is provided by TREC 2014 Temporal Summarization track. But this year track offers two datasets one after another. Two datasets are KBA Stream Corpus 2014 and TREC-TS-2014 corpus, in fact we use the TREC-TS-2014 corpus which one is a specially filtered subset of the full 2014 Stream Corpus for use in the Temporal Summarization (TREC-TS) track and 41.36% of data is English corpus.

   This year the office offers 10 training queries and 15 testing queries. All these queries can be sorted in two kinds:

   - Natural disasters such as earthquakes, hurricanes, cold wave, etc.
   - Man-made accidents such as shooting, explosion, parades, etc.

2) Due to the change of dataset ,we choose the second dataset and followed the blow steps:

   - Decryption and Uncompressing. KBA organizers for the convenience of multiplatform multi-language operation, the use of the original document of Apache Thrift framework for data serialization, then to xz serialized binary data compression, GPG encryption, eventually to synthesize into a GPG file.

- Deserialization, we extract stream_id、clean_visible、token、offset、POS NER and co-reference resolution information from the data of thrift serialization

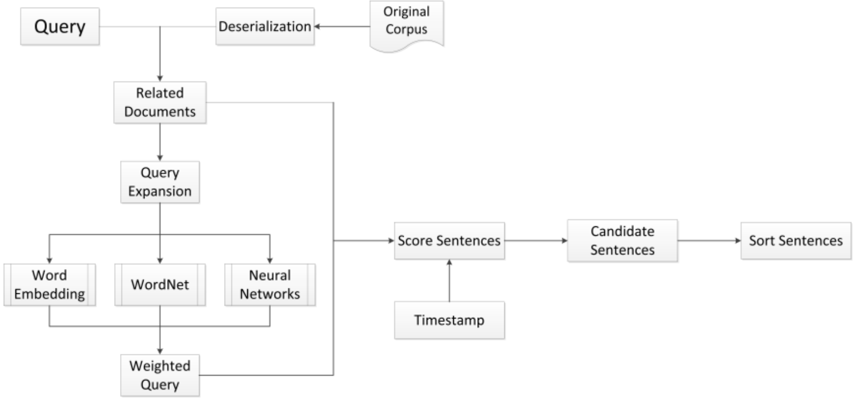
This time we don’t build Index for the time isn’t enough. But at the time of preliminary search relevant documents, to imitate the thinking of Boolean query, the document must be present query and within the scope of the prescribed time, it was about 100 g related documents. These include text content, time, information Token information, id information.

3. Sequential Update Summarization

This task also focuses attentions on keywords mining and sentence scoring as last year. The framework of Sequential Update Summarization is illustrated in Figure 1

3.1 System overview

After preprocessing, we expand official queries through the query expansion in different angles. Through the relevant search, we get related document collections, and according to the sentence scoring formulas to score sort related documents. After getting high confidence sentences, we then do some post-processing on sentences. The system overview is shown in Fig.1.



3.2 Query expansion Module

The query expansion module uses three methods to enrich the official queries and allow to find more relative documents after retrieving.

1. WordNet
   In WordNet[1] , for each entry can have part of speech of each word description and a distinguished from other distinguishing meanings of words are called ‘lemma’. We extract as a noun or a verb meanings of the lemma as candidate query words for each word in the existing query. We filter candidate word frequency is lower than 3 word query in the final, the rest of words treat as extensions .
2. Word Representations[2]
   We use Google's open source tools Word2Vec to train word vector, and use cosine distance metric to measure alternative query word and the distance from the query

words, choose the nearest TOP N as extensions, and add the distance as the weight information of scoring formulas.

3. Spatial analysis method

   In learning text semantic characteristics, recently, more and more people focus on neural network learning model, more and more people are remarkable achievements in the neural network model, such as the above Mikolov's word representations, the spatial analysis method is a combination of word representations and linear discriminator. Through linear discriminator to analyze and select dimension feature of word representation.

3.3 Sentences Scoring Module

We utilize a keyword shooting method to evaluate sentences. The keyword shooting method is described as following[3][4]:

$$\mathrm{Score}(s_i) = \frac{K}{\|s_i\|} \times \left(1 - \frac{2}{\pi}\tan^{-1}\left(\frac{u_t - n_t}{\alpha}\right)\right) \qquad , \alpha = 3600 * 6 \qquad (1)$$

Where K is the number of keywords that found in the sentence s;

α is latency-step (6 hours);

$\mathrm{n_t}$ is nugget time (time of Wikipedia edit from which nugget was extracted)

$\mathrm{u_t}$ is update time (in words)

3.4 Post-processing

1. Timely sentence ranking

This step ranks all sentences by decision-making time, this step makes sure that no repetition of stream_id

2. Different threshold conditions

Due to the different methods of query expansion and keyword shooting method, different threshold conditions of sentence length

4. Experiment Result

Table 1 shows the retrieval performance of our submitted four runs for Sequential Update Summarization task. The primary evaluation metrics for this year's vital filtering are Expected Gain (nEG), comprehensiveness (and latency-comprehensiveness) of the system, E(latency) and H.

We can see from the table that BUPT_PRIS_Cluster 1 have better performance of C(S) than others and also E(Latency). But BUPT_PRIS_Cluster 4 have better rperformance of nEG(S) and H**.**

Table 1 The performance of submitted runs for TS

Run ID	**nEG(S)**	**C(S)**	**E[Latency]**	**H**
BUPT_PRIS_Cluster 1	0.0033	**0.4369**	**0.0059**	0.0127
BUPT_PRIS_Cluster 2	0.0059	0.3728	0.0101	0.0222
BUPT_PRIS_Cluster 3	0.0115	0.338	0.0208	0.0407
BUPT_PRIS_Cluster 4	**0.0155**	0.2692	0.0314	**0.0508**

5. Conclusion

In this paper, we present our systems for TREC 2014 Temporal Summarization Track. In the Sequential Update Summarization task, we apply three methods to enrich the official queries and allow finding more relative documents after retrieving.

We don’t have enough server an time to build Index this year, this leads to losing of important features of scoring formula.

6. Acknowledgments

The work reported in this paper was supported by 111 Project of China under Grant No. B08004, key project of ministry of science and technology of China under Grant No. 2011ZX03002-005-01, National Natural Science Foundation of China (61273217) and the Ph.D. Programs Foundation of Ministry of Education of China (20130005110004)

Reference

[1] http://wordnet.princeton.edu/wordnet/

[2] Tomas Mikolov; Kai Chen ;Grep Corrado; Jeffery Dean (September 2013).”Efficient Estimation of Word Representations in Vector Space”. International Conference on Learning Representations 2013

[3] University of Glasgow at TREC 2013: “Experiments with Terrier in Contextual Suggestion, Temporal Summarisation and Web Tracks”

[4] Chunyun Zhang;Weiyan Xu ;”The Information Extraction systems of PRIS at Temporal Summarization Track”