The Information Extraction systems of PRIS at Temporal Summarization Track

Chunyun Zhang, Weiyan Xu, Fanyu Meng, Hongyan Li, Tong Wu, Lixin Xu

Abstract:

This paper describes the information extraction systems of PRIS at Temporal Summarization Track. The Temporal Summarization Track includes two tasks: sequential update summarization and value tracking. For the first task, we focus attention on keywords mining and sentence scoring. The system utilizes hierarchical Latent Dirichlet Allocation (LDA) to do keywords mining and score sentences with keywords shooting method. For the second task, we define the value extracting as a sequence labeling problem and build a discriminative undirected graph model (CRF model) to extract attribute values of all topics.

1. Introduction

This is the first year of the Temporal Summarization track. The track includes two tasks: sequential update summarization and value tracking. The goal of first task is broadcasting useful, new, and timely sentence-length updates about a developing event. The goal of value tracking is tracking the value of important event-related attributes (e.g. number of fatalities, financial impact).

In order to complete the two tasks, we build an information extraction system for each task. For the sequential update summarization task, the system focuses attention on keyword mining and sentence scoring. For the value tracking task, the system defines the task as a sequence labeling problem and tries to learn a discriminative undirected graph model—CRF model to solve the information extraction problem.

2. Sequential Update Summarization task

The sequential update summarization system focuses attention on keywords mining and sentence scoring. The framework of Sequential Update Summarization system is illustrated in Figure 1.

2.1 Preprocessing module

The preprocessing mainly refers to decryption, uncompressing and deserialization process to the initial data. After preprocessing, we use the open source indexing tool--Elasticsearch to build index. Due to the big scale of initial data, we just take the content of token label as information from initial data to build index.

2.2 Keywords Mining Module

The keywords mining module uses hierarchical Latent Dirichlet Allocation [1] (LDA) to find potential topics and returns the most representative words of each topic as keywords.

We gather all supporting documents of events as training data. we firstly use the LDA toolkit to discover two topics and choose the most representative words for each topic;

secondly, discover 5 new topics by the same method under the topic discovered in the last step and choose the most representative words of each topic; lastly, choose keywords manually in the two level representative words of each topic.

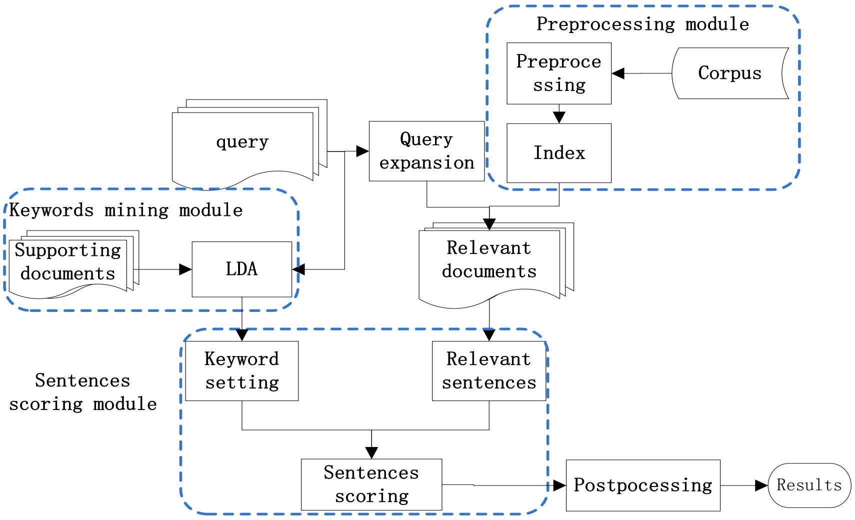

Figure 1. The framework of sequential update summarization system

2.3 Sentences Scoring Module

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

$$Score(S_i^k) = \frac{|V_{keyword}^k \cap S_i^k|}{|S_i^k|} \tag{1}$$

Where $V_{keyword}^k$ is the keyword vector of the topic *k*. S_i^k is the *i*th related sentences of topic *k*.

2.4 Postprocessing and Results

After getting high confidence sentences, we then do some postprocessing on sentences.

1. Timely sentences filtering
   This step filters out outdated information in each hour. For all sentences, detect the time information to make sure these sentences are in current hour.
2. Duplicate removal
   For sentences in current hour, the postprocessing module first finds same sentences with different sentences id, then compares the stream id of all sentences and choose the one with the earliest time information as the submission sentence.
3. Redundancy removal
   Based the idea that a update is a timestamped short text string comparable in length to a sentence, the system reserves sentences with the length shorter than 20.

3. value tracking

We define the value tracking task as a sequence labeling problem. Then the value tracking system learns a discriminative undirected graph model—CRF model to do value extraction. The framework of value tracking task is illustrated in Figure 2. The preprocessing module and postprocessing module are same to task 1. So we do not describe these two modules in the following sections.

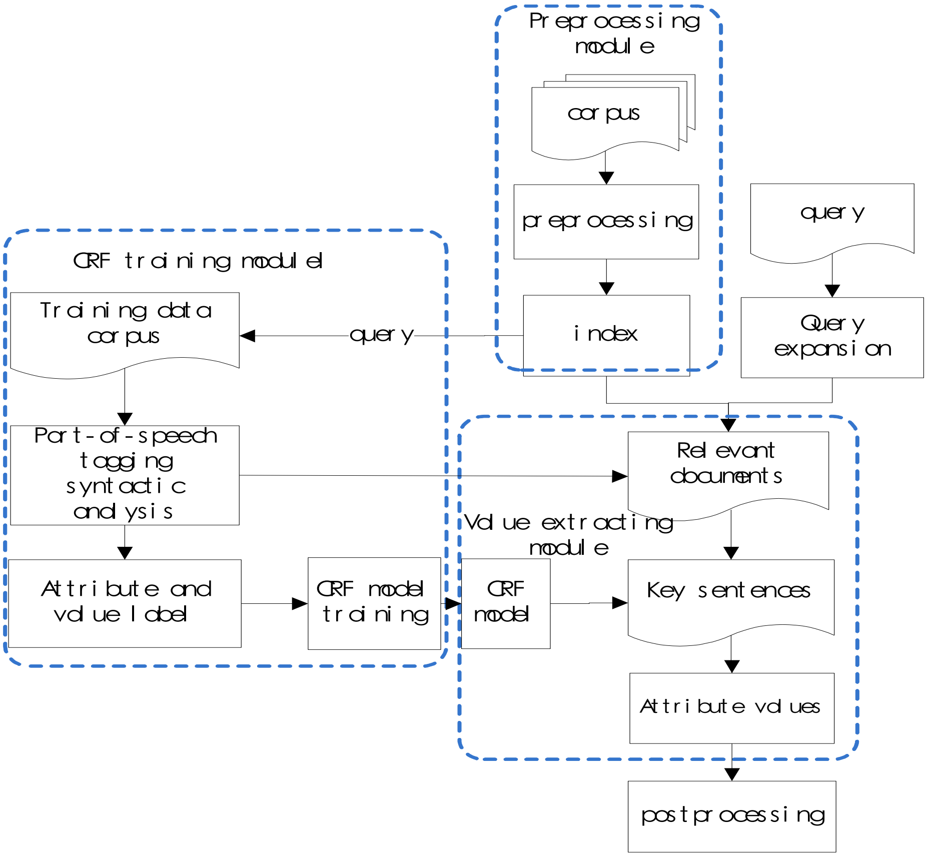

Figure 2. The framework of value tracking task system

3.1 CRF training module

The goal of this module is training a CRF model [2] for attribute value labeling. The training data is the event of 2012 East Azerbaijan earthquakes which is issued by the organizer. We first obtain relevant documents by searching in the initial data and then do syntactic and dependence tree analysis on relevant documents, lastly, for sentences including attribute values, we label attribute-values manually with labels of FV, LV, DPV, DV, IV corresponding to financial impact value, location value, displaced value, death value and injury value. Lastly, train the CRF model with training data to form the best parameter model which will be used in the value tracking module.

Specially, we use the dependency features besides lexical and part-of-speech features. The dependency features are the root node and the dependency with the root node in dependency tree.

3.2 Value extracting module

In this module, we firstly retrieve relevant documents of all events and do syntactic and dependence tree analysis on relevant documents; secondly, choose sentences including keywords of each attribute as key sentences. Thirdly, apply the trained CRF model on the key sentences and label all attribute values with corresponding labels. Then we can obtain the initial attribute value sets with temporal information.

Reference

[1] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). "Latent Dirichlet allocation". In Lafferty, John. Journal of Machine Learning Research 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

[2] Lafferty, J., McCallum, A., Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. pp. 282–289.

[3] Settles, B. (2004). "Biomedical named entity recognition using conditional random fields and rich feature sets". Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. pp. 104—107.