International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-2, Issue-3, March 2016 Pages 37-39

Research on News Video Multi-topic Extraction and Summarization

Di Li , Hua Huo

***Abstract*—Comparing to other videos, news video contains more important information. Processing a news video with semantic extraction, abstract and annotate its sub-theme can facilitate our operations such as reading and retrieval, make the news video more easy to management and application. In this paper, we use news video text area positioning and extraction technology to obtain text corpus, put these sentences in Chinese word segmentation and syntactic analysis system, then computation sentence weighting and ranking them, calculate the similarity with the sentences. Finally, conduct a study about multi-topic extraction and summarization. The experimental result show this method with high pertinence which can better achieve the targets.**

I. INTRODUCTION

In the information age, news video is one of the important ways we get information. Added semantic[1] generalization to the news video playback time axis on each child news theme can make us more convenient know the news content, rapid positioning point of interest, is advantageous to the query and management. Because the news video is a special kind of video[2], it's information such as text, subtitle has a high degree of generality and can be used as important auxiliary when we explore the video semantic. Therefore, we first to extract of news video text[3], summarize a news video as a text-involved corpus, then analysis it combination with video time series[4]. Analysis of sentence similarity, and the relationship between the text and classification for each subject of news video, then summarize the child theme[5].

Currently widely used text summarization technology generally have the following kinds:

1. Method based on test sets, this kind of method based on machine learning, learning from the text and the relevant rules. The shortcoming of this method is to use to generate the training of the generated rules, rely too much on training text set, our Chinese news video project's episode is very concise, so effect is not very good.
2. Method based on statistical model[6], this method is the word frequency, title, location, syntactic structure and the Key words according to the importance of elements such as the judge sentence to extract important sentences as a summary. But lacks the analysis of the text theme and content.
3. Method based on domain ontology[7], this method through ontology for semantic information in the level of application field and internal relationship ontology to text analysis, filter out irrelevant information, which make the theme more prominent. But due to the lack of Chinese domain ontology, we need to build a domain ontology, workload is big, and does not apply to our video news corpus.
4. Method based on text diagram[8], this approach in the paragraph of text relationship between figure, is summarization method based on the passage, it concludes a paragraph of words and other paragraphs consistency of vocabulary, in order to query the paragraph topic covered, this algorithm does not apply to our short video news corpus, and the time complexity is higher.

In order to improve all kinds of faults of these method, and combined with our research topic, we put forward the combination of time series news video summarization system.

II. SEMANTIC ANALYSIS AND MULTI-TOPIC SUMMARIZATION

A. The Preparation of Text Corpus

In order to extract the caption of a news video, we use the method of news video text area positioning and extraction based on coner detection algorithm[9]. Because most of the text of news video can describe its semantic information, and according to the characteristics of text in news video, we use an improved Trajkovic corner detector to detect the corner related to text. The adaptive thresholding method of scale in combination with standard deviation (SCSD) can be applied to determine the corners, which aims to accurately obtain the corners. The breadth-first clustering algorithm was also been used to distinguish and plan the detected corners which are in same range to determine the text area in news video frame. We apply this method to get the text area of the news video and marked text area as ROI (Region of Interest), then OCR (Optical Character Recognition) technology is utilized to extract text.

Di Li, Information Engineering College, Henan University of Science and Technology, Luoyang, Henan, China;327017465@qq.com

Hua Huo, Information Engineering College, School of Software,Henan University of Science and Technology,Luoyang, Henan, China;

B. The Text Pretreatment and Syntactic Analysis

Sentence: 英国观察家发表积极评论

（British observers comment positively）

After POS: 英国/ns 观察家/n 发表/v 积极/ad 评论/v

Figure1. Example of Chinese Word Segmentation

The text segment we extracted can be thought of a simple summary of news video. As a news video contains multiple news events, each news event also contains multiple description, we need to calculate each news event and all of the fields, draw the summary of news events, combine time series with the shot boundary as the cut-off point to semantic summarizing and labeling.

The smallest semantic unit of Chinese is word[10], so, to deal with it, the first priority is to find the smallest semantic unit. In this paper, we use ICTCLAS 2015 system to word segmentation and part-of-speech tagging for a sentence. As shown in Figure 1. Is the sample of word segmentation result.

In the processing of Chinese information interdependence syntactic analysis study, Chinese scholars put forward five axioms of dependency:

1. There is only one component in a sentence is independent;
2. Other ingredients directly depends on a particular component;
3. Any component could not dependent with two or more than two components;
4. If A directly dependent on B, and C in the sentence is located between A and B, then C or directly dependent on B, or directly dependent on one between A and B;
5. Two sides of the Center composition did not concern about other ingredients.

Dependency grammar analyzing the relation between the language units in composition to reveals the interdependence of the syntactic structure. It advocating core verb in the sentence is the center of the composition, control other components and it itself is not subject to any other ingredients, all controlled components in some dependencies belong to the dominant.

C. Multi-topic Summarization

The news video corpus we studied not only have more types of news video, such as politics, finance, science and technology, sports, etc. And each type of news video may also contain more sub-theme.

If only the use of statistical learning methods, according to the weight factor to extract the summary, may cause the loss of part of the child theme. Therefore, in this paper, we will use the statistical method and text diagram method, and according to the differences of the timeline feature to make the sub-theme summary.

We representation of a sentence in vector, news video text after the word segmentation can filter out some words which less relation to subject, such as Prepositions, function words, numerals, etc, in addition to the influence of useless information. We only on verbs, nouns, and other important parts of sentences, switch sentences into key words in the form of feature vector.

Assuming that video 'A' corpus consists of the sentence' $Q_1, Q_2, \ldots, Q_n$ ', the sentence Q_i is made up of a set of keywords ' $N_k = \{N_{k1}, N_{k2}, \ldots, N_{kn}\}$ ', the key words set of a news video text corpus can be expressed in The Equation 1.

$$N = \bigcup_{i=1}^{m} N_i = \{N_1, N_2, \ldots, N_m\} = \{W_1, W_2, \ldots, W_n\} \quad (1)$$

We defined the weight of key words in the text corpus as The Equation 2.

$$D_i = T_i \times \log\left(\frac{n_i}{m} + 1\right) \quad (2)$$

The ' T_i 'is the word frequency of the key word ' W_i ' in the text corpus, 'M' is the amount of sentences in the text, and ' n_i 'is keywords frequency, so ' n_i / m' expressions the keywords distribution.

The vector representation of sentence' Q_i 'is" $(QD_{i1}, QD_{i2}, \ldots, QD_{ij}, \ldots, QD_{in})$", the key word W_j'weight express as QD_{ij} .

As the Equation 3 shown, so the weight of sentence' Q_i 'is ' $WQ_i = QD_{i1} + QD_{i2} + \ldots + QD_{im}$'.

$$QD_{ij} = T_{ij} \times D_j \times \frac{n_j}{m} \quad (3)$$

The next step is sentence similarity calculation, we adopt the method based on Euclidean distance to calculate the similarity between each sentence of corpus. Assuming sentence $Q_i, Q_j \in S$, as the Equation 4 shown, if $dist(Q_i, Q_j) < \lambda$, Q_i, Q_j are similar. According to our experiment many times, in the aspect of news video text corpus, we can achieve good effect when $\lambda = 0.63$.

$$dist(Q_i, Q_j) = \sqrt{|QD_{i1} - QD_{j1}|^2 + |QD_{i2} - QD_{j2}|^2 + \ldots + |QD_{in} - QD_{jn}|^2} \quad (4)$$

Through the similarity calculation of the entire news video corpus, we can according to the results to division sub-theme about each news event, higher similarity of sentences could be divided into a cluster, corresponding to the full text is divided into several sub theme. In the interior of the sub-theme, order was based on the weight of the sentence, the important node to the first place, the importance of sentences can be calculated according to the weight of it.

III. Experimental Analysis

News video sub-theme summarization didn't have a general evaluation method, but due to the semantic structure of the news video summarization is clear, it does not need more text corpus. We can treat a news video corpus as multi-topic text, then extract the sub-theme. We can evaluate the automatically-extracted result with actual sub-themes and the time sequence. Therefore, we selected 6 different categories of news video from the corpus, such as politics, finance, technology, society, military, sports, etc. Then compare the result.

We evaluate result through two aspects below:

1. Use precision and recall rate and F measure to evaluate the quality of sub-theme, the higher the three value, the better the results.
2. Using the subject extraction rate to evaluate the sub-theme extraction effect, subject extraction rate is the ratio of experimental result and practical theme.

MS Word automatic abstract method were compared with our method, through selected a number of different categories of representative news video to evaluate experimental result, the compression ratio is 0.2 and experimental result is shown as table1.and table2.

TABLE I. ACCURACY ASSESSMENT

Video category	Our method			MS Word		
	P	***R***	***F***	***P***	***R***	***F***
politics	0.836	0.792	0.813	0.512	0.701	0.653
finance	0.783	0.721	0.751	0.677	0.632	0.654
technology	0.811	0.823	0.817	0.709	0.679	0.694
society	0.715	0.779	0.746	0.683	0.641	0.661
military	0.781	0.692	0.734	0.632	0.597	0.614
sports	0.687	0.752	0.718	0.614	0.645	0.629

TABLE II. RATE OF SUBJECT EXTRACTION

Video category	Our method			MS Word		
	Total	***Result***	***Rate***	***Total***	***Result***	***Rate***
politics	4	4	1.00	4	3	0.75
finance	5	4	0.80	5	3	0.60
technology	3	3	1.00	3	2	0.68
society	6	5	0.83	6	4	0.68
military	4	4	1.00	4	2	0.50
sports	5	4	0.80	5	3	0.60

IV. CONCLUSION

Through a review of the experimental results, our method is better than MS-Word summarization method. The quality in summarization is much better, and more accurate when aim to the news video text. In the section of extracting sub-theme, due to the combination of the information such as video shot and timeline, our method can complete almost all sub-theme extraction, and news video multi-topic summarization.

ACKNOWLEDGMENT

This study was supported by the National Natural Science Foundation of China under the grant 60743008, the Key Scientific and Technological Projects of Henan Province in China under the grant 142102210045 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

REFERENCES

[1]. I. iaqi Fu, H. Hu, R. Chen, and H. Ren, Video Semantic Analysis[J]. Information Technology Journal, 11(10) 2012. p. 1381-1390.
[2]. W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang. "Integrating visual, audio and text analysis for news video." in Image Processing, 2000. Proceedings. 2000 International Conference on. 2000. IEEE.
[3]. Q. Ye and D. Doermann, Text detection and recognition in imagery: A survey[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 37(7) 2015. p. 1480-1500.
[4]. Y. Nakamura and T. Kanade. "Semantic analysis for video contents extraction—spotting by association in news video." in Proceedings of the fifth ACM international conference on Multimedia. 1997. ACM.
[5]. Q. Mei and C. Zhai. "Discovering evolutionary theme patterns from text: an exploration of temporal text mining." in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005. ACM.
[6]. G. Erkan and D. R. Radev, LexRank: Graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004. p. 457-479.
[7]. C.-S. Lee, Y.-J. Chen, and Z.-W. Jian, Ontology-based fuzzy event extraction agent for Chinese e-news summarization[J]. Expert Systems with Applications, 25(3) 2003. p. 431-447.
[8]. G. Salton, A. Singhal, M. Mitra, and C. Buckley, Automatic text structuring and summarization[J]. Information Processing & Management, 33(2) 1997. p. 193-207.
[9]. M. Trajković and M. Hedley, Fast corner detection[J]. Image and vision computing, 16(2) 1998. p. 75-87.
[10]. N. Xue, Chinese word segmentation as character tagging[J]. Computational Linguistics and Chinese Language Processing, 8(1) 2003. p. 29-48.

Di Li, he is a postgraduate student of Information Engineering College, Henan University of Science and Technology. His research interests in intelligent information processing and image processing..

Hua Huo, he is a professor with a Ph.D. in Information Engineering College and School of Software, Henan University of Science and Technology. His research interests in intelligent information processing and video semantic extraction.