Study on Multi-document Summarization Based on Text Segmentation

Meng Wang
Assoc. Prof., Computer College, Guangxi University of Technology, CHINA
Email:mwang007@163.com

Xinlai Tang
Assoc. Prof., Lushan College, Guangxi University of Technology, CHINA
Email: lz_txl@163.com

Xiaorong Wang
Assoc. Prof., Computer College, Guangxi University of Technology, CHINA
Email:ccnuxnxs@163.com

***Abstract*—This paper introduces a novel approach of automatic multi-document summarization based on text segmentation. The approach acquires concepts with a How-Net oriented tool and calculates the importance degree of sentences by means of employing the improved DotPlotting model and establishing a sentence-based vector space model (VSM). A conclusion is made according to the importance degree and similarity of sentences. Experimental results show that the performance on ROUGE of the approach put forward hereby is effective and significant.**

***Index Terms*—Text segmentation, summarization, How-Net**

I. Introduction

Today electric text messages of various kinds have emerged in great numbers, with pages available on the Internet almost doubled every year in numbers [1]. For instance, in January 2012, the number of hosts advertised in DNS is 888,239,420 [2]. To help people cope with the ever-increasing text documents, advanced technologies facilitating text summarization have been developed. Automatic text summarization aims to automatically produce a short and well-organized summary of a single or multiple documents. As a fundamental and effective tool for document understanding and organization, the multi-document summarization enables better information services by creating concise and informative reports for a large collection of documents [3,4].

Recently, more and more models have successfully been applied to summarization. Jean-Yves Delort adopts an unsupervised probabilistic approach to model the novelty in a document collection and applies it to the generation of update summaries [5]. James Gung uses temporal information to improve abstracted summarization [6]. Kristian Woodsend adopts a method where such individual aspects are learned separately from data but optimized jointly by employing an integer linear program to abstract summary [7]. Seonggi Ryang presents a new approach to automatic text summarization based on Reinforcement Learning, which models the construction of a summary within the framework of reinforcement learning and attempts to optimize the given score function with the given feature representation of a summary [8].

Till now, there are two types of summarizations, one of which is the abstractive summarization and the other is the extractive one. Extractive summary consists of selecting important sentences and paragraphs etc. from the original document and combines them into a shorter form based on statistical and linguistic features of sentences. The abstractive summarization aims to represent main concepts and ideas of a document by paraphrasing the source document in a clear natural language. Most of the recent works have concentrated on the extraction summarization method where there are two main techniques for feature extract, namely, sentence-based and keyword-based text summarization [9]. The former identifies the most salient sentences in a document while the latter summarizes documents by topics. Each of the approaches is featured in a set of keywords.

In this paper, a special Chinese automatic summarization method is proposed on the basis of Text segmentation. The method consists of three main parts: 1) feature finding: using concepts as minimal semantic unit rather than words, and using HowNet as a tool to obtain concepts in the text. 2) Text segmentation: using an improved DotPlotting method to segment texts. The method not only gives consideration to the defects of traditional DotPlotting, but also improves the speed of text segmentation by using a concept matrix. 3) Automatic summary. According to the segmentation results, the system can obtain the summary of the text on the basis of similarity. Summarization evaluation metric ROUGE motivated by the MT evaluation metric is used. Experimental results indicate clear superiority of the proposed method over the traditional ones in the proposed evaluation scheme.

The rest of this paper is organized as follows: Section 2 describes how to obtain concepts by using HowNet;

doi:10.4304/jcp.9.5.1241-1246

Section 3 demonstrates text segmentation based on the improved DotPlotting method; Section 4 introduces the way to abstract summary; Section 5 presents some experiments and their numerical results.

II. Concept-Obtained

A. Introduction to HowNet

How-Net is a knowledge database which has been released recently on the Internet [2]. In How-Net, the concepts expressed in Chinese or English are described and the relations between concepts and the attributes of concepts are revealed. This knowledge database is used as the resource of evaluating the sememe, for it can offer some useful information The format (which is defined as HowNet tool) can be described as follows:

NO. = serial number
W_X = word
G_X = part of speech
E_X = example of word
DEF = definition of word

Examples of lemma in HowNet can be represented as follows:

NO.=005987
W_X= blow up
E_X= the plane will blow up， Boat blowed up
G_X=V
DEF={FormChange| shape change:StateFin={Out Of Order| shatter}}

A lemma in HowNet presents part of speech and definition of a word. In the definition of words, the basic sememe ({FormChange|形变(shape change)}) and the related sememe ({StateFin={OutOfOrder|坏掉(shatter)}) are defined respectively. The former reflects the meanings of a word, while the latter represent the frame feature about a word. Both of the sememes can help to obtain the word concept in the text.

B. Concept Acquisition Based On HowNet

From the structure of How-Net, one learns that the DEF item expresses the meaning of words very well. Words with the same DEF item are regarded as sememes for they share the same word meaning. Words composed of a set of single elements are different words with the same concept. Two problems are processed when the actual concepts are being acquired. The first problem is: the obtained principle is the same DEF item in the process of obtaining a word concept for polysemous words. The actual sememe item of polysemous word cannot be distinguished, which then influences the accuracy of concept acquisition. The second problem is: the distinction of DEF is too strict, thus some related information will be probably missed if alignment-search depends entirely on DEF item.

We can firstly solve the selection problem of DEF item of polysemous words; the word concept is obtained by using improved DEF item. We use the ICTCLAS platform of ICT (Institute of Computing Technology, Chinese Academy of Sciences) to conduct words segmentation and part-of-speech tagging for the document. Some words will be deleted, such as prepositions, numerals and function words which have little influence on text summarization. Some key words will be extracted, such as nouns and adjective. The text with segmented and part-of-speech tagging will be obtained. There are two cases for selecting DEF item of polysemous words. One is that the part-of-speech of some polysemous words is varied in different contexts. The DEF items of these polysemous words can be determined by tagged part-of-speech. Another case is that the same part-of-speech in different DEF items for polysemous words, but different part-of-speech of words will pair up different words in different contexts. Just taking two words (NO. is 005987 and NO. is 005990) in HowNet for instance. The probable meaning of the first word usually adopts a grammar form of N+V, while the probable meaning of the second word usually adopts a grammar form of V+N. Different contexts will have different grammatical forms, and therefore DEF item of this type of polysemy can be determined in this view.

The detailed process is as follows:

1) The DEF item is redefined. The DEF item is extended to the union of contained basic sememe and relation sememe of this word. If the meaning of abstract sememe in How-Net is too large and broad, the abstract sememe will be filtered, such as “attribute”, “event” and “entity”.

2) The document model before concept acquisition is established by sentences. The document model is expressed as $S_j(W_1,W_2,\cdots W_n)$(the document contains the j sentence, with each sentence containing n words).

3) We scan the sentence where the vector space model is established. We assume that the scanning sentence is currently the jth sentence.

4) We scan the word W_i of the sentence, and find the corresponding DEF item of the word. At the same time we scan the sentence to search whether some words have the same meaning as the sememe of the DEF item. If the search result is a negative one, we will tag the concept of the word W_i, and scan the next word W_i +1 of the sentence, and proceed to Step (4). When all the words of the sentence are scanned, we will scan the next sentence and proceed to Step (3). If the search result is a positive one, we will proceed Step (5).

5) The word W_k is extracted and then the corresponding DEF item of the word W_k is found. If the DEF item sememe word of the word W_k does not contain the word W_i, the word W_i’s concept and the word W_k’s concept will be tagged with the word W_i’s DEF item. If the DEF item sememe word of the word W_k contains the word W_i, we will compare the sememe distance of the two words in the DEF item. The DEF item of the word which is closer to the basic sememe will be selected as the concept of the two words. Then we scan the next word W_i+1, and proceed to Step (4). When all the words of the sentence are scanned, we will scan the next sentence, and proceed to Step (3).

When all the steps are completed, concepts of all words are contained. The word concept contained by the above method solves the digestion problem of the polysemous words. Meanwhile, the words which have the same relationship in the same context are treated as a concept. This can ensure the orthogonal relation of each conceptual element in concept vector space model based on concept, and help generate high quality text summarization.

This paper selects the “H7N9 Bird Flu” as the topic， and downloads 100 documents from http://news.qq.com/zt2013/H7N9/ as the test corpus. Statistics concepts and words use concept-based method and word-based method respectively. The results show that compared with word frequency statistics method, the number of concept has reduced greatly by concept statistical algorithms. Only consider document collection frequency greater than 2 times words and concepts, the document include 2,869 words and 1,789 concepts. Moreover, 1,789 concepts include 3,345 words. This illustrates that more of the words could be included in less of the concept when the concept of statistical method is used. This cannot miss the word frequency statistics appearing in the small amount of articles and express is an important word concepts.

Figure 1 shows the comparison between the numbers of words and concepts using word vector space model and concept vector space model under three different themes. The number of concepts is significantly less than that of words.

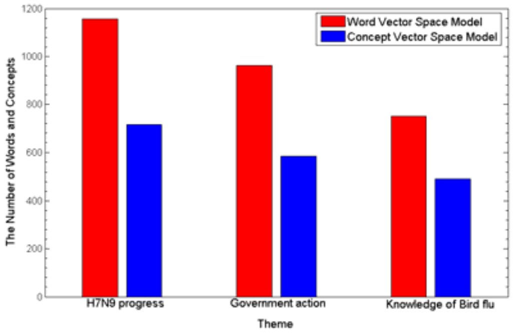

Figure 1. Number of words and concepts

III. Text Segmentation Method Based on Concept Density

A. The Traditional Doptlotting Model

The DotPlotting[10] is a famous model in the field of text segmentation. This method is based on the vocabulary degree of polymerization and the image analysis technology. It identifies the semantic paragraph border by point diagrams reflecting overall distribution of document vocabularies. If one word appears at x and y position repeatedly in the document, the word will be marked by a point in four coordinate points (x, x),(x, y),(y, x),(y, y) in the point diagram. Then, all vocabularies of the document will be marked in the point diagram. A symmetrical two-dimensional point diagram will be formed based on this method. The two-dimensional point diagram can clearly reflect sub-topic distribution of one document and measure theme consistency by establishing density evaluation function.

The density function of Dotplotting is expressed as:

$$f_D = \sum_{j=2}^{|P|} \frac{V_{P_{j-1},P_j} \cdot V_{P_j,n}}{(P_j - P_{j-1})(n - P_j)} \qquad (1)$$

Here n represents the length of the document. P_j represents the position of the jth semantic paragraph boundary. |P| represents the semantic paragraph number of the document. $V_{x,y}$ represents the word frequency vector of text fragment from the x^{th} word to the y^{th} word.

In the traditional DotPlotting model, if we only use the vocabulary as a basic semantic unit, the two-dimensional point diagram will have a lot of coefficient matrix, which will be unable to accurately abstract the border of some semantic paragraphs when density function evaluation is adopted. In the formula (1), every single density is $\frac{V_{P_{j-1},P_j} \cdot V_{P_j,n}}{(P_j - P_{j-i})(n - P_j)}$ .The density of each position P_j is to calculate the vocabulary similarity from its previous semantic paragraph to its back in all texts. So the density of each position P_j is determined based on its previous semantic paragraph border and the end position n of the document. That causes an asymmetry density function, resulting in completely different text segmentations between scanning the document from front to back and from back to front. Sincewe do text segmentation by evaluating density function in a symmetric two-dimensional point diagram, we must solve the problem of density function asymmetry for the traditional DotPlotting model.

B. Improved Dotplotting Model

From the analysis and research for the traditional DotPlotting model, this paper will use concepts of the second part instead of words to create symmetric two-dimensional point diagram. At the same time, density function is improved to solve the problem of density function asymmetry for the traditional DotPlotting model.

$$f_D^{'} = \sum_{j=2}^{|P|} \frac{V_{P_{j-1},P_k} \cdot V_{P_j,n}}{(P_j - P_{j-1})(n - P_j)} + \sum_{j=1}^{|P-1|} \frac{V_{0,P_j} \cdot V_{p_j,P_{J+1}}}{P_j(P_{J+1} - P_j)} \qquad (2)$$

The second part of formula (2) is “Backward” density which aims to solve the density function of symmetry. By modifying the formula (1), the “Backward” density of P_j is determined by next semantic paragraph boundary P_j+1 and the start position 0 of the document. This density function can get the same density function value whenever the document is scanned from front to back or from back to front.

C. Text Segmentation Algorithm

The semantic paragraph boundary determination method of the DotPlotting model is: If B is the established semantic boundary set, the remaining boundaries are candidate semantic boundaries; the remaining boundary set is the candidate boundary of the next round which is composed of the candidate boundary set C. For each candidate boundary i of C, P=B∪ {i}, we calculate the overall density by P division recording to the formula (2). We will select the overall density of the smallest candidate boundary as the next best semantic paragraph boundary, and combine it with the set B. The specific description of the algorithm is as follows:

(1) For a given document W, we have to pretreat it. We acquire word concept according to the concept acquisition method of the second part, establish a two-dimensional point concept diagram, and determine the semantic paragraph partition number K.

(2) Initialize the semantic boundary set B as an empty set; each paragraph is a boundary which is seen as a candidate segmentation point. We establish a candidate boundary set C based on the candidate segmentation point and we use S to record the best segmentation variable.

(3) We repeat operations (4)-(5) from segmentation paragraph 1 to segmentation paragraph k.

(4) For each boundary candidate point i of the set C, P=B ∪ {i}, we calculate the overall density d by P division according to the formula (2). If dmin is greater than d, dmin =d. We will record S=i.

(5)The boundary S will be a target boundary added to set B. At the same time, S will be deleted from the candidate boundary set C.

Semantic paragraph boundaries are successively added in this algorithm. The end of natural paragraph in the document is set as candidate semantic paragraph segmentation point. We check each candidate boundary when selecting new semantic paragraph boundaries. We try to add each candidate boundary to the boundary set B and form the new boundary set P. We evaluate segmented mode composed of the boundaries from the new boundary set by density function. The candidate boundary which has the minimum value of density function is selected as a segmentation boundary and it is added to segmentation boundary set until the number of boundary is equal to K.

IV. Automatic Summary based on Text Segmentation

For those original documents, the system should exclude those useless words, such as prepositions, empty words and numerals etc during pretreatment, and only some important nouns and adjectives are treated. In this section, the proposed method will be introduced in details. The process of abstracting Summary by text segmentation is displayed in Figure 1.

(1) Calculate Importance of Concepts

We apply TF*IDF to assign weight to the individual concept and the importance of each concept is defined as follows:

$$Wd_{it} = TFd_{it} * \log \frac{N}{Nd_t} \qquad (3)$$

Wd_{it} in formula (3) is TF*IDF of concept t in the i-th document. TFd_{it} in formula (3) denotes concept frequency of t in the i-th document. N is the number of documents and Nd_t is the number of documents where t occurs.

(2) Calculate Importance of Sentences

After the CVSM S_j (C_1，W_{1j}; C_2，W_{2j}; C_n, W_{nj}) of all sentences in the text are established, the importance of each sentence is defined as follow:

$$W(S_j) = \lambda \frac{\sum_{i=1}^{n} F_{ij} \times w_i(d_t)}{M} \qquad (4)$$

Wherein $W_i(d_t)$ is the importance of C_i, F_{ij} is the frequency of appearance of C_i in sentence S_j, M is all the words that sentence S_j contains; λ is the correct factor when the sentence is at the beginning or ending of paragraphs. It is 1.5 in this system.

(3) Compute Similarity of Sentence

In order to avoid overlap sentences in summary, we work out the cross-sentence word overlap according to the following formula:

$$\mathrm{R_s} = 2* \frac{(\#\text{overlapping words})}{(\#\text{words insentence1} + \#\text{words in sentence2})} \qquad (5)$$

The system sets 0.7 as threshold. If R_s exceeds the value, we deem that each pair has the same semantic, and select higher sentence value as summary to remove the following score sentence.

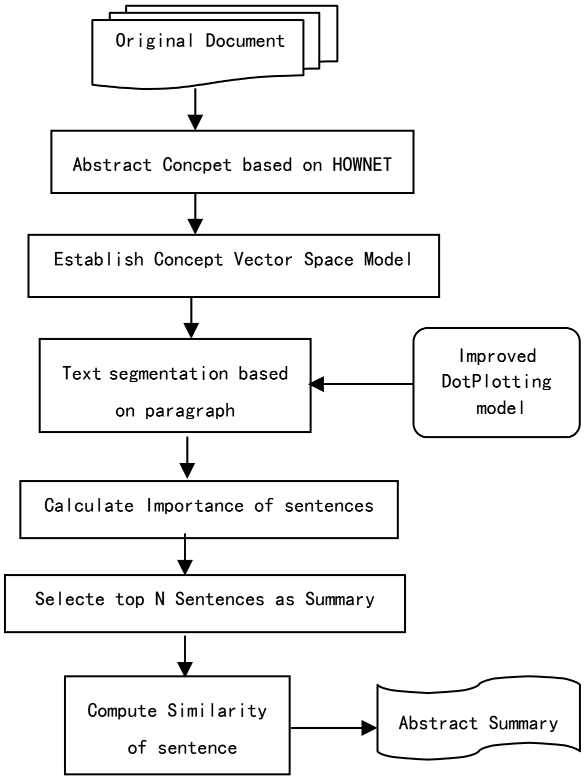
Original Document
Abstract Concpet based on HOWNET
Establish Concept Vector Space Model
Text segmentation based on paragraph
Improved DotPlotting model
Calculate Importance of sentences
Selecte top N Sentences as Summary
Compute Similarity of sentence
Abstract Summary

Figure 2. The process of Abstracting Summary

V. System Evaluations

A. Performance Evaluations

Generally, summaries can be evaluated by using intrinsic and extrinsic measures. While intrinsic methods attempt to measure the quality of summary by using human evaluation thereof, extrinsic methods measure the same through a task-based performance measuring such information retrieval oriented task. We adopt the former to evaluate the quality of summarization by defining the following parameters for evaluation.

This system uses intrinsic evaluation method to verify the algorithm of this paper. Traditional intrinsic evaluation indexes mainly include recall rate, accurate rate and F-Score. At present, the intrinsic evaluation method is generally automatic summarization evaluation method ROUGE [11,12] as proposed by Lin Chin-Yew et al. This method has been gradually adopted in DUC automatic summarization evaluation since 2006, but the testing data of DUC is in English. However, when we conduct automatic summarization evaluation for Chinese texts, the according corpus must be established. After that, we use the ROUGE method to evaluate the text automatic summarization.

(1) We use three parameters which are recall, precision and F_measure to evaluate the summarization system. Recall refers to the ratio of accurate recognition by system; precision refers to the ratio of exact recognition. The formula: recall R = N_{hm}/N_h, precision P = N_{hm}/N_m , N_{hm} is the number of sentences abstracted by the summarization system and experts simultaneously, N_h is the number of sentences abstracted by experts and Nm is the number of sentences abstracted by the summarization system, $F\_Score = \frac{2\times P\times R}{P+R}$.

(2) ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. There are five different ROUGE measures: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and Rouge-Su. Three of them have been used in the Document Understanding Conference (DUC) 2004, namely, Rouge-N, Rouge-S and Rouge-Su. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-S is Skip-Bigram Co-Occurrence Statistics, but Rouge-SU is an Extension of ROUGE-S which resolves the problem when ROUGE-S does not give any credit to a candidate sentence if the sentence does not have any word pair co-occurring with its references. This system will use Rouge-2 and Rouge-SU4 as evaluation criteria.

B. Evaluated Summary System

1 Coverage Baseline: choose the first sentence in first document, then choose the first sentence in the second document, and choose the first sentence in the n-th document; select the second sentence in the first document, then choose the second sentence in the second document…, until the summary is long enough. (Method 1)

2 Centroid-based summaries: this system is proposed by Dragomir R. Radev in Centroid-based summarization of multiple documents. (Method 2)

3 Text Segmentation based summary (TSS): the author describes the system. (Method 3)

C. Evluation Result And Analysis

Summary evaluation is a very important aspect for text summarization. Our evaluations on the three proposed summarization methods have been conducted based on a database of China's National Linguistics Work Committee which covers 200 articles covering economics, newspaper and literacy aspects.

We select three independent human evaluators which are employed to conduct manual summarization on the 200 documents contained in the evaluation database to obtain an objective summary. Each evaluator was requested to select exactly five sentences which he/she deems the most important for summarizing every document. Because of the disparities in the evaluators' sentence selections, 5 to 15 sentences in each document can be selected by at least one of the evaluators. Evaluation of recall, precision and F_measure parameter of each method are shown in Figures3-5.

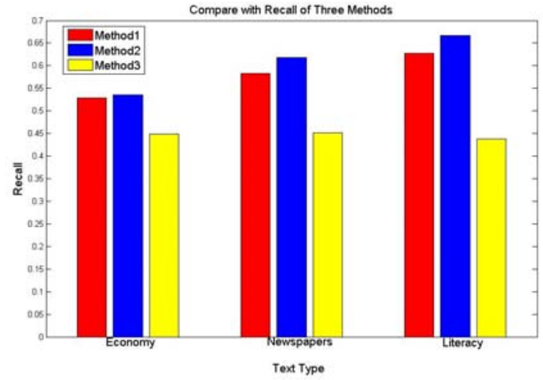

Figure 3. Result of Recall

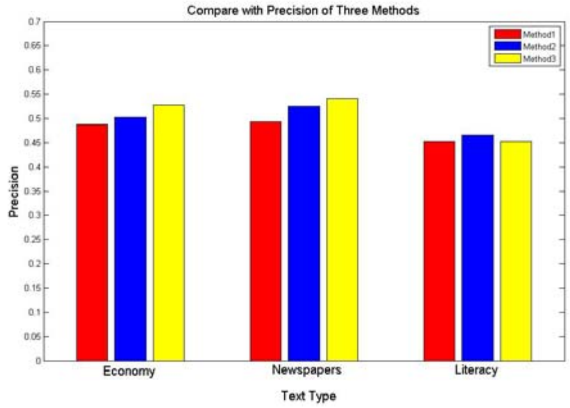

Figure 4. Result of Precision

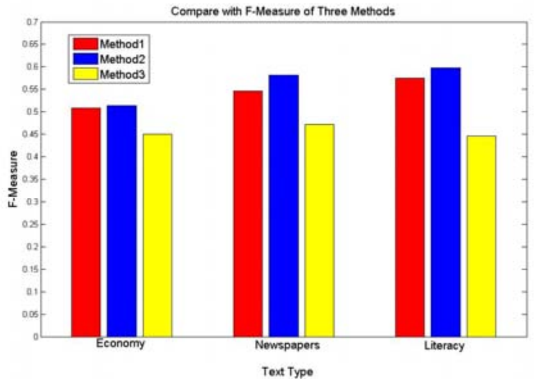

Figure 5. Rresult of F-Measure

From Figures 3-5, it distinctly shows that the economic type achieves a comparatively ideal result, while the newspaper and literature types shall be improved. The reason is that emphasis for texts of economy is explicit and the structure of those texts follows a regular pattern.

On the contrary, texts of newspapers and literature have no such features.

As is shown in Table 1, on this data set, the score of our system (TSS) is close to the mean score of all the participating systems (DUC2006) on ROUGE-2 and ROUGE-SU4. The mean scores of ROUGE-2 and ROUGE-SU4 are 0.0736 and 0.1288 respectively in DUC2006. TTS is slightly lower than the mean scores, but TTS generates a summary in Chinese instead of English. Different languages have different characteristics. The syntactic structure in Chinese is more complexly than that of English. In addition, most of participating systems (DUC2006) have adopted language tool, external corpora and knowledge database to help them understand the content of documents. This system uses basically statistical linguistics technology independent of any external resources, thus it isfaster and more independent.

TABLE 1.
RESULTS OF ROUGE-2 AND ROUGE-SU4

System Type	Rouge-2	Rouge-SU4
Method 1	0.0662	0.1112
Method 2	0.0691	0.1189
Method 3	0.0735	0.1281

From Tables 1-2, we can learn that TTS is more effective and efficient in performance than the other two systems, which means that using statistical linguistics technology to process documents can distinctly improve the quality of summary in a cost-effective manner.

ACKNOWLEDGMENT

This work is supported in part by a grant from Guangxi Natural Science Foundation (NO. 2013GXNSFBA019280) and Educational Commission of Guangxi Province of China (2013YB180).

REFERENCES

[1] D. R. Radev and W. Fan, "Automatic summarization of search engine hit lists", Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong Kong, 2000, pp. 99

[2] ISC "ISC Internet Domain Survey", Available http://ftp.isc.org/www/survey/reports/current/

[3] Zhou Xi, Wang Li, Pan Fuping , Dong Bin and Yan Yonghong. Automatic Scoring for English Spoken Question and Answer [J]. International Journal of Advancements in Computing Technology, 2013, vol 5(3), pp 448-455.

[4] Zhu Junwu, Jiang Yi, Li Bin and Sun, Maosheng. Ontology-based Automatic Summarization of Web Document [J]. International Journal of Advancements in Computing Technology, 2012, vol 4(14), pp 298-306.

[5] Enrique Alfonesca, and Jean-Yves Delort. A topic-model Based Approach for Update Summarization. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics ECAL 2012, Avignon, France, pp. 214–223.

[6] James Gung and Jugal Kalita. Summarization of Historical Articles Using Temporal Event Clustering. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 631–635.

[7] Seonggi Ryang and Takeshi Abekawa. Framework of Automatic Text Summarization Using Reinforcement Learning. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (July 2012), pp. 256-265.

[8] Kristian Woodsend and Mirella Lapata. Multiple Aspect Summarization Using Integer Linear Programming. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (July 2012), pp. 233-243.

[9] Seniz Demir,Sandra Carberry and Kathleen F. McCoy. Summarizing Information Graphics Textually. Computational Linguistics (2012), pp. 527-574

[10] Reynar JC. Topic Segmentation: Algorithms and Applications [D], University of Pennsylvania, USA, 1998.

[11] Lin C Y. ROUGE: A Package for Automatic Evaluation of Summaries // Proc of ACL Workshop on Text Summarization. Barcelona,Spain, 2004: 74 – 81

[12] Lin C Y, Hovy E. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics / / Proc of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton, Canada, 2003: 71 - 78

Meng Wang received the master's degree in Computer Science and Technology from Central China Normal University, in 2005. He is a Ph.d. candidate of Wuhan University of Technology and Associate Professor at Guangxi University of Technology. His interests are in Natural Language Processing and Feature Selection.

Xinlai Tang received the master's degree in Computer Science and Technology in Huazhong University of Science and Technology, in 2006. He is a Ph.d. candidate of Wuhan University of Technology and Associate Professor at Guangxi University of Technology. He is interested in Natural Language Processing.

Xiaorong Wang received the master's degree in Computer Science and Technology from Central China Normal University, in 2005. She is a Ph.d. candidate of Wuhan University of Technology and Associate Professor at Guangxi University of Technology. Her interests are in Natural Language Processing.