The Research on the Application of Text Clustering and Natural Language Understanding in Automatic Abstracting

Qinglin Guo[1], Cunbin Li[2]

1. School of Computer Science and Technology, North China Electric Power University, Beijing, 102206, China
2. School of Business Administration, North China Electric Power University, Beijing, 102206, China
qlguo88@sohu.com

Abstract

A method of realization of Automatic Abstracting based on Text Clustering and Natural Language Understanding is brought forward, aimed at overcoming shortages of some current methods. The method makes use of text Clustering and can realize Automatic Abstracting of multi- documents. The algorithm of twice Word Segmentation based on the Title and First-Sentences in Paragraphs is brought forward. Its precision and recall is above 95%. For a specific domain on plastics, an Automatic Abstracting system named TCAAS is implemented. The precision and recall of multi-document's Automatic Abstracting is above 75%. And experiments do prove that it is feasible to use the method to develop a domain Automatic Abstracting System, which is valuable for further study in more depth.

Keywords: *automatic abstracting, text clustering, natural language understanding*

1. Introduction

As one important research field of Natural Language, Automatic Abstracting has been necessary need in Internet time[1].Automatic Abstracting means that a computer can produce exact, laconic and even abstract from original text automatically. In some sense, information searching becomes more important than information itself.

1.1 Shortage of current automatic abstracting

There are four methods for Automatic Abstracting research and realization [2]: excerpted abstracting, abstracting based on comprehension, abstracting based on extracting as well as based on structure. But they are all unconvincing. For instance, excerpted abstracting may pile some sentences of original text, lacking consistency; abstracting based on comprehension is limited to some fields; abstracting based on extracted is usually similar and inanimate; abstracting based on structure needs to analyze text structure, so being complicated. In addition, current Automatic Abstracting system cannot work for multi-document but only for single-document. In fact, there are many documents that present the same problem [3]. Current Automatic Abstracting system can neither find these documents from vast electronic documents nor compile one abstract that can reflect the main idea automatically. So Automatic Abstracting based on text clustering and Natural Language Understanding is brought forward.

1.2 Structure and composition of automatic abstracting based on text clustering and natural language understanding

Automatic Abstracting system based on Text Clustering and Natural Language Understanding adopts multiplayer structure according to logic, namely Web server, applications server and database server. Web server supplies Input/ Output service, Boolean calculation service is offered for Web server by applications server; database server manages applications database, semantic knowledge base and domain base, etc.

The system is composed of: 1) Word Segmentation and module tagging; 2) single document abstract sentence module establishing based on statistic; 3) paragraph abstract sentence module establishing based on Sentence Clustering; 4) single document abstract sentence module establishing based on Sentence Clustering; 5) abstract sentence smoothing processing module based on NLU; 6) Multi-document Text Clustering module; 7) Multi-document Automatic Abstracting module; 8) knowledge base and regulation base.

2. Realization of automatic abstracting based on text clustering and natural language understanding

2.1 Automatic Word Segmentation

Presently, Word Segmentation methods mainly include Maximum Matching, Word-by-Word Matching [4], etc. Although there are over 10 methods, they cannot solve the identification of unknown word very well. Therefore, The method of twice Word Segmentation based on the Title

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

and Paragraph-First Sentence is brought forward in this paper (TTPFS for short).

The title and Paragraph-First Sentence in an article usually reflect the main idea. In addition, professional terms, shortened form and the words created by the author sometimes occur in title, subheads and Paragraph-First Sentence. Thus, Word Segmentation exactness must be ensured, and especially for those professional terms, shortened form and the words created by the author. So the method of twice Word Segmentation based on the Title and Paragraph-First Sentence is brought forward (TTPFS). The process of realization is: 1) Firstly find title, subheads and First-Sentences in paragraphs. 2) Do the first Word Segmentation for these sentences by electronic dictionary. The first Word Segmentation has 3 steps: Word Segmentation tagging, elementary Word Segmentation based on professional dictionary and the maximum bi-directional scanning Word Segmentation with the function of return as well as association. Thus, blank professional terms, shortened form and the words created by the author and other uncommon words may be written in the dictionary for the twice Word Segmentation. The twice Word Segmentation, namely Word Segmenting for the whole article by above-mentioned electronic dictionary. In this way, not only special lexis can be distinguished from a document, but also the efficiency of Word Segmentation can be improved greatly.

Accuracy and recall are important indexes to weigh the quality of Automatic Word Segmentation and Automatic Abstracting. Twenty articles on plastic in People's Daily in 1998 and 2000 have been searched. There are 15,083 words in all, 379 unknown words therein. Automatic Word Segmentation by TTPFS Arithmetic, 15,336 words were segmented, among those 14,661 correctly; 409 unknown word were identified, among those 352 correctly. The experimental result may be stated as the following Table 1.

Table 1. Experimental result of TTPFS

Category	Accuracy/%	Recall rate /%
Word Segmentation	95.6	97.2
Identification of unknown word	86.1	92.9

2.2 Realizing Model

Text Clustering is nuclear module in Automatic Abstracting system, which can gain the information of sentences clustering degree through computing the distance between sentences. In a certain paragraph, the topic sentence is usually the sentence that relates to other sentences closely [5]. In order to compute the distance between sentences, sentence vector parameter being confirmed as all non-functional words, namely vector denotation for sentences, is needed. As for sentence s_i and s_j, in the first place their vector denotation $(w_1,w_2,\cdots,w_m)$and $(w_1,w_2,\ldots,w_m)$are gained by Word Segmentation, $m<>n$. Because the two vectors have different dimensions, the data on the same dimension have different attributes and each element is not numerical value but word,$|w_{ik}-w_{jk}|$can not get directly. Therefore continental distance formula should be altered. Suppose:

$$d\,(w_{ik})=\min(d_{\mathrm{w}}(\,w_{ik},w_{jl}))\,. \qquad \textbf{(1)}$$

Therein: $d\,(w_{ik})\in[0,1]$; w_{ik} is one word in s_i, w_{jl} is one certain word in s_j, $d_{\mathrm{w}}(\,w_{ik},w_{jl})$ is the semantic distance between Word w_{ik} and w_{jl}, which can be found in semantic distance table. If there are not words to match w_{ik} in s_j, then $d(w_{ik})$ is 1. Thereby continental distance formula is altered as:

$$d_{\mathrm{s}}(s_i,s_j)=\Big[\sum_{k=1}^{m}[d(w_{ik})]^2\Big]^{1/2}\,. \qquad \textbf{(2)}$$

Obviously, the smaller for the numerical value of $d_{\mathrm{s}}(s_i,s_j)$, the nearer for the semantic distance between sentence s_i and s_j, so the higher for their clustering degree.

2.3 Realizing Process

The Automatic Abstract realizing process based on text clustering is as follows:

1) Construct semantic distance table. One sentence is composed of words. According to Formulas (1) and (2), semantic distance computing may involve relevant Acceptation distance in two sentences consequentially. Acceptation distance table constructed may be stated as the following Table 2.

Table 2. Acceptation distance

Word	w_1	w_2	…	w_j	…	w_n
w_1	0					
w_2		0				
|						
w_i				$d_{\mathrm{w}}(w_i,w_j)$		
|						
w_n						0

Planar coordinates elements of this table is composed of Word $w_1,w_2,\cdots,w_n$. Element $d_{\mathrm{w}}(w_i,w_j)$ means the element which is located in horizontal Line I as well as in the vertical Line j, it denoting acceptation distance between w_i and w_j. $d_{\mathrm{w}}(w_i,w_j)\in[0,1]$. Two sorts of extreme value of acceptation distance are: $d_{\mathrm{w}}(w_i,w_j)$=1 denotes the two words are antonym; while $d_{\mathrm{w}}(w_i,w_j)$=0 denotes they are synonym. All words in acceptation distance table are non-functional words, and besides, acceptation correlative description of certain words in the former thesaurus base

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE
IEEE COMPUTER SOCIETY

is specially amended according to work field the system faces. For example, the acceptation distances between professional lexis in the plastic industry are all defined below 0.5.

2) Sentence weighting. Text Clustering analysis is not needed to all sentences, but to the important sentences. In Sentence weighting, 4 sorts of sentences that can get upper weight are: a) title sentence, namely the sentence which involves the effective words occurring in the title; b) Sentence having high-frequency effective words; c) Sentence located in the important position in an article, such as the first and the last sentence in a certain paragraph, and also the sentence in the first paragraph, etc. d) Sentence having suggestive phrases, such as the phrase in a word, therefore, so, discuss, express, etc.

3) Sentence Vector denoting [6] and Clustering Analyzing. According to functional vocabulary and unused vocabulary, firstly to do obvious tagging for the functional words and unused words in each sentence, secondly to do Word Segmentation and Part-of-Speech Tagging (POS Tagging) that are based on TTPFS arithmetic for other strings. The treated sentences are denoted by vector, consequently getting vector model of each sentence （$w_1,w_2,\ldots,w_n$）. According to Eq. (2), semantic distance between sentence s_i and sentence s_j can be computed; then the sentences whose aggregate degree with other sentences is within the allowable confine value; and the abstracting sentences can be considered. Therein, α is the semantic distance confine value, being computed dynamically from average semantic distance between sentences by the Automatic Abstracting system. β is the number of abstracting sentence, being determined by the allowable length of abstract. Given that document D is the processing problem space, it has n sentences sample $s_i(i=1, 2, \cdots, n)$, then $D = \{s_1, s_2, \cdots, s_n\}$. Sentences Clustering for a document is to divide $D=\{s_1, s_2, \ldots, s_n\}$ into some subsets CS_1, CS_2, $\cdots$, CS_m, m being the number of category regulated in advance , besides it needs to satisfy

$$CS_1 \cup CS_2 \cup \ldots \cup CS_m = D \text{；}$$

$$CS_i \cap CS_j = \Phi(i \neq j) .$$

Sentence Clustering adopts system clustering. That is to say, sentences are separated into some categories by clustering (the number of category can be man-made), and then Rough Set of the abstract will be composed of important sentences in each category.

4) Abstract sentence smoothness processing based on NLU. Piling the abstracting sentences sometimes may not be coherent or fluent, so the smoothness processing will be needed. Automatic Abstracting system (TCAAS) adopts abstract sentence smoothness processing based on NLU.

2.4 Experimental Results

The abstracts of 20 articles on plastic in *People's Daily* (electronic edition) of 1998 and 2000 have been experimented with the help of Turing method. Firstly get their abstracts by TCAAS, and then get man-made abstracts by two teachers in Department of Chinese severally. Mix the 60 abstracts together. Finally, the Chinese experts were asked to mark the three abstracts of each article unwittingly. The results can be seen in Table 3.

Table 3. Experimental results of TCAAS

Category	Best		Better		Average	
	Number	Scale/%	Number	Scale/%	Number	Scale/%
Teacher1	6	30	9	45	5	25
Teacher2	7	35	5	25	8	40
TCAAS	7	35	6	30	7	35

From the experimental results of Table 3, we find that abstracting procedure of TCAAS is slightly better than manual abstracting: for one teacher at least, the exceeded scale was above 65%, while for two other teachers, the scale was 35%. TCAAS has attained applied level.

3. Abstract sentence smoothness processing based on NLU

3.1 Syntax analyzing model of smoothness processing

Syntax Analyzing of TCAAS adopts a kind of improved probability dependency model named Lexical Semantic Form (LSF for short) [7]. Suppose LSF is an analysis result of the character string $s=w_i\ldots w_j$, SR(R, h, w_i), meaning Word w_i in LSF depends on Word h through semantic relation. SR(i)=SR(R, h, w_i) can be got. The basis of this model is analyzing semantic relevant probability P(SR(i)|h, w_i) between words in Tree Set. This model supposes that there is high correlation between dependency relation R and sub-node, and then the inconsistency of data sparseness will be less. Thus, $w_i \cdots w_j$ analysis probability relative to LSF can be given.

$$\text{P(LSF}|w_i \cdots w_j\text{)}= \prod_{k=i, w_k \neq h}^{j} \text{P(SR}(k)|h, w_k\text{)} . \qquad \mathbf{(3)}$$

Given input $s=w_1\ldots w_n$, the task of random analyzer is to find the best analysis T^*:

$$\text{T}^*\text{=arg max} \underset{T \in Parse(w_1^n)}{p} (\text{T}|\, w_1^n). \qquad \mathbf{(4)}$$

In the formula: Parse($w_1{}^n$) is all the possible structure analysis for the input sentence s, P(T, $w_1{}^n$) is defined as the product of probability in all used LSF. This is a kind of probability analyzing model based on dualistic lexical dependency relation. While doing parameter training by the means of model, the scale of training corpus should be considered. It is nearly impossible that the words repeat in sentence analyzing. Thereby smoothness processing for the statistic results must be done [8]. So magnify lexical information through Hyponym part of speech so as to reduce the sparseness degree of data except for the close part of speech such as preposition, adverb.

3.2 Contents of abstract sentence smoothness processing

Abstract sentence smoothness processing based on NLU includes: 1) Cutting. The two sentences having the same subject can be combined together. If there are other same components, they should be omitted necessarily. 2) Combining. If two conjoint sentences only have one different component, then they can be combined as a sentence with multiplex components. 3) Parenthesis. Inserting some phrases may emphasize and correct signification, decorate text, or avoid different meanings. 4) Replacing. If the two sentences have the same subject or object, then the relevant component of the second sentence may be changed with a certain pronoun.

To do the above abstract sentence smoothness processing, semantic analysis and syntactic analysis for the abstract sentence are needed. In addition, language knowledge, regulation knowledge and concept hierarchy network theory are also needed.

4 Multi-document Automatic Abstracting model

4.1 Realization of multi-document's Automatic Abstracting

The core of Multi-document's Automatic Abstracting module is Document Clustering as well as the establishment of Multi-document abstract sentences. In Document Clustering, it is difficult to establish special vector model of the documents. TCAAS chooses vector model parameter [9] according to effective Title words, effective High-frequency words and effective words in the First Sentence of one paragraph. The parameters of every document have the same quantity, such as 15. As for these 15 effective words, different words weights were evaluated (range from 1~15), therein the effective Title words having the highest weight, effective High-frequency words higher, effective words in the First Sentence of one paragraph lowest. Computing the degree of vector similarity for documents may adopt the following formula:

$$\mathrm{S}(D_i, D_j)=[\sum_{k=1}^{m}[q(w_{ik})d(w_{ik})]^2]^{1/2} \quad . \qquad (5)$$

in the formula: D_i and D_j are two documents; w_{ik} is a certain effective word in D_i, $q(\ w_{ik})$ is the word weight; $d(w_{ik}\,)$ is the Acceptation distance between the effective word w_{ik} and its closest word in D_j, the formula is:

$$d\ (w_{ik}\,)=\min[d_{\mathrm{w}}(\ w_{ik},w_{jl})]\ . \qquad (6)$$

Hereinto: $d\ (w_{ik}\,)\in\ [0,1]$; w_{ik} is an effective word in D_i; w_{jl} is a certain effective word in D_j; $d_{\mathrm{w}}(\ w_{ik},w_{jl})$ is the Acceptation distance between w_{ik} and w_{jl}, $d_{\mathrm{w}}(\ w_{ik},w_{jl})$ can be found out in Acceptation distance table.

Degree of Clustering can be judged from the degree of similarity among documents. Several documents that have been chosen as the highest degree of Clustering are the objects of multi-document's Automatic Abstracting [10]. In multi-document's Automatic Abstracting, firstly abstract sentences of every document are determined by single document Automatic Abstracting module, secondly abstract sentences of multi-document are decided by Sentence Clustering and topic degree of asymmetry Analyzing (analyze the degree of departure with the user), finally multi-document's Automatic Abstracting smoothness processing should be done.

4.2 Experimental results and analysis of multi-document's Automatic Abstracting

From the training corpus of “domain information abstract & expert evaluating system”, 100 articles on plastic are samples waiting for test. 50 questions are designed according to these 100 articles. At beginning, ask experts to tag the association ratio for every question and every article. Then, the result of man-made tagging is to test by the means of system. Because of the limit of length, the experimental data of the former 5 questions on document recall and Document Clustering gives document accuracy in Table 4. The above-mentioned 5 questions are: market analysis for plastic of China in 2003, price trend forecast for plastic in 2004, market analysis and forecast for plastic in Shandong Province in the near future, price of plastic, where to buy needed PE & PB plastic products.

Table 4. Experimental result of multi- document’s Automatic Abstracting（1）

Category	Document Recall/%	Document Accuracy /%
Question 1	85.71	75.00
Question 2	100.00	100.00
Question 3	75.00	75.00
Question 4	93.75	88.24
Question 5	100.00	80.00

It is obvious that the average clustering document recall is above 90% and the lowest accuracy is 75% in multi-document's Automatic Abstracting module. This module can satisfy user's searching need.

The relation of document’s recall, document’s accuracy and document’s title association ratio has been studied. Document’s title association ratio means the degree of relevancy between document contents and document title. First, ask experts to mark every article with regard to association ratio (range from 0~1), and then do the statistic aimed at the relation of document recall, document accuracy and document title association ratio. The following Fig.1 denotes experimental result.

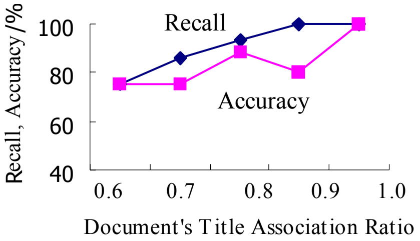

Fig. 1. Relation between document’s recall and document’s title

From Fig.1, it can be seen that the relation of document recall, document accuracy and document title association ratio is direct proportion relation, namely the higher the document title association ratio, the higher the document recall and accuracy. It is as a result of the weight of title is the highest while choosing the special vector model parameter for the document. In reality, might as well set an adjusting Function to adjust parameter weight. For instance, the title weight on technological paper and news may be adjusted higher, while for the scribble may be lower.

Table 5 gives experimental results of multi- document’s Automatic Abstracting aimed at news corpus, title weight 10~15.

Table 5. Experimental results of multi- document’ s Automatic Abstracting（2）

Category	Recall/%	Accuracy /%
Question1	92.7	90.9
Question 2	93.2	91.4
Question 3	83.5	78.3
Question 4	87.1	85.8
Question 5	89.2	86.6

5. Conclusion

Automatic Abstracting realizing methods based on Text Clustering and NLU are set forth, several innovative outcomes are briefly summarized as follows: 1) Text Clustering is introduced to Automatic Abstracting, thereby may conquer the shortage of routine Automatic Abstracting method; 2) multi-document’s Automatic Abstracting can be realized; 3) Twice Word Segmentation method based on Title and Paragraph-First Sentence (TTPFS) is put forward. TTPFS solved the recognition of “unknown word” in plastic industry. At the same time, Automatic Abstracting system (that is TCAAS) has been realized. TCAAS has built a website on domain information abstracting and expert evaluating system combined with a certain company’s website. This has been a successful endeavor.

Further in-depth research should be undertaken in the following areas. First, in smoothness processing of abstract sentence, methods to analyze sentence with the help of semantic block and sentence model and in-depth analysis of Chinese sentence analyzing theory based on semantic block and sentence model on the basis of semantic web should be explored. Second, construction of large-scale domain ontology should be studied. Third, transplant research outcome into automatic abstracting systems of other fields should be researched.

ACKNOWLEDGMENT

We would like to acknowledge the support from the National High Technology Research and Development Program (863 Program) (2004AA1Z2450) , the National Natural Science Foundation of China (70572090, 60305009) and the Ph.D Degree Teacher Foundation of North China Electric Power University (H0585) in China.

REFERENCES

[1] M.E. Califf and R.J. Mooney, “Relational learning of pattern-match rules for information extraction”, *Proceedings of the 19th National Conference on Artificial Intelligence*, Pittsburgh in USA, vol. 19, no. 1, January 2003,pp. 87-90.

[2] P. Brown and V. Della, “Class-based n-gram models of natural language”, *Computational Linguistics*, the MIT Press, Boston in Massachusetts, vol. 30, no. 4, April 2004, pp. 477-480.

[3] B. Terje and A. Jon, “Natural language analysis for semantic document modeling”, *Data and Knowledge Engineering*, Elsevier Science, Amsterdam in Netherlands, vol. 42, no. 1, January 2005, pp. 45-62.

[4] R. Bellman and L.A. Kalaba, “Abstraction and pattern classification”, *JMAA*, January 2005, vol. 52, no. 1, January 2005, pp. 1-7.

[5] E. Charniak, “ Statistical Techniques for Natural Language Parsing”, *AI Magazine*, AAAI press, Los Angeles in California, vol. 26, no. 4, April 2005, pp. 33-43.

[6] J. Philip, *A tutorial on Techniques and Application for Natural Language Processing*, 2rd ed., Pennsylvania: Carnegie-Mellon University Press，May 2004, pp.26-29.

[7] A. Blum, J. Lafferty, “Cluster kernels for semi-supervised learning”, *Proceedings of the 21th International Conference on Machine Learning*, Beijing in China, August 2004, pp. 92–100.

[8] A. Demiriz, K. Bennett, and M. Embrechts, “Semi-supervised clustering using genetic algorithms”, *Intelligent Engineering Systems*, Sydney University of Technology press, Sydney in Australia, vol. 10, no. 3, March 2004, pp. 809-814.

[9] B. Abdelhamid and P. Witold, “Data Clustering with Partial Supervision”, *Data Mining and Knowledge Discovery*, Berlin Heidelberg New York: Springer-Verlag, vol. 12, no. 8, August 2006, pp. 47-78.

[10] C. Charu and H. Jiawei, “On High Dimensional Projected Clustering of Data Streams”, *Data Mining and Knowledge Discovery*, Berlin Heidelberg New York: Springer-Verlag, vol. 11, no. 8, August 2005, pp. 251-273.

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE