Front. Electr. Electron. Eng. China (2006) 4: 460–464
DOI 10.1007/s11460-006-0088-y

RESEARCH ARTICLE

GUO Qing-lin, FAN Xiao-zhong, LIU Chang-an

The research and realization about automatic abstracting based on text clustering and natural language understanding

Abstract A method of realization of automatic abstracting based on text clustering and natural language understanding is explored, 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 investigated. 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 %. Also, the experiments prove that it is feasible to use the method to develop a domain automatic abstracting system, which is valuable for further in-depth study.

Keywords automatic abstracting, text clustering, natural language understanding

1 Introduction

As one important research field of natural language, automatic abstracting has been a necessary need in these days of the Internet [1]. In some sense, searching for information becomes more important than information itself.

Automatic abstracting means that a computer can produce exact, laconic and even abstract information from original text automatically [2].

Translated from *Transactions of Beijing Institute of Technology* (*Natural Science Edition*), 2005, 25(8): 705–709 (in Chinese)

GUO Qing-lin(✉), FAN Xiao-zhong , LIU Chang-an
Department of Computer Science,
North China Electric Power University, Beijing 102206, China
E-mail: qlguo88@sohu.com

FAN Xiao-zhong
Department of Computer Science and Engineering,
Beijing Institute of Technology, Beijing 100081, China

2 Shortages of current automatic abstracting

There are four methods of automatic abstracting research and realization: excerpted abstracting, abstracting based on comprehension, abstracting based on extracting as well as abstracting based on structure [3–5]. But they are all unconvincing. For instance, excerpted abstracting may pile some sentences of the original text, thereby lacking consistency; abstracting based on comprehension is limited to some fields; abstracting based on extracting is usually similar and inanimate; abstracting based on structure needs to analyse text structure, this makes it complicated. In addition, current automatic abstracting system cannot work for multi-documents but only for a single-document. In fact, there are many documents that discuss the same topic. 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 studied.

3 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 a multiplayer structure according to logic, namely Web server, applications server and database server. The Web server supplies Input/Output service; applications server offers Boolean calculation service for the Web server; database server manages applications database, semantic knowledge base and domain base and so on

The system is composed of

1) Word segmentation and tagging module.

2) Single document abstract sentence establishing module based on statistic.

3) Paragraph abstract sentence establishing module based

on sentence clustering.

4) Single document abstract sentence establishing module 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.

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

4.1 Automatic word segmentation

Presently, word segmentation methods mainly include maximum matching, word-by-word matching and so on. Although there are over ten methods, they cannot solve the identification of unknown words very well. Therefore, the method of twice word segmentation based on the title and paragraph-first sentence(TTPFS) is discussed in this paper.

4.1.1 Word segmentation realization based on TTPFS algorithm

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, especially for those professional terms, shortened form and the words created by the author. So the method of twice word segmentation based on TTPFS is vital. The process of realization is as follows. First, title, subheads and first-sentences in paragraphs is identified. Second, the first word segmentation for these sentences is done by using electronic dictionary. The first word segmentation has three steps: word segmentation tagging, elementary word segmentation based on professional dictionary and the maximum bi-directional scanning word segmentation with the functions 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 twice word segmentation. Then, the twice word segmentation, namely word segmenting for the whole article by the aforementioned electronic dictionary is implemented. In this way, not only can special lexis be distinguished from a document, but the efficiency of word segmentation can also be improved greatly.

4.1.2 Experimental result of automatic word segmentation TTPFS algorithm

Accuracy and recall ratio are important indexes to weigh the quality of automatic word segmentation and automatic abstracting. Twenty articles on plastic on People's Daily in 1998 and 2000 have been searched. There are 15 083 words in all, 379 unknown words therein. Using automatic word segmentation by TTPFS arithmetic, 15 336 words were segmented, among which 14 661 were correct and 409 unknown words were identified, among which 352 are correct. The experimental result is stated in 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

4.2 Realization of automatic abstracting based on text clustering

4.2.1 Realizing model

Text clustering is a nuclear module in automatic abstracting system, which can gain the information of sentences clustering degree by computing the distance between sentences. In a certain paragraph, the topic sentence is usually the sentence that relates to other sentences closely. In order to compute the distance between sentences, vector denotation for sentences is needed and sentence vector parameter being confirmed as all non-functional words[6]. As for sentence s_i *and* s_j, in the first place their vector denotation(w_1, w_2, ⋯, w_m) and (w_1, w_2,⋯, w_m)are gained by word segmentation, $m \neq n$. 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, $|w_{ik} - w_{jk}|$ cannot be got directly. Therefore continental distance formula should be altered. Suppose :

$$d(w_{ik}) = \min(d_w(w_{ik}, w_{jl})) \tag{1}$$

where $d(w_{ik}) \in [0, 1]$; w_{ik} is one word in s_i, w_{jl} is one certain word in s_j, $d_w(w_{ik}, w_{jl})$ is the semantic distance between word w_{ik} and w_{jl}, which can be found in semantic distance table. If there are no words to match w_{ik} in s_j, then $d(w_{ik})$ should be 1. Thereby continental distance formula should be altered as:

$$d_s(s_i, s_j) = \left[\sum_{k=1}^{m}[d(w_{ik})]^2\right]^{\frac{1}{2}} \tag{2}$$

Obviously, the smaller the numerical value of $d_s(s_i, s_j)$, the nearer the semantic distance between sentence s_i and s_j; therefore, the higher their clustering degree.

4.2.2 Realizing process

The automatic abstract realizing process based on text clustering is as follows:

1) Construct semantic distance table. One sentence is composed of some words. According to Eqs. (1) and (2),

semantic distance computing may involve relevant acceptation distance in two sentences consequentially. The acceptation distance table constructed is shown in Table 2.

Table 2 Acceptation distance

Word	w_1	w_2	⋯	w_j	⋯	w_n
w_1	0					
w_2		0				
⋮						
w_i				$d_w(w_i,w_j)$		
⋮						
w_n						0

Planar coordinate elements of this table are composed of word w_1, w_2, ⋯, w_n. Element $d_w(w_i, w_j)$ means the element that is located in horizontal line I as well as in the vertical line j, denoting acceptation distance between w_i and w_j. $d_w(w_i, w_j) \in [0, 1]$. Two sorts of extreme value of acceptation distance are as follows: $d_w(w_i, w_j) = 1$ denotes that the two words are antonyms while $d_w(w_i, w_j) = 0$ denotes that they are synonyms. All words in acceptation distance table are non-functional words, and besides, acceptation correlative description of certain words in the former thesaurus base is specially amended according to the 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 for all sentences, but only for the important sentences. In sentence weighting, four sorts of sentences that can get higher weightage are:

a) Title sentence, namely the sentence that involves the effective words occurring in the title.

b) Sentence having high-frequency effective words.

c) Sentence located in an important section of an article, such as the first and the last sentence in a certain paragraph, and also the sentence in the first paragraph and so on.

d) Sentence having suggestive phrases, such as the following phrases: in a word, therefore, so, discuss, express and so on.

3) Sentence vector denoting and clustering analyzing. According to functional vocabulary and unused vocabulary, first, obvious tagging for the functional words and unused words in each sentence should be done. Second, word segmentation and part-of-speech tagging (POS Tagging) based on TTPFS arithmetic for other strings should be carried out. The treated sentences are denoted by vector, consequently providing the vector model of each sentence (w_1, w_2, ⋯, 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 α and β as the abstracting sentences can be considered. Here, 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 problem space that is to be processed, it has n sentence samples $s_i (i=1, 2, \cdots, n)$, then $D = \{s_1, s_2, \cdots, s_n\}$. Sentences clustering for a document is to divide $D = \{s_1, s_2, \cdots, s_n\}$ into some subsets CS_1, CS_2, ⋯, CS_m, where m is the number of category regulated in advance. Besides it needs to satisfy

$$CS_1 \bigcup CS_2 \bigcup \cdots \bigcup CS_m = D;$$

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

Sentence clustering adopts system clustering. That is to say, sentences are separated into some categories by clustering (the number of category can be set). 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 is not coherent or fluent, so the smoothness processing is needed. Automatic abstracting system (TCAAS) adopts abstract sentence smoothness processing based on NLU.

4.2.3 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 the Turing method. First, their abstracts were extracted by TCAAS, and then several abstracts were drafted manually by two teachers in the department of Chinese. The 60 abstracts were combined. 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.

4.3 Abstract sentence smoothness processing based on NLU

4.3.1 Syntax analyzing model of smoothness processing

Syntax analyzing of TCAAS adopts a kind of improved probability dependency model called lexical semantic form (LSF for short). Suppose LSF is an analysis result of the character string $s = w_i \cdots w_j$, $S_R(R, h, w_i)$, meaning Word w_i in LSF depends on word h by semantic relation. $S_R(i) = S_R(R, h, w_i)$ can be got. The basis of this model is the analysis of

semantic relevant probability $P(S_R(i)|h, w_i)$ between words in tree set. This model supposes that there is high correlation between dependency relation R and sub-node; therefore, the inconsistency of data sparseness will be less [7]. Thus, analysis probability of $w_i \cdots w_j$ relative to LSF can be given as follows.

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

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

$$T^* = \arg\max_{T \in P_A(w_1^n)} P(T \mid W_1^n) \tag{4}$$

In the formula, $P_A(w_1^n)$ 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 relations. While doing parameter training by the means of this model, the scale of training corpus should be considered. It is nearly impossible that the words are repeated in sentence analyzing. Thereby smoothness processing for the statistic results must be done. So it is necessary to magnify lexical information through Hyponym part of speech to reduce the sparseness degree of data except for the close part of speech such as preposition, adverb and so on.

4.3.2 Contents of abstract sentence smoothness processing

Abstract sentence smoothness processing based on NLU includes the following.

1) Cutting: the two sentences having the same subject can be combined together [8]. If there are other same components, they should be omitted necessarily.

2) Combining: if two conjoint sentences only have one different component, 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 aforementioned 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.

4.4 Multi-document automatic abstracting module

4.4.1 Realization of multi-document's automatic abstracting module

The core module of multi-document's automatic abstracting is document clustering as well as the establishment of multi-document abstract sentences. In document clustering, it is difficult to establish the spacial vector model of the documents. TCAAS chooses the vector model parameters according to effective title words, effective high-frequency words and effective words in the first sentence of one paragraph. The parameters of every document are of the same quantity, such as 15. As for these 15 effective words, different words weights were evaluated (range from 1 to 15), therein the effective Title words having the highest weight, effective high-frequency words having a higher weight, effective words in the first sentence of one paragraph having the lowest weight. Computing the degree of vector similarity for documents may adopt the following formula:

$$S(D_i, D_j) = \left[\sum_{k=1}^{m} \left[q(w_{ik}) d(w_{ik})\right]^2\right]^{\frac{1}{2}} \tag{5}$$

In Eq. (5) 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

$$d\ (w_{ik}) = \min\ [d_w(\ w_{ik}, w_{jl})]$$

$d\ (w_{ik}) \in [0, 1]$. w_{ik} is an effective word in D_i. w_{jl} is a certain effective word in D_j. $d_w(\ w_{ik}, w_{jl})$ is the acceptation distance between w_{ik} and w_{jl}, $d_w(\ w_{ik}, w_{jl})$ can be found out in the acceptation distance table.

Degree of clustering can be judged from the degree of similarity among documents [9]. Several documents that have been chosen as the highest degree of clustering are the objects of multi-document automatic abstracting. In multi-document automatic abstracting, first, abstract sentences of every document are determined by single document automatic abstracting module. Second, abstract sentences of multi-documents are decided by sentence clustering and topic degree of asymmetry analyzing (by analyzing the degree of departure with the user). Finally multi-document's automatic abstracting smoothness processing is done.

4.4.2 Experimental results and analysis of multi-document's automatic abstracting module

From the training corpus of "domain information abstract and expert evaluating system," 100 articles on plastic were tested. Fifty questions were designed according to these 100 articles. In the beginning, experts tag the association ratio for every question and every article. Then, the result of manual tagging is tested by the system. Because of the limit of length, the experimental data of the former five questions on document recall and document accuracy are given in Table 4 by document clustering. The aforementioned five questions are as follows: 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 and PB plastic products.

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 need while searching for information.

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

The relation between document's recall, document's accuracy and document's title association ratio has been studied. A 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 to 1), and then do the statistic aimed at the relation of document recall, document accuracy and document title association ratio. Figure 1 depicts the result.

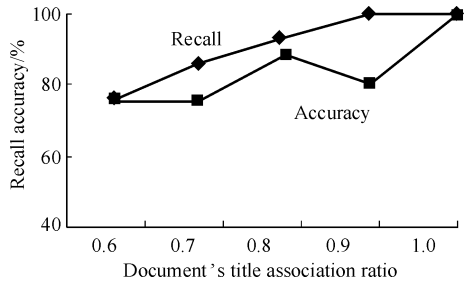

Fig. 1 Relation between document's recall and document's title

From Fig.1, it can be seen that the relation of document recall, document accuracy and document title association ratio is a directly proportionate relation, namely the higher the document title association ratio, the higher the document recall and accuracy. It is because the weight of title is the highest while choosing the spacial vector model parameter for the document. In reality, an adjusting function can be set to adjust parameter weight. For instance, the title weight on technological paper and news may be adjusted higher, while on scribble, it may be set lower.

Table 5 provides the 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/%
Question 1	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 Conclusions

Automatic abstracting realizing methods based on text clustering and NLU were explored. Several innovative outcomes were briefly summarized:

1) Text clustering was introduced to automatic abstracting, which may conquer the shortage of routine automatic abstracting method.

2) Multi-document's automatic abstracting was realized.

3) Twice word segmentation method based on TTPFS was studied.

TTPFS solved the recognition of "unknown word" in plastic industry. At the same time, automatic abstracting system (TCAAS) was 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.

Acknowledgements This study was supported by the National Natural Science Foundation of China (No. 70572090, No. 60305009), the Ph. D. Degree Teacher Foundation of North China Electric Power University.

References

1. Califf M. E., Mooney R. J., Relational learning of pattern-match rules for information extraction, Proceedings of the 19th National Conference on Artificial Intelligence, 2003, 19(1): 87–90
2. Li Lei, Zhong Yi-xin, The application of comprehensive information theory in automatic abstract system, Chinese Journal of Computers, 2000, 23(1): 4–7 (in Chinese)
3. Terje Brasethvik, Jon Atle Gulla, Natural language analysis for semantic document modeling, Data and Knowledge Engineering, 2001, 38(1): 45–62
4. Brown P., Della Pietra V., Class-based *n*-gram models of natural language, Computational Linguistics, 2002, 28(4): 477–480
5. Liu Ting, Wang Kai-zhu, Four kinds of main methods of automatic abstracting, Journal Information, 1999, 18(1): 11–19 (in Chinese)
6. Wu Si, Cluster analysis and Its application in the automatic information extraction from agricultural texts, Xiang tan: Xiang Tan University Press, 2001, 22–28 (in Chinese)
7. Yao Tian-shun, Natural language understanding, Beijing: Tsinghua University Press, 2002: 98–101 (in Chinese)
8. Li Jin-qian, Zhang Dong-mo, Yao Tian-fang, The optimization of sentence structure in natural language generation, The Research of Computer Application, 1998, 19(1): 53–54 (in Chinese)
9. Liu Chang-yu, Tang Chang-jie, Bayes discriminator for BBS documents based on latent semantic analysis, Chinese Journal of Computers, 2004, 27(4): 567–568 (in Chinese)