The Application and Problem Analysis of Computer Technology in English Information Processing

Zhaoli Wu¹,*

¹Shandong Women’s University, China

*Corresponding author e-mail: 163bujianbusan@163.com

Abstract. The traditional method of automatic summarization is based on the statistical extraction of abstract sentences from the grammar, without semantic analysis of the text, resulting in low summarization accuracy. In order to overcome the shortcomings of traditional methods, this article proposes an automatic summarization method based on topic concepts. An English automatic summarization system is designed and implemented based on concept statistics and analytic hierarchy. Concept statistics replace traditional word frequency statistics, based on the main super concept constructs a vector space model, calculates the importance of the sentence, selects the distribution of the main super concept on the concept hierarchy tree, analyzes the text structure and divides the meaning block, and extracts the abstract with the meaning block as the unit. Preliminarily solved the problem of unbalanced abstract structure of multi-topic articles. The experimental results show that through the methods of concept statistics and semantic hierarchy analysis, the abstracts generated by the English information processing system are more accurate and reflect the main content of the original text more comprehensively.

Keywords: Computer Application, Chinese Information Processing, Concept Planning, Topic Concept, Vector Space Model

1. Introduction

Automatic summarization can quickly condense and refine large-scale electronic texts, which is an accurate and efficient means to speed up reading and obtaining information resources. Mechanical automatic summarization is usually carried out in four steps: (1) Calculate the weight of the word; (2) Calculate the weight of the sentence; (3) Arrange all the sentences in the original text in descending order of weight, with the highest weight Ruqian Sentences are abstract sentences: (4) Output all abstract sentences in the order of their appearance in the original text.¹ The key is to calculate the importance of the sentence. In the research of automatic summarization at home and abroad, the following information is often used in the calculation of the importance of ten days: word frequency, position information, clue words, etc. The most basic information is the word frequency. According to the word frequency, the key words and words of the article are determined, and then the weight of the sentence can be calculated according to the number of valid words in the sentence. This is the basic basis of the mechanical abstraction method.² For clue words, Edmundson's abstract system has a pre-compiled clue word dictionary. The clue words in the dictionary are divided into three types:
Bonus Words with positive values, such as "important"; derogatory words with negative values (Stigma Words), such as "impossible"; Null Words, such as "hardly". The weight of the sentence is equal to the sum of the weights of each clue word in the ten-year period. The Xiangdong Space Model (VSM) based on word statistics is the basic method to calculate sentence importance. In this method, each sentence of the article is mapped to a vector \( S(T_i; W_i; T_2, W_2; \ldots, T_n, W_n) \) in an n-dimensional space, where \( T_i \) is a different word in the text to be processed, and \( W_i \) is usually taken as the frequency of \( T_i \) appearing in sentence \( S \) \([3]\). This method is based on pure morphology and believes that the more words appear in the text, the more important it is, and the semantic connection between different words is ignored \([4]\). The most basic assumption of VSM is that the vector terms \( T_i \) are orthogonal, and in real text, the topics reflected by words can be expressed through different morphological expressions. There is often a great correlation between the words as sub-items, which is not completely Independent \([5,6]\). For example, in an article about Bayesian Network Technology (Bayesian Network Technology), the theme concept "NetWork" is "network" that appears 6 times. As expressed by "net" which appears 2 times and "system" which appears 4 times, the method of word counting ignores the influence of "net" and "system" when investigating whether "NetWork" is a subject heading, causing part of it to reflect the theme of the article. The loss of keywords in the article cannot accurately reflect the content and structure of the article, thereby affecting the quality of the abstract.

In view of the shortcomings of traditional methods, this paper proposes to replace word statistics with concept statistics, build a vector space model based on topic concepts and extract abstract sentences.

2. Concept statistics and topic concept extraction algorithm

The concept of the article must first have a considerable connection with the content of the article, and at the same time, it should have a certain degree of inductive ability, not too general. For this reason, this algorithm defines three basic parameters-concept S-frequency, concept T-frequency, and concept induction degree to comprehensively investigate the possibility of a concept becoming a subject concept. The three parameters are described in detail below:

2.1. Concept selection

The degree of concept selection is a measure of the possibility of a concept becoming a theme concept, and is the basis for systematically determining the theme concept. The three parameters defined in the above analysis reflect the possibility of a concept becoming a subject concept from different aspects. Among them, S-frequency and T-frequency reflect the content of the article covered by this concept, so they are combined as a calculation factor, with different weights. The degree of induction reflects the inductive ability of the concept node and the relative importance to the child nodes, which is multiplied by the frequency factor as an independent factor. The value range of the induction degree is 0–1. If you multiply it directly, the difference in the selection degree will be too large and the expression of the frequency factor will be weakened. Therefore, two adjustable parameters are set to adjust it to a suitable range. Combining these factors, this article determines the formula for the selection degree \( Sel(C) \) of concept \( C \) as:

\[
Sel(C) = \alpha \cdot \log(F_s(C) + 1) + \beta \cdot \log(F_t(C) + 1) \cdot \gamma \cdot R(C) + \delta
\]

(1)

Among them, \( F_s(C) \), \( F_t(C) \), \( R(C) \) are the S-frequency, t-frequency and induction degree of concept \( C \), respectively. Because the logarithm cannot be taken when the frequency is 0, \( F_s(C) \) and \( F_t(C) \) in equation (4) are increased by 1 and then log is taken. \( \alpha, \beta, \gamma, \delta \) are weighting coefficients, which are used to adjust the weight of each parameter, which is determined based on experience and combined with experimental results. In this system, \( \alpha = 1, \beta = 0.25, \gamma = 1, \delta = 0.5 \). The greater the degree of selection, the more likely it is to be selected as the topic concept of the article.
Take the concept subject-matter in the concept hierarchy tree shown in Figure 1 as an example, \( F_s \) ("subject_matter")=0; the subtree with the concept "subject-matter" as the root node is shown in the dashed box, then \( F_r \)("subject_matter") = 11, \( R"subject_matter") = 1-6/11 = 0.45, \( Sel"subject_matter") = (4\log l+0.25\log 12)(0.45+0.5)=0.256.

3. Application of Concept Statistics in English Information Processing

3.1. Sentence VSM based on topic concept
After extracting the corresponding theme concept set, as shown in Figure 1, each theme concept can be used as the meaning of the VSM to establish the space vector model VSM. For each sentence S in the text to be processed, each word contained in S is classified into the corresponding topic concept, and a corresponding vector \( S(T_i;W_i;T_j;W_j;\cdots;T_n;W_n) \) is established, where \( T_i \) is the topic concept contained in the sentence, and \( W_i \) is the frequency corresponding to \( T_i \). In the article shown in Figure 1, suppose sentence S contains the word "application", and "application" belongs to the subtree with the subject concept "subject_matter" as the root node, so wherever "application" appears in S, "subject -matter" to represent, "subject_matter" is the corresponding item of Xiang Dong S.

After the vector space model is established, the importance of each sentence is calculated, and the most relevant important sentences are extracted to form an abstract. The formula for calculating the importance of sentence S is:

\[
W(S) = \lambda_{pos} \cdot \sum_{i=1}^{n} W_i \cdot I(T_i) / n
\]

(2)

Where \( \lambda_{pos} \) is the position weighting coefficient of S, \( \lambda_{par} \) is the importance of the paragraph where S is located, and \( I(T_i) \) is the importance of \( T_i \). After the importance of all sentences is determined, they are sorted by importance, and sentences with high importance are selected as abstract sentences, and they are arranged in the order of appearance in the original text to generate the first draft of the abstract.

![Figure 1. Undirected graph structure of the same theme concept](image)

3.2. An extraction example and comparison of results
Still taking the article "Building Natural Language Generation Systems" as an example for testing, extracting abstracts using thematic concept method and word statistics method respectively, the results show that compared with the word statistical method, the number of meaning items in the topic concept algorithm is greatly reduced. The keywords are grouped into 26 thematic concepts.

It can be seen from the results that the topic concept algorithm has a more ideal frequency distribution of meaning items compared with the traditional word statistics:

1) The frequency of high-frequency sense items has been strengthened; the high-frequency words of word frequency statistics are concentrated around 0.8, and after the topic concept induction, the high-frequency sense items are concentrated around 1.0. This shows that the topic concept algorithm effectively integrates and summarizes the semantic concepts contained in the processed text. The
high-frequency concepts contain certain semantic related words, and the frequency has been strengthened to make the topic concept more prominent and better reflect the article. theme.  

2) The high-frequency sense items increase; the highest peak of high-frequency words in word frequency statistics is 18%, and after the topic concept induction, the highest peak of high-frequency sense items increases to 25%. This shows that through the topic concept extraction, some low-frequency words have formed some new high-frequency meaning items through the semantic relationship induction, which is conducive to expressing the deep topics of the article that cannot be found by morphological statistics.

The final summary result of this article is: the recall rate of the theme concept algorithm $R=9/11$, the precision rate $P=9/12$, F-measure=0.78; the recall rate of the word statistical method $R=8/11$, the precision rate $P=8113$, F-measure=0.67, indicating that the abstract generated by the theme concept algorithm is closer to the expert abstract.

3.3. Expert Digest Corpus Test
measure=0.67 In order to test the effect of abstracts on a larger scale, we selected Ziff_Davis computer corpus from the TIPSTER artificial abstract corpus developed by LDC, 20 news corpus of San Jose Mercury & AP Newswire and 20 narrative corpus of USPatents. Two methods are used to extract abstracts of different lengths, and the test results are shown in the following table:

Table 1. The F-measure value of the two methods and the comparison test of the expert literature

| Genre     | Method         | 0   | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|-----------|----------------|-----|-----|-----|-----|-----|-----|
| News      | Word statistics| 0.09| 0.42| 0.45| 0.61| 0.48| 0.49|
|           | Theme concept  | 0.11| 0.46| 0.48| 0.67| 0.59| 0.49|
| Technologi| Word statistics| 0.22| 0.28| 0.40| 0.41| 0.48| 0.39|
| y         | Theme concept  | 0.25| 0.31| 0.42| 0.47| 0.53| 0.39|
| Narrative | Word statistics| 0.31| 0.37| 0.30| 0.19| 0.21| 0.11|
|           | Theme concept  | 0.18| 0.22| 0.35| 0.24| 0.20| 0.11|

Among them, the "length" column indicates the percentage of the extracted abstracts to the length of the text to be processed, and 0 means that the system only extracts one sentence as the abstract. It can be seen from the table that for articles of various genres, especially news and technology articles, the Fmniyiae value of abstracts obtained by the theme concept algorithm is higher than that of the word statistical method. This advantage is when the abstract ratio is moderate. It is more obvious when

4. Conclusion
This paper proposes a new method of mechanically extracting abstracts-topic concept algorithm. The algorithm replaces traditional word frequency statistics with concept statistics and semantic induction, uses topic concepts instead of word forms to build a vector space model, calculates sentence weights based on the importance of topic concepts, and extracts key sentences. Tests on a large number of corpus show that the abstracts obtained by the algorithm are closer to expert abstracts, which greatly improves the accuracy of the system's English information processing, and preliminarily solves the problem of unbalanced distribution of abstracts for multi-topic articles.

Acknowledgements
This work was supported by Exploration and Practice of Three-dimensional Comprehensive Teaching Evaluation System: A Case Study of Online and Offline Blended Teaching Mode of College English.

References
[1] Stanton N A , Rothrock L , Harvey C , et al. Investigating information-processing performance of different command team structures in the NATO Problem Space[J]. Ergonomics, 2015, 58(12):2078-2100.
[2] Karimi-Rouzbahani H, Vahab E, Ebrahimpour R, et al. Spatiotemporal Analysis of Category and Target-related Information Processing in the Brain during Object Detection[J]. Behavioural Brain Research, 2019, 1(2):42-50.

[3] Nakamura, Masahiro, Yoshida, Kciihiro. Non-transitory storage medium encoded with computer readable information processing program, information processing apparatus, information processing system, and information processing method[J]. Physics Today, 2017, 35(8):60-61.

[4] McGlew K, Shaw V, Zhang M, et al. An annotated database of Arabidopsis mutants of acyl lipid metabolism[J]. Plant Cell Reports, 2015, 34(4):519-32.

[5] Do, Cong Thuan, Choi, Hong Jun, Soo, Dong Oh. NTB branch predictor: dynamic branch predictor for high-performance embedded processors[J]. The Journal of Supercomputing, 2016, 72(5):1679-1693.

[6] Lerman G, Mccoy M B, Tropp J A, et al. Robust Computation of Linear Models by Convex Relaxation[J]. Foundations of Computational Mathematics, 2015, 15(2):363-410.