Computational Linguistics and Chinese Language Processing
Vol. 4, No. 1, August 1999, pp. 85-102

An Approach towards English Automatic Abstraction

Wu Yan*, James N.K.Liu+, Wang Kaizhu*

Abstract

This paper presents a hybrid approach for automatic abstraction of English text. This approach is based on statistical analysis and understanding of the text. An abstraction algorithm is introduced and its application discussed. Experiment results demonstrate that this hybrid approach absorbs the advantages of two kinds of automatic abstraction and achieves a better result.

Keywords: Automatic abstraction, statistical abstraction, understanding abstraction

1. Introduction

With the rapid development of science and technology, a recent trend is to access information via computers. As the amount of electronic information is increasing rapidly, it has become necessary to extract and condense information. Subsequently, automatic abstraction of text becomes an important research topic in Natural Language Processing (NLP).

Automatic abstracts are used to express the main ideas or certain aspect content of documents using computers. In the late 50's, Luhn attracted wide attention with his design for the world's first automatic abstracting system [Luhn 1958]. Since then, many famous automatic abstracting systems, such as the ADAM system [Pllock and Zamora, 1975] and ACSI-matic of IBM [IBM Corporation 1961] have been studied, and research has been carried out around the world [Eael, 1971; Paice, 1981; Zechner, 1996; Watanable, 1996; Fisher and Soderland, 1996; Riloff, 1994].

So far, the methods for automatic abstraction can be classified into two types, i.e., mechanism abstracting method based on statistics and understanding-abstracting method based on automatic text understanding. Mechanism abstracting method is simple. It

* Dept. of Computer, Harbin Institute of Technology. E-mail: wuy@insun.hit.edu.cn

+ Dept. of Computing, Hong Kong Polytechnic University. E-mail: csnkliu@comp.polyu.edu.hk

directly extracts important sentences from text as candidate sentences by statistical information, such as word frequency, sentence location, and clue words. The main advantage of this method is that it can process the text of an arbitrary field and can be realized easily. However, the quality of the abstraction result is not high and lacks logic among sentences. Unlike mechanism abstraction, understanding-abstracting method is based on NLP technology. The sentences of abstraction result are automatically produced by the abstracting system and are different from original text sentences. The quality of understanding abstraction is higher. The disadvantage of understanding abstracting method is that it can only process texts of certain fields and its realization is not easy.

2. Conceptual Framework

The approach discussed here combines the advantages of other automatic abstracting methods. First, it extracts important sentences of a text using only the statistical method. Then, it processes these sentences using a case relational model, analyses the text structure and understands these sentences. Finally, it outputs the result. This is an hybrid approach for automatic abstraction. It can process a text on any subject with better results.

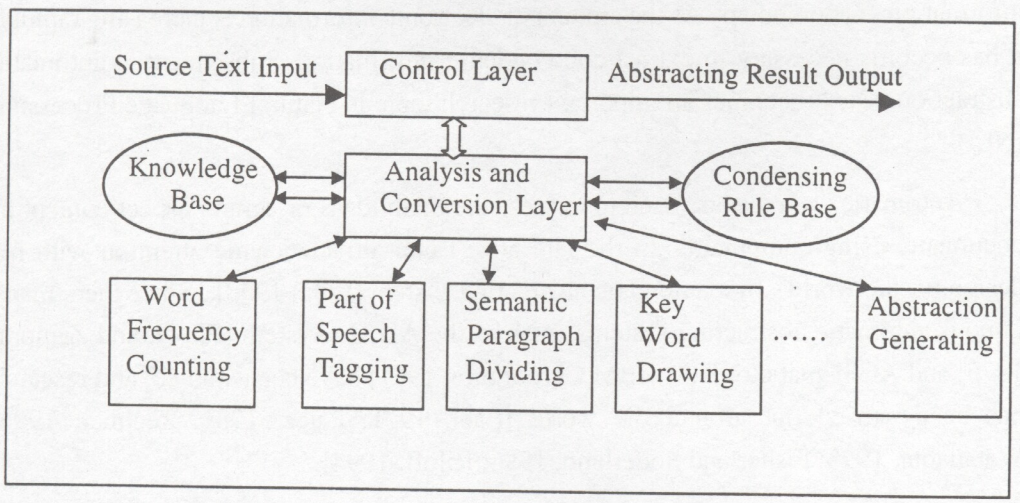

Figure 1 *Figure 1 The structure of an English automatic abstracting system.*

The processing mechanism is divided into two layers: the Control Layer and the Analysis Layer. The Control Layer involves works such as inputting the text, scanning and maintaining the Knowledge Base (KB) and Rule Base (RB), and outputting the

result; the Analysis Layer involves the following works: preprocessing the source text, calculating the word frequency, part-of-speech tagging, dividing the semantic paragraphs, extracting the important words and sentences, and condensing the abstracted sentences. Its structure is shown in Figure 1.

3. System Realization

3.1 Processing Flow

The procedure involves ten steps as follows: inputting the source text, preprocessing, calculating the word frequency, part-of-speech tagging, dividing the semantic paragraphs, extracting the content words, efficient words and key words, calculating the sentence weight, producing the Abstracting Base Set (ABS), analyzing the sentences of ABS, condensing the sentences of ABS, and outputting the abstraction result. Figure 2 shows the processing steps.

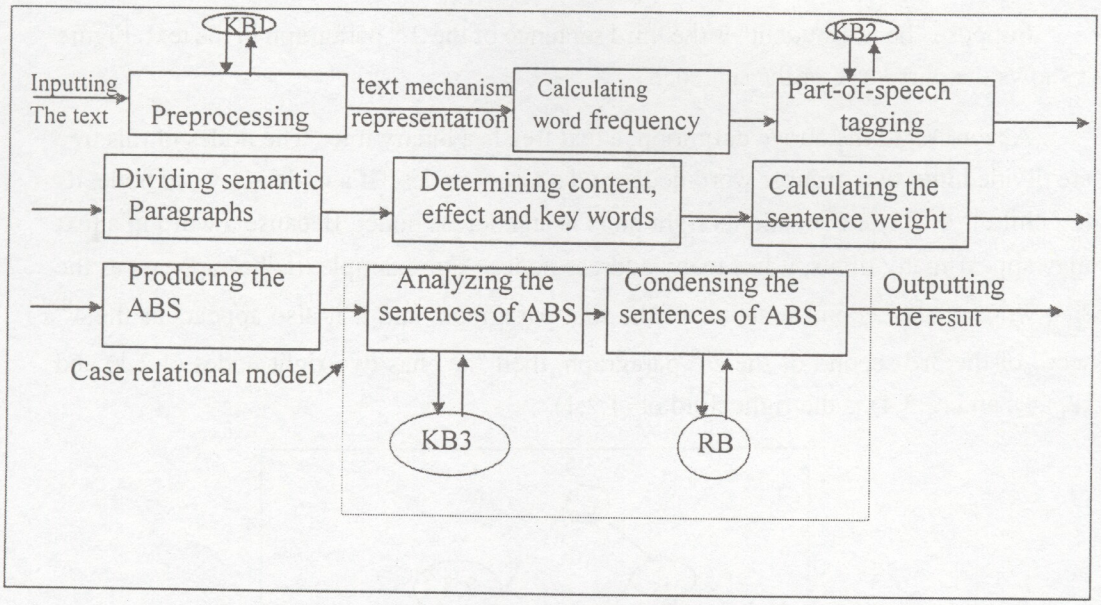

Figure 2 *The processing flow of English automatic abstraction.*

3.2 Realizing process

3.2.1 Preprocessing

The processor scans the source text twice. The first scan is used to screen out figures and tables in the text, and other auxiliary sentences, for example, "for instance", "perhaps",

"e.g.", "could not" and so on. *KB1* is a clue list and is manually established. It is good for any type of text. After this process is performed, the length of the source text can be shortened by 5% - 10%.

The second scan is used to produce the text structure tree. Let $T=S_1,S_2,...,S_n (n \geqq 1)$ be the source text, $S_k=w_1,w_2, ...,w_m (m \geqq 1)$ be a sentence of T, and w_m is the punctuation. Suppose ϕ_1 is a set of English words, and ϕ_2 is a set of other symbols of the text; then $\phi=\phi_1 \cup \phi_2$ is the set of all symbols (words, punctuation) of the text. The text structure tree is defined as follows:

1) the root is the first symbol appearing in the text;

2) if $w_i, w_j \in \phi$, and $w_i, w_j \in S_k$, $j=i+1$, then w_j is the left child;

3) if $w_i \in \phi$, then the right child of w_i is its location sequences;

4) if a node is the right child of another node, then it only has one child.

Suppose "he is a student" is the third sentence of the 2nd paragraph in the text. Figure 3 shows the parse tree of the sentence.

According to the above definition, a text tree is a binary tree. The nodes of this tree are divided into two groups: word nodes and address nodes. If a node is a word node, its left child is a word node; and its right node is an address node. Because a word in a text may appear many times, it has many address nodes. For example, if "he" appears as the first word of the second sentence of the first paragraph, and if it also appears as the 4th word of the 3rd second of the 2nd paragraph, then "he" has two right nodes (1,2,1) and (2,3,4), and (2,3,4) is the right child of (1,2,1).

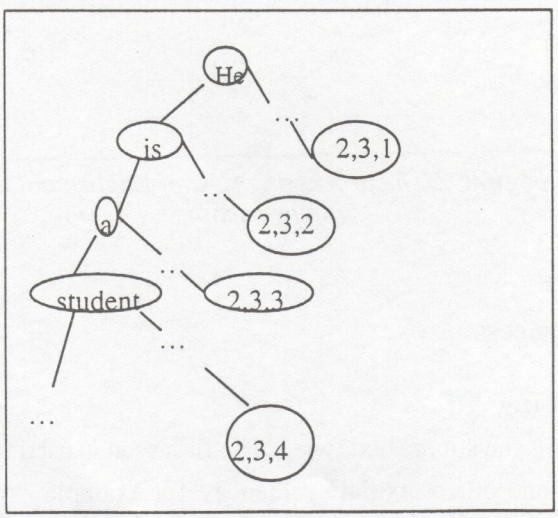

***Figure 3** An example of a text structure tree.*

3.2.2 Part of speech tagging

The part of speech restricted model is designed to determine the part of speech shown in Figure 4.

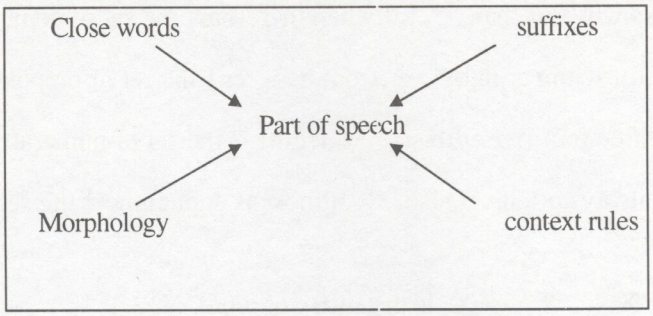

Figure 4 *Part of speech restricted model.*

The part of speech can help us determine key words, and it is the foundation for understanding and transformation of the ABS. In practice, the size of tagset in the system is taken to be ten. The experimental results indicate that the accuracy of tagging of unseen testing text was higher than 96%; the accuracy of tagging of visible testing text was higher than 97%. Furthermore, the accuracy of noun tagging was up to 99%.

- Theoretical model *TMOD*

The grammar of Chomsky I is the theoretical model of part of speech tagging, and its definition is as follows:

$TMOD = (V_I, T, P, S)$.

Here, $V_I=\{x1,x2,....,xn\}$ is a finite set of non-terminated symbols, and x_i indicates that the part of speech of the i^{th} word has not yet been determined. $T=\{N,V,J,F,D,U,Y,P,C,Z\} \cup$ FB is a finite set of terminal symbols. *N*--*Z* is the code for the part-of-speech. ***S*** is the set of all English sentences.

N--noun *V*--verb *J*--adjective *F*--adverb *D*-- definite article BD--indefinite article

U--indication *Y*--pronoun *P*--preposition *C*--connection *Z*--intersection

FB=DA ∪ YA ∪ VIA ∪ VAA ∪ CA ∪ PA ∪ SA ∪ FA ∪ SUA

Here, DA={the,a,an} the set of art words.

YA={this,that,we,it,our,my,us,his,they. ⋯ ,their} the set of pronoun words.

VIA={been,be,am,is,are,was,were} the set of copula words.

VAA={shall,cannot,may,have,had,might,can, ⋯ ,could} the set of auxiliary words.

CA={unless,while,or,than, ⋯ ,if,when,therefore} the set of conjunction words.

PA={to,of,for,with,per,in,by, ⋯ ,from,between} the set of preposition words.

SA={two,three,four,five,fifth,six, ⋯ ,tenth} is the set of numeral words.

FA={thus,always,often,yet,also, ⋯ ,almost,as,enough,not} the set of adverb words.

SUA={ (X_1 , X_2) X_1 is the suffix of word, $X_2 \in T_1$ }

Let T_1={N,V,J,F,D,U,Y,P,C,Z} be a syntax category set .

P is a set of deduction equations; the abstract representation is as follows:

$\alpha_1 A \alpha_2 \rightarrow \alpha_1 \beta \alpha_2$, here $\alpha_1, \alpha_2 \in T \cup [\,], \beta \in T_1, A \in V_1$.

KB2 includes three kinds of rules: close word matching rules, morphological matching rules, syntax and semantic matching rules. For example:

$$Y+X1+BD \rightarrow Y+V+BD.$$

That is, if a word is behind a pronoun and succeeded by an indefinite article, its part-of-speech is Verb.

The number of rules in *RB* is 1322; all of them were obtained through experiments. *RB* was manually established.

3.2.3 Dividing the semantic paragraphs

Text structure is the internal organization and construction of subject content so that the text has explicit layers and a logical structure. We define:

text meaning = word meaning + structure meaning.

Content words, effect words and key words denote Word meaning. Structure meaning is expressed by semantic paragraphs. Semantic paragraphs consist of consecutively natural paragraphs, which express the same subject.

Dividing semantic paragraphs requires an understanding of the structure meaning.

Definition 1. $A(w)=\left\{w \middle| w \in \bigcup_{i=1}^{n} f_i(w), f_i(w) \text{ is the word set of ith paragraph}\right\}$

If $B(w)$ is a set of closed words in the text, then $C(w) = A(w) - B(w)$ is a set of content words in the text.

Effect words make up a set of content words that its' frequency is bigger than L_1; key words make up a set of effect words whose frequencies are bigger than L_2. The title and subtitle are the important sentences, so the nouns appearing in the title and subtitle are the key words. The key words should include such words as "this paper proposed...", "the purpose of the paper...", and "to sum up...".

Let $\phi=\{\alpha_{ij}\}$ be a matrix of word repetition, i,j=1,2, $\cdots$, n

$$\alpha_{ij}=\begin{cases}\left|C_i(w)\bigcap C_j(w)\right| & \text{if } i<j\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

Here, $C_i(w)$ is the set of content words of the i^{th} paragraph, and $C_j(w)$ is the set of content words of the j^{th} paragraph. Following is the rule for dividing a semantic paragraph.

Rule: Suppose $p_1, p_2, \cdots, p_n$ are the paragraph codes of text T. p_i-- p_j can be treated as a semantic paragraph if they satisfy the following conditions:

1) $\alpha_{i-1,i}<\alpha_{k,k+1}$,k=i, $\cdots$,j-1.

2) $\alpha_{j,j+1}<\alpha_{k,k+1}$, k=i, $\cdots$,j-1.

3) $\left|C_i-\bigcup_{i_1=m}^{i-1}C_{i_1}\right| \div |C_i| \geq 0.6$, where m is the starting paragraph of the current semantic paragraph. There is a new content word in the first paragraph of the semantic paragraph.

4) $\left|\bigcap_{i_1=i}^{j}C_{i_1}\right| \div \left|\bigcup_{i_1=i}^{j}C_{i_1}\right| \geq 0$. There is a content word that appears in all the paragraphs of the semantic paragraph.

5) Suppose p_{k+j} is the semantic paragraph for $j=1,2 \cdots m$. If they satisfy the condition

$$\left|C_{k+1} \cap C_{k+m}\right| \geq 1$$

then $p_{k+1}, p_{k+2}, \cdots p_{k+m}$ is a semantic paragraph.

6) Suppose $p_1, p_2, \cdots, p_n$ is a semantic paragraph, and p_i satisfies the following condition:

$$\left| C_i - \bigcup_{i_1=1}^{i-1} C_{i_1} \right| \div \left|C_i\right| \geq 0$$

$$\left|C_i \cap C_{i+1}\right| > \left|C_{i-1} \cap C_i\right| \; and \; \left|C_i \cap C_{i+1}\right| > \left|C_{i-1} \cap C_{i+1}\right|$$

Then, p_i is the starting paragraph of the semantic paragraph.

Condition 1) and condition 2) ensure that the relations of the paragraphs in the semantic paragraph are the closest. Condition 3) ensures a new concept represented by each semantic paragraph. Condition 4) ensures that there is a concept, which is present in all of the paragraphs of the semantic paragraph. Condition 5) is used to merge the semantic paragraphs, each of them includes one paragraph.

For some text, conditions 1)-5) are rigorous, so the text has one semantic paragraph. Condition 6) is used to resolve above problem.

The precise of this approach is up to 84%. Experimental results indicate that the semantic paragraph can be used to extract the information from the text quite effectively.

3.2.4 Calculating the sentence weight and producing the ABS

There are numerous ways to calculate the sentence weight. For example, the method of Luhn first determines the effect word part and then calculates the weight of every part of the sentence. The sentence weight is the maximum weight of all the partial weights. The equation of dynamic calculation of the sentence weight is as follows:

$$F = H \cdot L \cdot \frac{m_1 \times \sum_{i=1}^{n_1} C_i + m_2 \times \sum_{i=1}^{n_2} E_i + m_3 \times \sum_{i=1}^{n_3} K_i}{S \times S_1 \times S_2} \qquad (2)$$

$$W_{word} = l^2 \cdot f \tag{3}$$

Here, H is a coefficient and is usually equal to 1; C_i, E_i, and K_i, respectively, denote the weight value of the i[th] content word, the i[th] effect word and the i[th] key word; equation (3) is their calculating equation; l is the number of words; f is the word frequency; $m=1$, $m_2=2$, and $m_3=3$ are, respectively, the weight coefficient of the content word, effect word and key word; n_1, n_2, and n_3 are, respectively, the number of content words, effect words and key words in a recent sentence; S is the total number of sentences; S_1 is the number of clauses; S_2 is the number of digital symbols in the sentence; $L=2$ indicates that the sentence represents a title or subtitle; $L=1.5$ indicates that the sentence is the first one in the paragraph, otherwise $L=1$. The algorithm of dynamic extraction is as follows:

Extraction algorithm

{

Apply equation (2) to calculate the weight value of S_i;

Sort the sentences in descending order according to the value of each individual sentence;

Extract the sentences with higher order;

Sort the extracted sentences according to the sentence location;

}

3.2.5 Analyzing the sentences of the ABS

After analyzing the sentences of the *ABS*, the system produces a sentence using the models *NP*, *VP*, and *PP*, and then determines the role of every part in the sentence using the Case Grammar of Filemore [Walter and Cook, 1979]. *KB3* includes two kinds of rules: phrase structure rules and case assigning rules. The phrase structure rules are built using phrase structure grammar. After processing, the structure of a sentence is as shown in Figure 5.

In Figure 5, M is the tense of the sentence; $K_1 \ldots K_n$ are symbols of the roles; $n=10$ is the role number. At present, research to determine the role of every part of the sentence is in the primary stage.

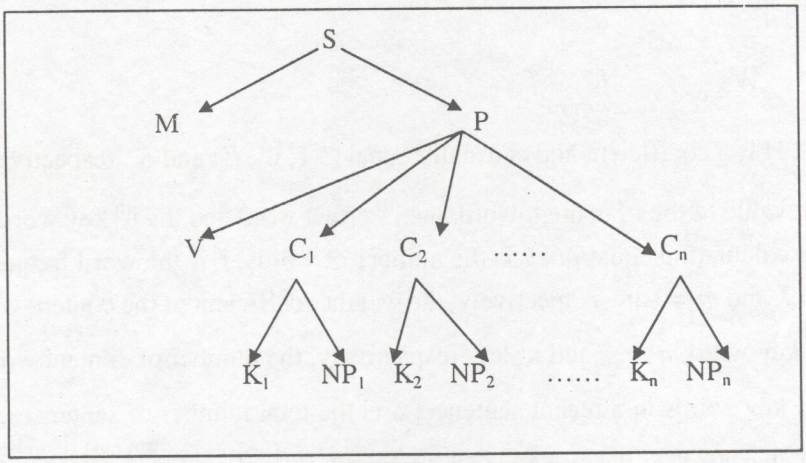

Figure 5 *Example of analyzing a sentence.*

3.2.6 Condensing the sentences in the ABS

RB is used to condense the sentences in the ABS. The condensing process includes the condensation of sentence classes and context classes. The former step is used to screen out the modifiers of nouns and verbs, such as adverb decorating adjectives in the attributes, and fashion adverbs and degree adverbs decorating verbs. The latter step is used to treat complex sentences. For example, if a complex sentence includes a query sentence and the response sentence is complete, the query sentence can be taken away. Unlimited attribute clauses and adverbial modifier clauses can be ignored. The process involves scanning of each sentence three times. The first scan is used to condense the sentence; the second scan is used to condense the context; the third scan is used to screen out any repeated sentences in the output.

There are two kinds of rules in the RB: sentence rules and context rules. The format of rules is as follows:

Ruler:=**IF**<Condition items> **THEN**<Operating items>

<Condition items>:=<Scanning form><Matching pattern>

<Scanning form>:=SL | SR

<Matching pattern>$\in (B \times (T_1 \cup T_2 \cup T_3) \times K') \cup [\,]$

<Operating items>:=DELETE<Parameter>

<Parameter>$\in T_3$

IF and *THEN* are judging sentences; *SL* and *SR* are left scanning and right scanning operations, respectively; **B** is a matching form, and it has three values: *B=1* denotes *AND*

matching, $B=2$ denotes *OR* matching, and $B=3$ denotes repeat matching; T_1 is the syntax category set just as above; T_2 is the syntax function set, $T_2=\{NP, VP, PP\}$; T_3 is an English word set; K is the role code defined by the system; *DELETE* is the deleting operation; $K'=K \cup \{\varepsilon\}$, $K=\{k_1,k_2,.....k_n\}$ is a set of symbols of roles.

RB includes two kinds of rules: sentence rules and context rules. For example:

Example1: $F\ J\ N \longrightarrow J\ N$.

That is, if an adjective is behind an adverb and succeeded by a noun, the adverb is deleted. For the English sentence "Mary is very beautiful girl.", "very " is deleted.

Example 2: $F,\ NP\ VP \cdots \longrightarrow NP\ VP \cdots$

That is, if an adverb is succeeded by a comma, and succeeded by a clause, the adverb is deleted. For the English sentence "Frankly, we are interested in going", "Frankly," is deleted.

Example 3: $NP \cdots, R \cdots, VP \cdots \longrightarrow NP\ VP \cdots$

The infinite attribute clause is deleted. For example, the English sentence "Mr. Green, who gives me piano lessons, has been ill recently" can be changed to "Mr. Green has been ill recently". For the English sentence "He was born in the year of 1949, when the new China was founded", the later clause that is an infinite attribute clause can be deleted.

4. The Experimental Results

4.1 Statistical results

With the method discussed above, the system called HIT-98 can process any text of an arbitrary field and produce an abstract with a specific content length requirement. Results from visible testing texts and unseen testing texts are given in Table 1 and Table 2, respectively. Note that the unseen texts were obtained from the Internet. The testing method was as follows: there were ten undergraduate students involved in our test. They read the abstraction result, then compared it with the original text, and scored the result with respect to readability and information. Finally, the quality of abstraction result was the average of the scores given by the testers.

The experimental results indicate that the method can preferably be used to process theses. Because a typical thesis usually focuses on one or a few problems, so the system can easily extract the important parts. However, there are more ideas in novels and

proses, so it is difficult to capture its gist.

4.2 Example

The following text was randomly extracted from some experimental texts.

① Preprocessing, calculating word frequency and part of speech tagging("/" is succeeded with word category, and the number is the word frequency).

"Money/N1 in/P5 the/D23 Soviet/N6 system/N2 is/G3 nothing/N1, "complained/V1 a/B20 highly/F1 paid/J1 writer/N2 who/R5 had/A3 never/F1 been/G2 allowed/V1 to/K24 go/V1 West/N2."You/T2 have/A5 to/K be/G2 able/J1 to/K spend/V1 it/T7. A/B Center/N1 Committee/N1 member/J2 does/A1 not/F5 get/V8 much/J2 pay/N1 but/R3 he/T7 gets/ 1 all/N3 kinds/N1 of/P16 things/N1 free/N1. He/T can/A1 get/V his/E4 children/N3 in/P the/D best/J1 universities/N3 or/I6 institutes/N6, or/I get/V them/T2 abroad/F2. "Then/F1 he/T paused/V and/I23 added/I1 sarcastically/F1, "They/T4 are/G3 all/ sending/V1 their/E5 children/N abroad/F now/N1, exporting/U1 them/T like/J1 dissidents/N1.

"There/F3 is/G also/F1 an/B3 informal/J1 network/N1 of/P connections/N3 that/Y3 enables/N1 a/B general/J1 to/K call/V3 a/B scientist/N3 to/K get/V his/E son/N2 admitted/V1 to/K an/B institute/N, a/B scientist/N to/K wangle/V1 a/B return/ 1 draft/N1 deferment/N1, or/I a/B movie/J1 scriptwriter/N1 who/R has/A1 produced/V1 a/B good/J2 Soviet/N spy/F1 film/N1 to/K call/V the/D security/N1 services/N1 to/K get/V permission/N1 for/P8 his/E wife/N1 and/I daughter/N3 to/K travel/V2 to/K the/D West/N. Blat/N3, as/P4 the/D Russians/N3 call/V influence/N1, is/G a/B constant/N1, vital/N1, and/I pervasive/J1 factor/N1 of/P Russian/N life/V1. "We/T1 have/A a/B caste/J1 system/N, "a/B senior/N1 scientist/N told/V2 me/T6. "Military/N2 families/N9 intermarry/N1. So/C1 do/A3 scientific/J1 families/N*NP, party/N7 families/N, writers/J families/N, theater/N1 families/N. Sons/N3 expect/V1 their/E fathers/N2 or/I fathers-in-law/N1 to/K promote/V1 their/E careers/N2 through/P2 blat/N, and/I fathers/N take/V1 it/T for/P granted/V1 that/Y they/T should/A1 do/V this/T2. Others/N1 do/V it/T. I/T3 did/V2 it/T for/P my/E4 son/N.

"Certain/J1 universities/N and/I institutes/N have/A become/ 1 known/N2 as/P the/D province/N1 of/P the/D party/N, government/N3, and/I military/N elite/V3 for/P their/E offspring/U1: at/P3 Moscow/N2 State/N1 University/N, the/D faculties/N1 of/P journalism/N1 and/I law/N1, since/P1 these/E2 are/G largely/F1 "political"/J1 fields/N1; and/I the/D Foreign/J3 Language/N1 Institute/N and/I the/D Moscow/N Institute/N of/P International/J1 Relations/N1 because/R2 they/T lead/ 1 toward/P1 foreign/N travel/V and/I foreign/J careers/N. These/E are/G known/J as/P places/N2 where/R1 some/N3

of/P the/D highest-ranking/N1 party/N and/I government/N people/N1 place/V their/E sons/N and/I daughters/N or/I grandsons/N1 and/I granddaughters/N1, frequently/F1 using/U1 blat/N to/K get/V flunking/U1 grades/N2 on/P2 entrance/N2 examinations/N1 falsely/F1 changed/V1 to/K A*s/N1.

"You/T have/A to/K have/V very/F1 good/J party/N and/I Komsomol/N1 recommendations/N1 to/K get/V into/P2 MIMO/N2, "one/T3 graduate/V1 told/V me/T, and/I he/T mentioned/V1 a/B score/N1 of/P sons/N and/I daughters/N of/P party/N and/I government/N officials/N1 who/R got/ 1 in/F through/P connections/N. He/T himself/T1 came/N1 from/P5 a/B party/N family/N, and/I said/V4 the/D whole/J1 student/N3 body/N1 had/V a/B clubby/N1, elitist/N1 atmosphere/N1.There/F were/G4 not/F many/J1 "ordinary"/N2 student/N because/R, although/R2 this/T was/G3 not/F a/B secret/J1 institution/N4, it/T was/G not/F listed/V1 in/P the/D normal/J1 handbook/N1 of/P Soviet/N institutions/N of/P higher/J1 education/N1 for/P prospective/N1 applicants/N1.

My/E friend/N3 said/V he/T knew/N1 of/P an/B instructor/N1 at/P MIMO/N, a/B party/N member/N, who/R had/A been/G fired/V1 for/P refusing/U1 to/K obey/V1 orders/N1 from/P the/D dean/N1 to/K give/V1 top/J1 grades/N to/K children/N from/P elite/V families/N, derisively/F1 know/ 1 among/P1 some/J Russians/N as/P Sovetski/N1 detki/N1, "the/D Soviet/N kids/V1". In/P his/E time/N1, he/T said/V, there/F were/G any/J1 number/N1 of/P students/N from/P ranking/U1 families/N who/R did/V poor/N1 work/V1 but/R were/G protected/V1 from/P expulsion/N1 by/P1 family/N connections/N.

One/T day/N1 some/J young/N1 friends/N offered/V1 to/K smuggle/V1 me/T into/P the/D institute/N for/P a/B look/N1 around/F1. It/T was/G one/T of/P those/E1 closed/W1 Soviet/N institutions/N, with/P3 no/J1 sign/N2 on/P the/D door/N1 to/K indicate/V1 its/E3 name/N1 or/I function/N1 and/I with/P guards/N2 posted/W2 to/K keep/V1 out/P1 the/D unwanted/W1. Although/R a/B sign/N at/P the/D entrance/N said/N plainly/F1, PRESENT/N1 YOUR/N1 PASSES-IN/N1 OPEN/N1 FROM/N1, my/E friends/N assured/V1 me/T --and/I I/T found/V2 they/T were/G right/J1--that/Y a/B firm/N1, knowing/U1 nod/N1 of/P the/D head/N1 and/I a/B steady/J1 stride/N1 would/A1 be/G enough/J1 to/K get/V me/T past/P1 the/D guards/N. My/E escorts/N1 showed/V1 me/T the/D posted/W academic/J1 curriculum/N1 and/I the/D library/N1 with/P its/E special/J1 "fund"/N1 of/P Western/N1 newspapers/N1 and/I books/N1. But/R I/T found/V it/T disappointingly/F1 similar/J1 to/K much/J more/N1 ordinary/N Soviet/N institutions/N, for/P all/ its/E elite/N status/N1.

The part of speech tagging accuracy is 97% 。

② Dividing the semantic paragraph. The matrix of content word repetition is as follows:

$$\Phi_c = \begin{Bmatrix} 0 & 5 & 2 & 1 & 3 & 2 \\ 0 & 0 & 7 & 6 & 5 & 2 \\ 0 & 0 & 0 & 4 & 2 & 2 \\ 0 & 0 & 0 & 0 & 6 & 3 \\ 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{Bmatrix}$$

The result of dividing the semantic paragraph is as follows: the first semantic paragraph includes the paragraphs 1~5, and the second semantic paragraph includes the 6th paragraph.

This result is the same as the manual processing result; the first semantic paragraph explains that the background of family is more important than money in Russia. The second semantic paragraph describes the experience of author.

③ Extracting characteristic words

Characteristic words are nouns, and their word frequencies are greater than and equal to 2.

④ Generating and analyzing basic extraction （ the extracting proportion is 30% of the whole text, , "/" is succeeded with phrases category, and Ki is the case role. ）

- The basic extraction is as follows:

"Money/NPK1 in/PP the Soviet system/NP is/VP nothing/NPK2," complained a highly paid writer/NPK1 who had never been allowed/VP to go/VP West/VPK5.

Blat/NP, as/PP the Russians call influence/NP, is/VP a constant/NP, vital/NP, and pervasive factor/NP of/PP Russian life/NPK2. "Military families/NPK1 intermarry/VP. So/NP do/VP scientific families/NPK1, party families/NPK1, writers' families/NPK1, theater families/NPK1."

Certain universities and institutes/NPK1 have become/VP known/NPK2 as/PP the province/NP of/PP the party/NP, government/NP, and military elite/NP for/PP their offspring/NPK4: at/PP Moscow State University/NPK5, the faculties/NP of/PP journalism and law/NP, since these/NP are largely/VP "political" fields/NPK2; and the Foreign Language Institute and the Moscow Institute of International Relations/NPK2 because they/NPK1 lead/VP toward/PP foreign travel and foreign careers/NP.

"You/NPK1 have to have/VP very good party and Komsomol recommendations/NPK2 to get/VP into/PP MIMO/NPK5," one/NPK1 graduate/VP told me/NPK2, He himself/NPK1 came/VP from/PP a party family/NPK6, and said/VP the whole student body/NPK1 had/VP a clubby/NPK2, elitist atmosphere/NPK2.

My friend/NPK1 said/VP he/NPK1 knew/VP of/PP an instructor/NPK2 at/PP MIMO/PPK5, who/NP had been fired/VP for/PP refusing to obey/VP orders/NP from/PP the dean/NPK6 to give/VP top grades/NPK2 to children/NP from/PP elite families/NPK6, derisively know/VP among/PP some Russians/NP as/PP *Sovetski detki*/NP, "the Soviet/NPK1 kids/VP".

One day some young friends/NPK1 offered/VP to smuggle/VP me/NPK2 into/PP the institute/NPK5 for/PP a look/NP around. It/NPK1 was/VP one/NPK2 of/P those closed Soviet institutions/NP, with/PP no sign/NP on/PP the door/NP to indicate/VP its name or function/NP and with/PP guards posted/NP to keep/VP out/PP the/D unwanted/NP. But I/NPK1 found/VP it/NPK2 disappointingly similar/NP to much more ordinary Soviet institutions/NP, for/PP all its elite status/NP.

⑤ Condensing basic extraction and generating abstract results (30%).

- Abstracting result is as follows:

This paper includes 2 parts:

Part 1: "Money in the Soviet system is nothing," complained a highly paid writer who had never been allowed to go west.

Blat ' is a constant, vital, and pervasive factor of Russian life. "Military families intermarry. So do scientific families, party families, writers' families, theater families."

Certain universities and institutes have become known as the province of the party, and military elite for their offspring: since these are largely "political" fields; and the Foreign Language Institute and the Moscow Institute of International Relations because they lead toward foreign travel and foreign careers.

"You have to have very good party and Komsomol recommendations to get into MIMO," one graduate told me, He himself came from a party family, elitist atmosphere.

My friend said he knew of an instructor at MIMO, "the Soviet kids". He said, there were any number of students from ranking families who did poor work but were protected from expulsion by family connections.

Part 2: One day some young friends offered to smuggle me into the institute for a look around. It was one of those closed Soviet institutions, but I found it disappointingly

similar too much more ordinary Soviet institutions, for all its elite status.

5. Conclusion

The development of automatic abstraction can be divided into two phases: mechanism abstraction from the late 1950's to the early 1970's and understanding abstraction from 1970's until now. In fact, Mechanism abstraction mainly relies on the statistical method. It has the advantage of processing arbitrary field texts, but has the shortcoming of lacking logical relation among sentences. Understanding abstraction makes use of NLP technology. After understanding the source text, the system produces an intermediary text expression. At last, the abstract is generated by the system. Understanding abstraction can only process texts on certain subject, although the quality of its abstracting result is better.

The paper presents an approach, which integrates statistics and text understanding to perform abstraction. It combines the advantages of both two abstracting approaches, and achieves better results and higher quality as shown in Tables 1-2.

Table 1. *Visible testing text results.*

Text field	The number of texts	the accuracy of POS Tagging	the accuracy of case analysis	quality of abstraction
Argument	50	97.3%	95.4%	98.3%
Explaining	45	98.2%	96.5%	98.5%
Novel	30	96.5%	95.1%	92.1%
Prose	10	97.1%	95.8%	81.2%

Table 2. *Unseen testing text results*

text field	The number of texts	the accuracy of POS Tagging	the accuracy of case analysis	quality of abstraction
Argument	40	96.3%	93.4%	97.3%
explaining	50	96.2%	93.5%	96.5%
novel	20	96.5%	94.1%	90.4%
prose	5	96.1%	92.8%	79.5%

References

David Fisher, Stephen Soderland , "DESCRIPTION OF UMASS SYSTEM AS USED FOR MUC-6," *Proceedings of the 6th Message Understanding Conference*, 1996.

Earl, L.L. etal. ,"*Annual Report: Automatic Information Abstracting and Extracting*," Lockheed, Missel and Space Company, Palo Albo, California, March, AD721066, 1971.

Hideo Watanabe, "A Method for Abstracting Newspaper Articles by Using Surface Clues," *Proceedings of 16 th International Conference on Computational Linguistics*,Vol.2, 1996.

IBM Corporation, "Advanced System Development Division," *ACSI-matic Auto-Abstracting Project, Final Report*, Yorktown Heights, New York, Vol. 3, 1961.

Klaus Zechner , "Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relavent Sentences," *Proceedings of 16th International Conference on Computational Linguistics*, Vol.2, 1996.

Luhn H.P., "An Experiment in Auto-Abstracting," Proceedings of International Conference on Scientific Information, Washington,D.C, 1958.

Paice,C.D. "The Automatic Generation of Literature Abstracts: An Approach Based on the Identification of Self-indicating Phrases," *Information Research*. London,1981.

Pllock,J.I.and Zamora,"Automatic Abstracting Research at Chemical abstracts,"*Journal of Chemical Information and Computer Science*, 1975, pp.226-232.

Riloff E.M., "Information Extraction as A Basic Set for Portable Text Classification System," *Ph.D. Thesis*, University of Massachusetts, Amherst, U.S.A., 1994.

Stephen Soderland, "Learning to Extract Text-based Information from the World Wide Web, *Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining(KDD-97)*,1997.

沈達陽，孫茂松，黃昌宁，局部統計在漢語未登錄詞辨識中應用和實現方法，第四屆計算語言學學會論文集， 1997.

王開鑄，李俊杰， HIT-863I 型非受限域中文自動文摘系統， "863" 智能接口與應用學術會議， 1995.7.