Corpus-Based Chinese Text Summarization System

Jun-Jie Li and Key-Sun Choi
CSLab, Center for AI Research, Korea Advanced Institute of Science and Technology
Taejon, Republic of Korea
Tel: +82-42-869-5565, Fax:+82-42-869-8700
E-mail:{jklee,kschoi}@world.kaist.ac.kr

Abstract

A Chinese Text Summarization system is developed, which is based on the surface information of context as well as the corpus based word segmentation and keyword identification. Unknown words identification is the most difficult topic on Chinese Word Segmentation. The context information is utilized here to resolve the unknown words and ambiguous segmentation problem by integrating word frequency and word length to dynamically weight the word weight, the theory and experiments show that this approach is superior than traditional dictionary based matching approach and pure word frequency-based statistical approach. The segmentation precision is 98% for real text. The keyword identification is not only based on word frequency but also word length, salient sentence determination is solved by using word weights, sentence length, number of clauses, numeric word and unknown words etc.., less relying on sentence position and surface cues. The evaluation measures of summary is studied and experimental results are provided.

1. Introduction

Text Summarization System is to identify and select the central content or user inquired content from the given original texts to form the summarized output with the sentences identical to the original input text or new generated. There are three kinds of approaches on developing Text Summarization System: the first one is based on the surface-clues of the current context such as the word frequency (Luhn, 1958), sentence position, word clue or indication , title sentence(Watanabe,1996), word association or rhetorical relations(Ono et al.,1994) and linear heuristic sentence weighting function (Zechner, 1996).Its advantages are simple and domain unconstrained, its shortcoming is inaccuracy in sentence abstracting due to the uncertain valve of word frequency for key words, varied distribution of important sentences and heuristic function itself. The second one is based on the knowledge-based natural language processing techniques, such as Script-based summarization system for given texts with multilingual output(Tait, 1985), CD-based domain constrained abstracting system with incomplete syntactic and semantic analysis (Dejong, 1979), rule-based summarization system with forward and backward scanning schema (Danilo,1982) etc.. Its advantages are more accurate and in depth language analysis and generation. Its shortcomings are domain constrained and difficulty in knowledge base maintenance. The third one is the corpus based methods (Li, 1995; Li and Choi, 1997). The corpus based sentence segmentation, non-linear sentence weighting function , collocation computation based word and sentence importance analysis and efficient raw corpus and text indexing method, give this method a prospective future.

In section 2, a full text indexing method called Natural Hierarchical Network (NHN) is illustrated. In section 3, the word segmentation algorithm is introduced, the text summarization system is illustrated in 4, the experimental results are given in the section 5, and finally the conclusion is given in section 6.

2. Natural Hierarchical Network

In order to make full use of context and text corpus information such as character and/or word frequency and collocation.We design a new full text indexing method, called Natural Hierarchical Network (NHN) shown as Fig. 1.

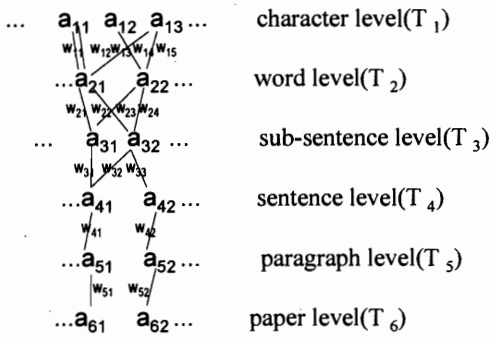

Fig.1 Description of Natural Hierarchical Network

The meaning of NHN is that every language unit(character,word, sentence, paragraph) have a vector to corresponds to its every occurrence in texts, and in turn, the texts(or raw corpus) can be indexed and represented by all the occurrences of elements in a certain level m (say character level) represented by the vectors as above. Certainly, we can omit some levels to make the vector shorter, however that will lose some useful information of text structure and language usage.

In practice, according to the sentence and sub-sentence ending symbols(i. e., punctuation such as . , : ? !) and the format and markers of text(such as writing rules or custom of paragraph, chapter , title and subtitle), the input text can be automatically converted into a series of vectors as (pp, pa, sn, ss, wd, ct, c_i), where pp, pa, sn, ss, wd and ct are respectively represent the sequential numbers of paper, paragraph, sentence, sub-sentence, word and character that character c_i appears in.

3. Word Segmentation

Chinese Word Segmentation(Li, 1995; Chen and Lee, 1996) is an ever-green topic due to the unknown word identification and ambiguity resolution problem. We provides an dynamic word weighting function which calculates the frequencies of words(exactly speaking, strings) in context(or textual corpus) in the run time, segmentation algorithm is designed as follows:

Algorithm Segment(s)

{

/* given a string s(initially is the whole sentence), segment it into some words by greedy algorithm.*/

Step 1. Computing weights of all the sub-strings in s.

Step 2. Pick up the string with the greatest weight, say s, which is the current abstracted word and store it. If s equals to s ,then exit, otherwise go to step 3.

Step 3. Segment(s-s),s-s is the left strings.

}

Algorithm Weighting(S)

{

/*Given a input string S= $c_1\ c_2 \cdots c_n$, in order to compute weights of all strings in s, it is necessary to build a collocation matrix A, where A(i, j) represents the frequency of string from character i to j, then the weight of that string can be calculated by the weighting function $W(c_i\ c_{i+1} \cdots c_j) = F(c_i\ c_{i+1} \cdots c_j) \times (j-i+1)^c$, where $F(c_i\ c_{i+1} \cdots c_j)$ is the frequency of string $c_i\ c_{i+1} \cdots c_j$, and its length is (j-i+1), c is a constant power of length, c>1. In practice when c equals 3, the segmented words are more probable to be correct;*/

Step 1. Search the data base to find out NHN set T_i of each c_i, i= 1,..n.

Step 2. For (j=1;j<=n-1;j++)
Step 3. For (i=1;i<=j-1;i++){
Step 4. $T_{ij}= T_{i,j-1} \wedge T_j$;// to compute collocation of column j in matrix A
Step 5. A(i,j-1) = $W(c_i\, c_{i+1} \cdots c_j)$ = $|T_{i,j-1}| \times (j-i)^c$; //to weight $c_i\, c_{i+1} \cdots c_{j-1}$.
}

where, T_{ij} is the NHN set of $c_i\, c_{i+1} \cdots c_j$, $T_{ij} = T_{i,j-1} \wedge T_j = ((((T_i \wedge T_{i+1}) \wedge T_{i+2} \wedge ...) \wedge T_j$ "$\wedge$" means collocation computation. For example,
let $(pp_1, pa_1, sn_1, ss_1, wd_1, ct_1) \in T_i$, $(pp_2, pa_2, sn_2, ss_2, wd_2, ct_2) \in T_{i+1}$, if$((pp_1 = pp_2)and(pa_1 = pa_2)and(sn_1 = sn_2)and(ss_1 = ss_2)and(wd_1 = wd_2)$ and$(ct_1+1 = ct_2))$, then it means that c_i and c_{i+1} collocate once with c_i appearing to the left side of c_{i+1}, let $(pp_2, pa_2, sn_2, ss_2, wd_2, ct_2) \in T_i \wedge T_{i+1} \cdot = T_{i,i+1}$

In practice, multiple segmentation technique is utilized. The first scanning of segmentation is to identify one character words , besides the numeric words(such as 1,2,三) and count unit words(such as 個) are also preprocessed by utilizing a numeric word list and a unit word list as well as matching based segmentation. Therefore in the first segmentation, the dictionary based approach is used based on a very small function word dictionary instead of a very big and complicated dictionary.

The segmentation on the second time is continue to process the unidentified strings based on the computation of the string frequency within context to find the frequently occurred(e. g. 2 or 3 times occurred) unknown words and solving ambiguous segmentation.

On the third time, segmentation is based on the string frequencies calculated in corpus to segment common words and make some low frequency unknown words isolated because, generally, the two words on the left and right of the unknown word is most likely to be common words which can be identified.

5. Text Summarization

5.1 Key Words Identification

A more efficient word weighting function is developed which is based on word frequency and word length, where the word frequency is refer to the frequencies calculated both in context and corpus, because key words is context related, the importance of context information and its utilization should be studied, besides the word length information is highlighted in weighting words, because key word is generally proper nouns which have a longer length than function words and other unimportant shorter content words, therefore longer words should be assigned higher weight, however pure frequency based method can not do that. The word weighting function is designed as follows:

$$T(w) = \frac{F_1(w)}{F_2(w)} \times L(w)^c$$

where $F_1(w)$ is the frequency of w in context, $F_2(w)$ is the frequency of w in corpus, $L(w)$ is the length of w , c is a constant power of length, in practice c=3.

5.2 Sentence Weighting Function

The important sentences generally illustrate themes or topics of contexts in a condensed and conclusive way, such as title and subtitle sentences, topic sentences and other conclusive sentences. The characteristics of important sentences are generally to contain more important words (or key words) and have a shorter sentence length and few number of sub-sentences. Therefore, the sentence weighting function is

designed as follows:

$$P(s)=\frac{T(w_1)+T(w_2)+\cdots+T(w_n)}{K\times L(s)\times N(s)}$$

where s is a sentence and w_i is a word of s, $T(w_i)$, i=1,..n, is the weight of w_i, $L(s)$ is the length of sentence s, $N(s)$ is the number of sub-sentences in s.

According to this function the shorter sentences with more important words will be given higher weight, so title and subtitle sentences, topic sentences and most of the conclusive sentences will have more chance to obtain higher weights than the other unimportant sentences.

Traditional approaches often use sentence position such as first and last sentence of a paragraph will statically assign a higher weight, however this assuption is not always true and salient sentences will often appeared in the middle of the paragraph and will not properly weighted. Besides, surface cues such as conjunctoins are often used to identify important sentence as well as rhetorical structures or referential links, which is powerful in analyzing inter- and intro sentence relations and guiding the salient sentence selection. However, the lack of /or superficial/or wrongly used surface cues often effect or mislead the rhetorical analysis, it should be noted that different styles of text such as literiture, editorial, technical paper, poetry and so on rely on and use surface cues in different weight and way.

There are also other interesting factors such as digital numbers(to answer how many/how much), time words(to answer when) unknown words such as name of people, organizations(to answer who/whom), place(to answer where) and this kind of factors is often user- oriented and text style related, and if used properly, it will make good effect. For example, if the user not only concern about the central content of the text but also concern the time or people involved, then the time and name of people should be designed to give a higher weight so that the sentences contain those words will have chance to be assigned higher weight.

5.3 Abstract Generation and Evaluation

The abstracts(or summaries) are generated by selecting the important sentences with higher sentence weight from the input text, and keeping their original sequential orders in the text.

It is not a settled problem on evaluating the quality of summarization. There are several criterias including recall, precision, brevity and ease in general, where recall means the percentage of central concepts or important sentences captured by an automatically produced summary, precision refers to the percentage of relevent concepts or sentences in this summary. Besides the above four criterias, there are other useful criterias such as Wh(when, where, what, why)&how(how, how many/how much), those criteria is useful when user are particularly interested in those contents or they acturally are the central contents of the text.

5. Experiments

5.1 Word Segmentation

The corpus used for word segmentation is a collection of texts without restrictions on domain, style and length.

A segmentation tests under 290,000-character sized textual corpus shows that the correctness rate of identified words ,denoted as PI, which is calculated by $PI=1-\frac{b}{w}$, where b is the number of wrong segmented words and w is the number of total words, is about 98% .

An comparison of 30000 words dictionary based Backwards Maximum Matching(BMM) algorithm and our textual corpus based segmentation algorithm,

shows that the number of wrong segmented words of BMM is 4 times of that of our method for open test.

5.2 Text Summarization

We have tested more than 50 texts in different domains and styles including editorials, technical papers, literiture papers, pose and news etc.. The results show that the better summarization results are generally domain unconstrained but influnced by writing styles. For example, if repitition is the main techniques to express emphasis and central concepts, then those kind of texts including editorials, technical papers and news will summarized properly, while pose or literiture texts will have worse summarization results.

6. Conclusion

we have illustrated the detailed algorithms of Chinese Text summarization system and shown some experimental results. The key techniques include the NHN(Natural Hierarchical Network) raw corpus(and text) indexing method, the word length and frequency based word weighting function and word segmentation algorithm. Keyword identification and salient sentence determination. The algorithms and ideas of our system are also quite meaningful for Korean and Japanese unknown words identification, and have been applied in Corpus-based Chinese -Korean Machine Translation(Li and Choi,1997a), Chinese Automatic Abstracting System(Li, 1995), Chinese-Korean Automatic Translation System(Li and Choi,1997b).

Acknowledgment

This research was supported by China National High -Tech Development Plan 863 and Center for Artificial Intelligence Research in Korea Advanced Institute of Science and Technology.

References

Jun-Jie Li and Kai-Zhu Wang, "Study and Implementation of Non-dictionary Chinese Segmentation," in NLPRS'95, Seoul, Korea, Dec.,1995, pp.266-271.

Hsin-Hsi Chen and Jen-Chang Lee, " Identification and Classification of Proper Nouns in Chinese Texts," in COLING'96, Aug., 1996, Vol.1 pp.222-229.

Jun-Jie Li and Key-Sun Choi, "Design and Implementation of An Example-Based Chinese-Korean Machine Translation System," in ICCPOL'97, Apr. 1997,Hong Kong.

Jun-Jie Li and Key-Sun Choi, "Corpus-Based Chinese-Korean Abstracting Translation System," in IJCAI'97, Nagoya, Japan, Aug. 1997.

Hideo Watanabe, " A Method for Abstracting Newspaper Articles by Using Surface Clues," in COLING96:947-979, Copenhagen, Denmark, Aug. 5-9, 1996.

Klaus Zechner, "Fast Generation of Abstracts from General Domain Text Corpora By Extracting Relevant Sentences," in COLING96:986-989, Copenhagen, Denmark, Aug.5-9, 1996.

Luhn,H.P., "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, Vol.2, No.2:159-165.

G. Dejong, "Prediction and Substantiation : Two Processes That Comprise Understanding," in Proceedings of IJCAI-79.

J.I. Tait, "Generating Summaries Using a Script-Based Language Analyzer," in Progress of Artificial Intelligence, 1985.

Danilo Fum. et al., " Forward and Backward Reasoning in Automatic Abstracting," COLING82.

Ono et al., "Abstract Generation Based on Rhetorical Structure Extraction," in COLING94, Vol.1:344-348.