Topic Segmentation for Short Texts

Tao-Hsing Chang
Department of Computer and Information Science
National Chiao Tung University
1001 Ta Hseuh Rd.,
Hsinchu, Taiwan 300, R.O.C
thchang@cis.nctu.edu.tw

Chia-Honag Lee
Department of Computer and Information Science
National Chiao Tung University
1001 Ta Hseuh Rd.,
Hsinchu, Taiwan 300, R.O.C
chl@cis.nctu.edu.tw

Abstract

Topic segmentation, which aims to find the boundaries between topic blocks in a text, is an important task for semantic analysis of texts. Although different solutions have been proposed for the task, many limitations and difficulties exist in the approaches. In particular most of the methods do not work well for such case as short texts, internet news and student's writings. In this paper, we focus on the short texts and present a method for topic segmentation. It can overcome the limitations in previous works. In preliminary experiments, the method show the accuracy of topic segmentation is increased effectively.

1. Introduction

Topic segmentation, which aims to find the boundaries between topic blocks in a text, is an important task for semantic analysis of texts. In general, they rely on such knowledge as word recurrence, collocations, thesaurus, linguistics cues, or the combination of those, to measure the similarity between sentences, and estimate whether topic shift occurs (Ferret(2002)). The studies are also classified into two schemes: one is based on locally comparing adjacent blocks of sentences, while the other method is based on considering topic block boundary decision as a global optimization (Bigi et.al(1998)).

Although the approaches for topic segmentation have been worked well for long texts, the assumptions it requires limit most of useful applications. First, before topic segmentation, the thresholds, coefficients, or parameters of formulas in some methods must be estimated beforehand depend on the characters of text sets or the experience of users. It not only limits applications to be fully automatic but also causes difficulty in many domains.

These methods perform especially poor when they are applied to such short texts as internet news and student's writings (Ponte and Croft(1997)). Since keywords from sentences are quite few and not reliable for short texts, the errors in measure naturally occur more frequently. It results in rapid decrease of the accuracy of topic segmentation.

The accuracy of these methods not only relies on the characteristics of the short texts, but also depends on the languages. For Chinese, the usage of punctuation marks such as comma mark is often ambiguous (Chen(1993)). For instance, the sentence ending with comma mark is not always considered a complete sentence on syntax or semantics. For incomplete sentences, the step of extracting reliable keywords is far more difficult. This causes, due to the characteristic of the language, further decrease of the accuracy of topic segmentation.

Apparently, short texts are troublesome and quite difficult for existing methods in topic segmentation. However, given a theme, a huge collection of reference texts for short texts can easily be obtained. This large corpus or thesaurus allows us to overcome the above mentioned limitations and define similarity between sentences. This paper will present a new method to segment topics for short texts. Section 2 reviews previous studies for text segmentation. Section 3 discusses the proposed technique in detail. Section 4 shows the performance of the method on some experiments. Section 5

discusses the conclusion.

2. Previous Work

There are two features in short texts processed in this paper. First, the content of the texts is complete but its length is short. That is, topic blocks in texts are sometimes composed of few sentences or one sentence. Secondly, the sentences in topic blocks may be incomplete on syntax or semantics. Though many of previous methods have been worked well, there are somewhat restrictions in the approaches applied to segment short texts. We will discuss detail below.

In earlier research, the topic blocks must be composed of several punctual paragraphs. Hearst(1997) introduces the *TextTiling* algorithm that measures the similarities between paragraphs with word sets of neighbor paragraphs, and then uses the similarities and known paragraph boundaries to place the position of topic shifts. It is useful to the lengthy texts in which article structure is punctual. Similarly, Salton et. al(1996) measures the similarity between paragraphs with term weight methods such as tf-idf weight, and links two paragraphs together when the similarity between them exceeds a threshold. Then, all links among paragraphs form a paragraph relation network, called *text relationship map*, and the sets of the paragraphs linked serially imply the same topic. The major difficulty of the method is that the principle for stating threshold is inexact. Moreover, the method applied to sentences does not work well because there are not many terms in the sentences. Nevertheless, it points out an interesting observation: that topic relationship not only exists in neighboring sentences but also in sentences in a topic block. In our studies, this idea will be used to develop a scheme to tolerate measuring errors.

In recent years, some of the approaches focus on detecting the occurrence of the topic shifts in sentence level. Beefman et. al(1999) develops a feature-based method with an exponential model to detect the places of topic shift in sentence. It uses cue-words feature, one of two features in the method, to detect occurrences of specific words that frequently occurs in the topic blocks boundaries. In our domain, such approach is often not applicable since cue-words may not appear in the few sentences in a topic block. Alternative approaches such focus on topic detection and tracking (TDT), which aims to process huge amounts of information as newswire. Hidden Markov model (HMM) (e.g. Blei and Moreno(2001), Yamron et. al(1998)) is a major technique for topic segmentation in TDT group. Ferret(2002) also presents and implements a method, called TOPICOLL, which combines word repetition and the lexical cohesion in texts for segmentation and link detection. The system processes texts linearly. However, it involves a delay state before deciding whether or not the current processing segment really ends. Nevertheless, the delay state is only effective for the tolerance of the similarity errors among several serial words.

Ponte and Croft(1997) presents a three-step method. It first uses lexical context analysis (LCA) to extract semantically related words in a sentence, and compute the number of related words in common between sentences. Secondly, it scores each possible topic block with both internal-external scoring model and Gaussian length model. Finally, it determines the position of the topic shifts by dynamic programming with the score of each possible topic blocks. In contrast to the previous methods mentioned earlier, it is a global optimization method. The method claims that it works well in spite of small segments with few common words. However, it must properly estimate the maximum segment size before segmentation by dynamic programming, and obtain the parameters of length model from a training set. When the variance of the size of topic blocks is large, the estimation of parameters is usually not reliable for the method. Furthermore, it is also unreliable to rely only on counting the numbers of related words in common between sentences to measure similarity.

Based on the observations, we develop a new approach which is both efficient for detecting topic shift in short texts and effective above for fault tolerance of the similarities between sentences. Moreover, it does not need any thresholds or parameters used by other methods.

3. Methodology

The method in this paper is divided into four steps: retrieving and expanding keywords, measuring similarity between sentences, scoring candidate topic blocks, and ranking segment sequences. In

retrieving and expanding keywords and similarity measure between sentences stages, the previous methods are usually classified into domain-specific methods, domain-independent methods, or hybrid methods according to the properties of their resource and application. Based on the properties of short texts as mention earlier, the method in this paper is designed for a domain-specific method in retrieving and expanding keywords stage. Moreover, the method determines the positions of topic shifts with global optimization. We discuss the detail below.

Previous studies often use a set of nouns or noun terms to represent both a sentence and implied bearings of a topic. This assumption is not effective or sufficient for short texts. Since a topic is usually changed with just a few sentences and the nouns in the sentence are either replaced with pronouns or even disappear. To overcome this difficulty, Ponte and Croft(1997), and Xu and Croft(1996) propose to expand the set of nouns extracted from a sentence to include other nouns for representation of a sentence. The expansion method is based on technique of query expansion in information extraction.

However, all these efforts and improvements are still not sufficient for many such other applications as student writings and internet pages. In our method, we define the original keywords of a sentence as the set of noun, verb, adjective words and phrases in the sentence based on the characteristics of the short texts. Furthermore, the original keywords will be extended the following method to include other terms and become a large set of keywords called expansion keywords.

3.1 Retrieving and expanding keywords

In this subsection, the step for expanding keywords is based on the concepts of co-occurrence and distance of two keywords in a passage which is composed of a fixed number of serial sentences. The concept of co-occurrence (Baeza-Yates and Ribeiro-Neto(1999), Ponte and Croft(1997), Xu and Croft(1996)) of two keywords in the same passage has been used to imply the possible existence of the same topic. We note that the frequency of the co-occurrence is an important cue of determining the degree of correlation between two keywords. Furthermore, the degree of correlation of two keywords also depends on their distance in a passage (Baeza-Yates and Ribeiro-Neto(1999)). We will define the distance as the number of the sentences between two keywords.

Fig. 1 shows an example for the concepts of co-occurrence and distance of two keywords. The distance between keyword *k2* and *k3* is 1, the distance between keyword *k3* and *k4* is 3. Because the length of the passage in Fig. 1 is 3, keyword *k2*, *k3*, and *k4* are co-occurrence words in the passage, but keyword *k1* and *k4* are not.

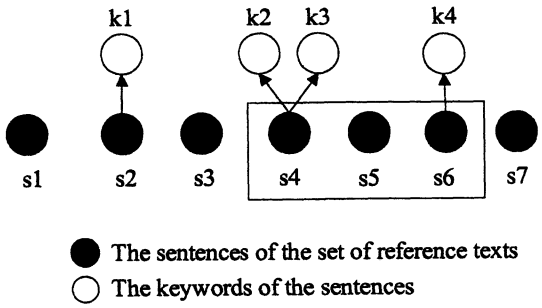

Fig. 1 An example for computing the relationship between the keywords

Based on the concepts of both co-occurrence and distance, our method will compute a matrix which describes the correlations among keywords. First, all keywords are retrieved from the set of reference texts. If the number of the keywords is n, then the correlation matrix R is a $n \times n$ matrix. The element $r_{i,j}$ of matrix R represents the correlation between keyword i and j, and is computed as following:

$$r_{i,j} = \sum_{t \in T} \sum_{p \in t} occ(i, j), \tag{1}$$

where t is the text in the set T of reference texts, p is the passage in the text t, and $occ(i,j)$ is computed

as following:

$$occ(i,j)=\begin{cases}\dfrac{1}{dist(i,j)}, & \text{when the keyword } i \text{ and } j \text{ both exist in passage } p.\\ 0, & \text{otherwise}\end{cases}, \quad (2)$$

where $dist(k_i,k_j)$ is the distance between keywords k_i and k_j in passage p.

Finally, the correlations are normalized according to formula (3):

$$normalized\ r_{i,j}=\frac{r_{i,j}}{r_{i,i}}. \quad (3)$$

Now, all correlations among keywords ranges from 0 to 1, and the correlation between a keyword and itself is 1. Usually, there are tremendous numbers of keywords in the set of reference texts. In order to increase processing efficiency, only a few highly correlated for each keyword is retained, the remaining ones are discarded. In contrast to LCA (Ponte and Croft(1997), Xu and Croft(1996)), our method uses the distance between the keywords and the frequency of co-occurrence to evaluate the correlation between keywords.

3.2 Measuring similarity between sentences

The expansion keywords in a sentence can represent the topic of the sentence. By retrieving the common expansion keywords between the sentences, we can measure the similarity between the sentences. The similarity between sentences m and n is defined below:

$$sim(m,n)=\frac{\sum_i\sum_j\sum_e r_{i,e}\cdot r_{j,e}}{|K(m)|\cdot|K(n)|\cdot SentDist(m,n)},\quad i\in K(m), j\in K(n), e\in E(i)\cap E(j), \quad (4)$$

where $K(m)$ and $K(n)$ is the set of the original keywords of the sentence m and n respectively, $E(i)$ and $E(j)$ is the set of the expansion words of the keyword i and j respectively, $|K(m)|$ and $|K(n)|$ is the number of the original keywords of sentence m and n respectively, and $SentDist(m,n)$ represents the distance between sentence m and n. For instance, the *SentDist* of the first sentence and the third sentence is equal to 1.

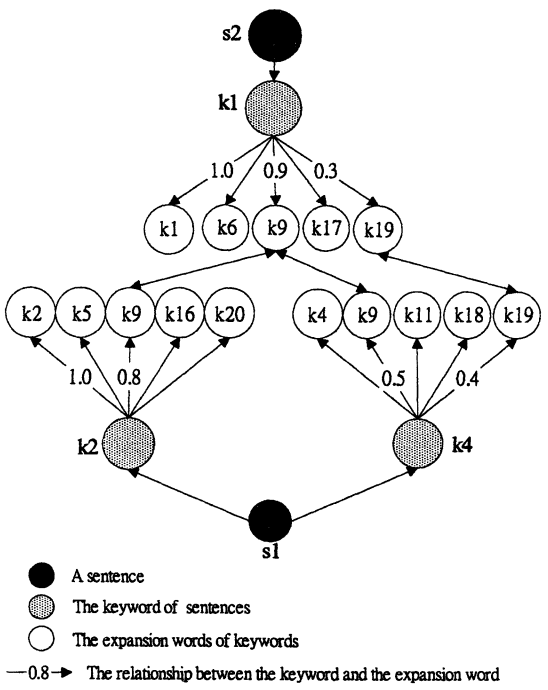

Fig. 2 An example of measuring similarity between sentences

Fig. 2 is an example for computing similarity between sentences. Both k2 and k4 are original keywords retrieved from sentence *s1*, and k1 is the original keyword retrieved from sentence *s2*. Moreover, *k1*, *k6*, *k9*, *k17*, and *k19* are the expansion keywords of *k1*, while *k2*, *k5*, *k9*, *k16*, and *k20* are the expansion keywords of *k2*, and so on. Using formula (4), the similarity between sentence *s1* and *s2* is equal to 0.645.

3.3 Scoring candidate topic blocks

Table 1 shows an example of the similarities among all sentences based on the above computations for a text consisting of eight sentences. In Table 1, there are various different ways of segmenting the text into topic blocks. For example, {S1,S2}, {S3,S4,S5,S6}, and {S7,S8} represents one segmentation, while {S1,S2,S3,S4} and {S5,S6,S7,S8} represents another segmentation. Such the topic blocks as {S1, S2} and {S5, S6, S7, S8} are called candidate topic blocks, and the sentences S1 and S2 are called the member sentences of the candidate topic block {S1, S2}. Using the table of similarities among sentences, the method assigns each candidate topic block a score.

Table 1. An example of the similarities among all sentences.

	S1	S2	S3	S4	S5	S6	S7	S8
S1	-	0.7	0.0	0.1	0.0	0.2	0.1	0.0
S2	0.7	-	0.1	0.1	0.0	0.1	0.2	0.0
S3	0.0	0.1	-		0.0	0.3	0.0	0.2
S4	0.1	0.1	0.8	-	0.0	0.6	0.1	0.3
S5	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0
S6	0.2	0.1	0.3	0.6	0.0	-	0.1	0.2
S7	0.1	0.2	0.0	0.1	0.0	0.1	-	0.9
S8	0.0	0.0	0.2	0.3	0.0	0.2	0.9	-

The score of a candidate topic block is the difference between internal and external score. The internal score of a candidate topic block is defined as the subtraction of internal variation term from internal cohesion term. Internal cohesion term is computed as

$$\frac{1}{2\|T\|}\sum_{i\in T}(\max_{k\in T} sim(i,k) + \max_{j<i,\, m>i} sim(j,m)), \tag{5}$$

where the first term represents the maximum similarity of sentence i with other sentences in block T; and the second term represents the maximum similarities of the preceding sentences with the succeeding sentences of i^{th} sentence. The second term, called tolerating term, is needed to smooth out the noise of the similarity measurements which occur quite often in short texts. Fig. 3 shows an example for computing internal cohesion score. Fig. 3(a) and (b) respectively represents the first term and second term in formula (5).

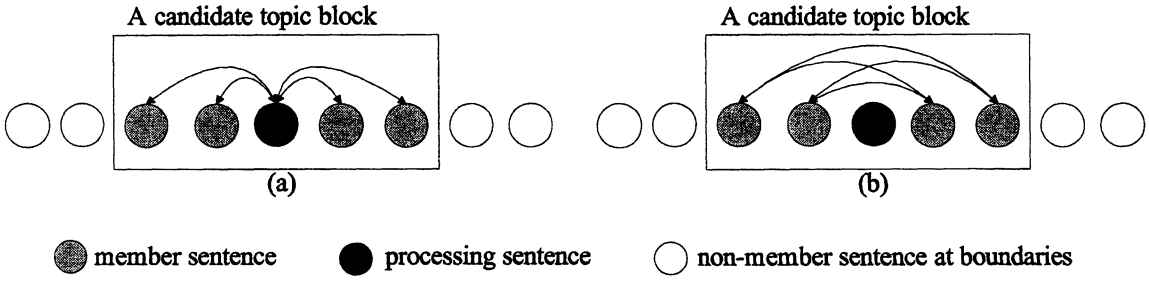

Fig. 3 An example for computing internal cohesion score.

The internal variation of a candidate topic block is computed by the following formula:

$$\frac{1}{\|T\|}\sum_{i,j\in T}\max_{\substack{k\in T\\ k\neq i,j}}\left|sim(i,k)-sim(j,k)\right| \tag{6}$$

The formula, in fact, is to compute the average of the difference between any two sentences in the topic block T. The difference will be zero if both sentences have exact same similarities with all other sentences in the block T.

The external score of a candidate topic block is computed as

$$\frac{1}{2}\max_{\substack{i\in LE\\ j\in LB}} sim(i,j)+\frac{1}{2}\max_{\substack{i\in RE\\ j\in RB}} sim(i,j), \tag{7}$$

where LB represents the set of left-side member sentences; LE represents the set of non-member sentences to the left of the block; RB represents the set of right-side member sentences; RE represents the set of non-member sentences to the right of the block.

When the external score is large, the boundary of the candidate topic block is not in place and obviously not a well chosen position. It is also used in Ponte and Croft(1997) and designed to be as a penalty from internal score.

3.4 Ranking Segment Sequences

A text can be randomly segmented into several serial candidate topic blocks, so a text can be treated as a segment sequence which is composed of these candidate topic blocks. Using the scoring approach as mentioned above subsection, the method can score all candidate topic blocks of a segment sequence and sum up the score of them to be the score of the segment sequence. The high-score segment sequence represents that there are high cohesion and low repulsion among its candidate topic blocks, and the low-score segment sequence represents that there are low cohesion or high repulsion among its candidate topic blocks. Because there are many possible segment sequences in a text, the method chooses the highest-score segment sequence as authentic segment sequence for the text. Using dynamic programming and recursive function can find out the best one from many segment sequences.

4. Preliminary Experiments

Many studies use different corpus or thesaurus to be the data set of experiment based on the study domain. In this paper, we use the set of student's writings to be the data set of the experiments because it is fit the properties of short texts as mentioned earlier. We collect 1285 writings from the eighth graders for theme "Recess at School". Randomly choose 1166 writings as a training set, and leave the remaining 119 writings as a test set. On average, each writing consists of 376 Chinese characters, 31 sentences and 7 topic shifts. The writings are segmented into topics using the methods in this paper and Ponte and Croft(1997), respectively. Moreover, human examines whether the topic shifts located by the method in the segmentations should occur.

The topic shifts in the segmentations can be classified into *hits*, *moves*, *insertions* and *deletions* (Ponte and Croft(1997)). An *insertion* is the method generates a shift that does not line up with real shift. A *deletion* is when a real shift exists but the method does not generate a shift. A *move* is two shifts line up but are not in the same position. In the experiment, both hits and moves shifts are called "exact match", and the sum of hits, moves, and deletions shifts are called "partial match".

Table 2 displays the result of examining segmentations. In Table 2, the exact match in the proposed method is 24%-29% higher than Ponte and Croft's method, and the partial match is 6%-20% higher than Ponte and Croft's method. This large improvement of the performance of our method over Ponte and Croft's method is obviously due to the reduction of the errors in the insertions.

Furthermore, Table 2 also shows how the tolerating term in internal score of the proposed method affects its accuracy. In Table 2, the exact match in the method with tolerating term are 8% higher than that without the term, and the insertions generated in the method with tolerating term are 5% lower than that without the term. Hence the tolerating term effectively reduces the wrong topic shifts generated by noise sentence. Based on the experiments, it indicates that the proposed method can

measure the similarities among the sentences more accurately and therefore segment short texts into topic segments.

Table 2. Results for using the method in this paper and Ponte & Croft's method.

	Exact Match	Partial Match	The Ratio of Insertions
The proposed method with tolerating term	0.62	0.74	0.26
The proposed method without tolerating term	0.54	0.69	0.31
Ponte & Croft's method with maximum size 5	0.38	0.54	0.46
Ponte & Croft's method with maximum size 10	0.33	0.68	0.32

5. Conclusions

In this paper, we present a new method for topic segmentation. It has two major advantages over traditional methods. First, it can measure similarities more accurately among sentences even though there are few keywords in the sentences. Secondly, it can tolerate noise sentences in topic segment. The proposed method are specifically designed and developed for short texts. In addition, it can be fully automated without human interaction and preprocessing. In our preliminary experiments, the method shows the accuracy of topic segmentation is increased effectively.

References

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern information retrieval.* ACM Press, New York.

D. Beeferman et. al. 1999. Statistical models for text segmentation. *Machine Learning*, 34:177-210.

B. Bigi et. al. 1998. Detecting topic shifts using a cache memory. *Proceedings of Fifth International Conference on Spoken Language Processing*, 2331-2334.

D. M. Blei and P. J. Moreno. 2001. Topic segmentation with an aspect hidden Markov model. *Proceedings of SIGIR '01*, 343-348.

H. H. Chen. 1993. A language model for parsing very long Chinese sentences. *Proceedings of 1993 IEEE International Conference on Tools with AI*, 290-297.

O. Ferret. 2002. Using collocations for topic segmentation and link detection. *Proceedings of ACL-COLING '02.*

M. A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics,* 23(1):33-64.

J. M. Ponte and W. B Croft. 1997. Text segmentation by topic. *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, 120-129.

G. Salton et. al. 1996. Automatic text decomposition and structuring. *Information Processing and Management,* 32(2):127-138.

J. P. Yamron et. al. 1998. A hidden Markov model approach to text segmentation and event tracking. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:333-336.

M. Utiyama and H. Isahara. 2001. A statistical model for domain-independent text segmentation. *Proceedings of ACL'01*, 491-498.

J. Xu and W. B. Croft. 1996. Query expansion using local and global document analysis. *Proceedings of SIGIR'96*, 4-11.