Subtopic Segmentation for Small Corpus Using a Novel Fuzzy Model

Tao-Hsing Chang and Chia-Hoang Lee

***Abstract*—Subtopic segmentation is a critical task in numerous applications, including information retrieval, automatic summarization, essay scoring, and others. Although several approaches have been developed, many are ineffective for specific domains with a small corpus because of the fuzziness of the semantics of words and sentences in the corpus. This paper explores the problem of subtopic segmentation by proposing a fuzzy model for the semantics of both words and sentences. The model has three characteristics. First, it can deal with the uncertainty in the semantics of words and sentences. Secondly, it can measure the fuzzy similarity between the fuzzy semantics of sentences. Thirdly, it can develop a fuzzy algorithm for segmenting a text into several subtopic segments. The experiments, especially for a short text with a small corpus in a specific domain, indicate that the method can efficiently increase the accuracy of subtopic segmentation over previous methods.**

***Index Terms*—Fuzzy modeling, fuzzy semantics, semantic similarity measurement, small corpus, topic segmentation.**

I. Introduction

SUBTOPIC segmentation is a fundamental task in numerous applications, including information retrieval [1], automatic summarization [2], [3], essay scoring [4], and others. As a vast amount of short documents have become available on the Internet, and in such media as broadcasting and newspapers, the need to find effective and fast segmentation approaches has become more urgent. Topic segmentation methods often first extract the semantics of both words and sentences and secondly measure similarities among the semantics of sentences; finally, they segment the text into subtopic blocks according to the similarity.

The performance of conventional approaches of subtopic segmentation with a small corpus is often undermined by the errors in computing the semantics of words and similarities among sentences. Although the semantics of words can be reliably determined using a universal corpus with many documents, semantics computed from the general corpus are neither adequate nor even useful in other specific domains. The difficulty arises from the fact that the semantics of a word often vary with the domain. Additionally, the computation of word semantics from a small corpus in a specific domain is often unreliable because insufficient data are available to compute the numbers of the co-occurrence of words in the corpus. Several studies have noted this dilemma [4]–[7] and developed techniques to eliminate errors and improve performance. The improvement provided by these methods, however, remains very limited because they try to increase accuracy without taking into account such other issues as fuzziness or the reliability of semantic association.

Fuzzy set theory [8] is often used to deal with uncertainties or fuzziness in linguistic terms. A fuzzy set can describe various possible values in a domain, so this paper explores the problem of subtopic segmentation by proposing a fuzzy model for representing the semantics of both words and sentences, for representing the degrees of semantic association with the domain of the fuzzy set, and for representing the uncertainty in the degree using the membership function. It comprises three steps. First, it uses a multidimensional fuzzy vector, in which each entry is a fuzzy set, to represent the semantics of a word and a sentence. Each membership in the fuzzy vector represents the uncertainty in the semantics. Secondly, it develops a fuzzy approach to measure the similarity and uncertainty between sentences. Thirdly, it designs a fuzzy algorithm to segment a text into several subtopic blocks.

The rest of this paper is organized as follows. Section II reviews some previous studies on the similarity of fuzzy semantics and studies on topic segmentation. Section III presents the system architecture and comprehensively surveys the proposed subtopic segmentation system. Sections IV and V discuss the proposed segmentation method in detail. Section VI presents the experimental results based on two corpora of essays. Section VII draws conclusions.

II. Related Work

Fuzzy set theory has been used to express the semantics of words [9]–[12]. Based on the assumption that semantic categories are fixed and known, these studies developed different fuzzy relations for the semantics of words in the known categories and present methods for measuring similarity for specific applications. The fuzzy approach is used herein because not all words necessarily fall in a single category. Takagi and Kawase [12] utilized conceptual fuzzy sets to specify the ambiguity of words and developed a conceptual matching algorithm for such applications as image retrieval and recommending television programs. Sun *et al.* [11] measured the similarity between sentences using a fuzzy semantic construction based on words and sentences, called information mass, and applied the measurement to a question-answering system. Akrivas *et al.* [9] described a query text in terms of a fuzzy set of semantic entities, generated using a fuzzy thesaurus, to retrieve useful information.

Manuscript received March 17, 2005; revised November 11, 2005 and February s1, 2006. This work was supported in part by the National Science Council, Taiwan, R.O.C., under Grant NSC-93-2213-E-009-069.

The authors are with the Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: thchang@cis.nctu.edu.tw; chl@cis.nctu.edu.tw).

Digital Object Identifier 10.1109/TFUZZ.2006.889911

Most approaches [13]–[26] for subtopic segmentation comprise three phases. First, the approaches extract such linguistic features from the corpus or external knowledge, as word repetition, co-occurrence of words, word frequency, and others. Secondly, the approaches measure similarities among features, sentences, or potential segments by such techniques as the cosine function, local context analysis, and others. Finally, the approaches determine the boundaries between topics, in a process called topic shift detection, using such techniques as valley scoring, dynamic programming, divisive clustering, and agglomerative clustering. For instance, Hearst [14] applied the cosine function to measure the similarity between adjacent blocks from the frequency of words contained in the blocks, and accordingly identified the boundaries of subtopics.

Recently, many studies [6], [20]–[26] have focused on the problem of segmenting such text streams as newswire feeds. Apart from the general steps described above, these studies also exploit various special features of the text streams, such as prosodic features, cue words, and references, to increase the accuracy of topic segmentation. Based on the detection of these features or the similarities among them, these studies have presented various models such as the hidden Markov model, the statistical model, decision trees, and the finite automaton to locate topic shifts.

All of the aforementioned approaches must exploit a large enough training corpus or knowledge bases to measure similarities with sufficient accuracy and precision. Approaches that rely on lexical cohesion neglect low-frequency words in a topical segment because the number of high-frequency words and phrases in the segment suffice and contain enough information for segmentation methods. However, for the domains of a short text with a small training corpus, the number of high-frequency words in a topical segment is often very lacking, and far less than the number of low-frequency words. When the above approaches are applied to the problem of a short text with a small training corpus, their performance is often greatly reduced, even when low-frequency words are used.

Some studies [4]–[7] have already noted that the uncertainty that arises in the computation of the similarity among sentences or words greatly degrades the performance of topic segmentation, to which several remedies have been proposed. Choi [5] proposed a sentence similarity matrix and employed a ranking scheme to increase the reliability. Ferret [6] included a delay state in finite automaton before determining whether the current processing segment ends. Chang and Lee [4] designed a fault tolerance term to smooth out the noise in the similarity measure. All of these approaches focus on refining the process of computing the similarity or evaluation of segmentation. None of these approaches works well for solving the problem of a short text with a small training corpus, because the uncertainty is magnified in the domain.

Ponte and Croft [7] stated that the computation of similarity is not robust because features are lacking in the problem of short text and the topic segments generally consist of only a few sentences. Local context analysis (LCA) has been used to increase the number of semantic features of sentences to overcome the issue [7]. First, LCA-based approaches treat each sentence as a query and obtain several words and phrases that are strongly related to the sentence, called concepts. Secondly, the similarity between two sentences is computed from the number of the concepts that are related to both of the sentences. Finally, the similarities are utilized to score individual segments of various sizes, and the sum of the scores of the segments in a potential segmentation represents the score of the potential segmentation. The segmentation with the highest score is chosen as the real segmentation. Therefore, the accuracy of segmentation for the problem of short text can be increased dramatically.

III. System Overview

This paper deals directly with the uncertainty based on modeling the semantics of words and sentences using a vector of fuzzy numbers. The computation of the similarity and the segmentation score is based on fuzzy arithmetic to address uncertainty.

Fig. 1 depicts the software architecture of the proposed subtopic segmentation. The system also consists of three main phases—training, estimating similarity, and determining segmentation. The training phase estimates the fuzzy semantics of every word based on a training corpus. The estimation of the similarity phase yields the fuzzy semantics of sentences from the fuzzy semantics of words, and then fuzzy similarities among sentences are estimated. The segmentation phase evaluates the scores of various candidate segmentations ranked using fuzzy similarities, dynamic programming, and defuzzification.

The first part of this work involves both the training phase and the estimation of the similarity. It presents novel approaches for modeling fuzzy semantics and estimating fuzzy similarities among sentences. The second part involves the evaluation of the segmentation phase, and develops a subtopic segmentation method using fuzzy models, based on some of the observations of Ponte and Croft [7].

The corpus and test document should undergo such preprocesses as word segmentation for oriental languages, POS tagging, and the filtration of stop words. These steps are not trivial, and their results influence the accuracy of segmentations. Interested readers should refer to [27]–[29], which have developed various preprocessing methods. The texts and corpus used herein are assumed to have been segmented and tagged.

IV. Fuzzy Model for Semantic Representation

Many studies have exploited a multidimensional vector or row matrix to represent the semantics of a word, a sentence, or a document. Each entry of the vector, representing the semantics of a word, consists of the frequency of the co-occurrence of the word with a reference word associated with the entry or dimension. For instance, if a corpus comprises 100 different reference words, the dimension of the vector would be of size 100, and each dimension is associated with one of the 100 reference words. For simplicity, in this paper, the ith word in the set of reference words is denoted as w_i, corresponding to dimension i.

The convention of using a multidimensional vector to represent semantics is adopted herein. However, each entry in the vector is a fuzzy set rather than a crisp value, as in other approaches. The set of reference words associated with the vector is limited to nouns, verbs, adjectives, and adverbs. Below, the

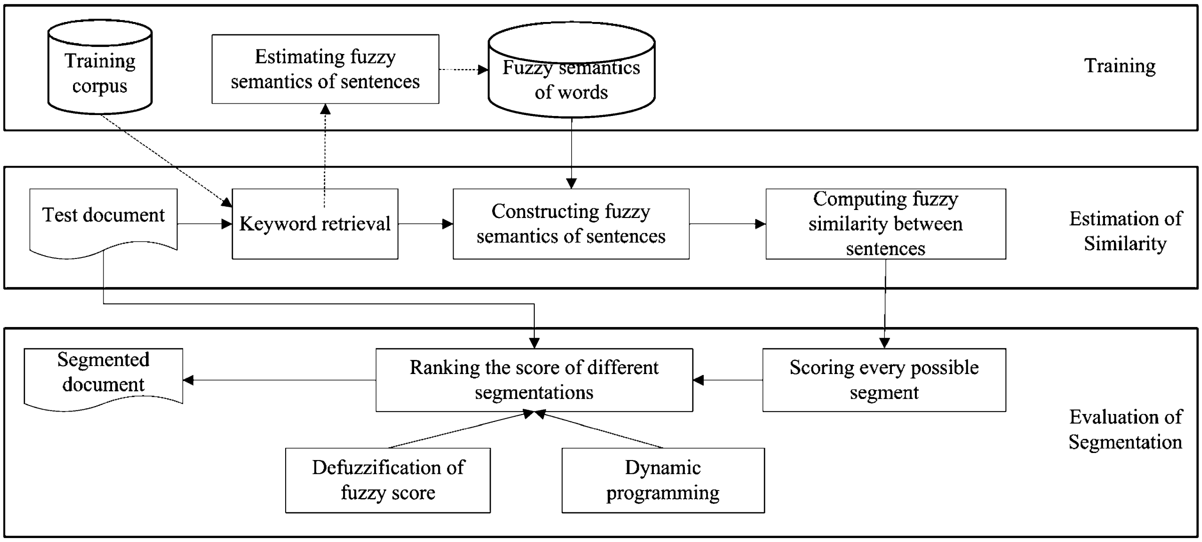
Training corpus
Estimating fuzzy semantics of sentences
Fuzzy semantics of words
Training
Test document
Keyword retrieval
Constructing fuzzy semantics of sentences
Computing fuzzy similarity between sentences
Estimation of Similarity
Segmented document
Ranking the score of different segmentations
Scoring every possible segment
Evaluation of Segmentation
Defuzzification of fuzzy score
Dynamic programming

Fig. 1. System architecture of the proposed subtopic segmentation method.

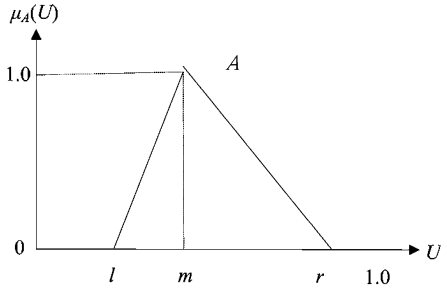

Fig. 2. Membership function for triangular fuzzy number A.

multidimensional fuzzy vector will be referred to as fuzzy semantics, in contrast to traditional word semantics. Assume there are n reference words in a corpus, and the semantics of word w_i in the corpus will be defined as

$$SW(w_i) = \langle M_{i,1}, M_{i,2}, \ldots, M_{i,j}, \ldots, M_{i,n} \rangle \qquad (1)$$

where $M_{i,j}$ represents the fuzzy set of the semantics of word w_i in the dimension associated with the reference word w_j.

This paper employs triangular fuzzy numbers to represent fuzzy sets. A triangular fuzzy number A can be parameterized by

$$A = (l, m, r) \qquad (2)$$

where $\mu_{\mathrm{A}}(m) = 1$ and l and r represent the left and right spreads, respectively. Fig. 2 plots the membership function for triangular fuzzy number A.

Given this model, topic segmentation is performed as three subtasks. First, a vector of fuzzy sets is generated for representing the fuzzy semantics of word; secondly, the fuzzy semantics of the words in the sentence are integrated to yield the fuzzy semantics of the sentence; and thirdly, the similarity between the fuzzy semantics of the two sentences is computed. The details are discussed below.

A. *Fuzzy Semantics of Words*

Assume there are n reference words in a corpus. The fuzzy semantics of a word w_i in the corpus can be obtained by substituting (2) into (1)

$$SW(w_i) = \langle (l_{i,1}, m_{i,1}, r_{i,1}), (l_{i,2}, m_{i,2}, r_{i,2}), \ldots, (l_{i,j}, m_{i,j}, r_{i,j}), \ldots, (l_{i,n}, m_{i,n}, r_{i,n}) \rangle. \qquad (3)$$

In (3), the $m_{i,j}$ of the fuzzy sets is defined as

$$m_{i,j} = \frac{\sum_{t \in T} \sum_{p \in t} \mathrm{occ}(w_i, w_j)}{\mathrm{freq}(w_i)} \qquad (4)$$

where $\mathrm{freq}(w_i)$ is the number of occurrences of word w_i in the training corpus; t is the text in the corpus T; p is a text segment in the text t; and $\mathrm{occ}(w_i, w_j)$ is shown in (5) at the bottom of the page, where $\mathrm{dist}(w_i, w_j)$ is the distance between words w_i and w_j in p.

$$\mathrm{occ}(w_i, w_j) = \begin{cases} \frac{1}{\mathrm{dist}(w_i, w_j)}, & \text{where the words } w_i \text{ and } w_j \text{both exist in } p \\ 0, & \text{otherwise} \end{cases} \qquad (5)$$

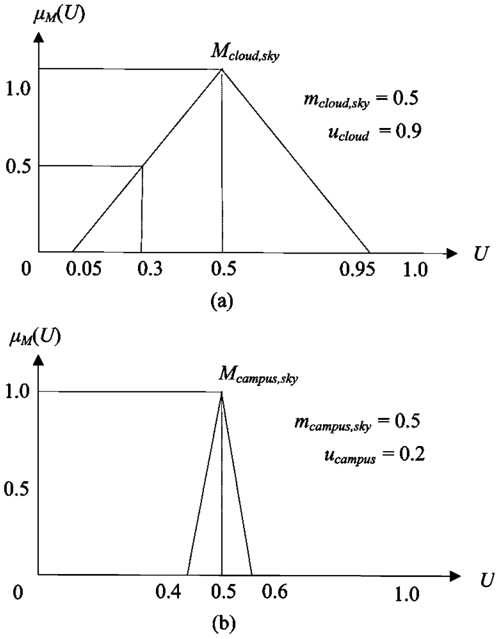

Fig. 3. Comparison of uncertainties for two words in the same dimension.

$l_{i,j}$ and $r_{i,j}$ in (3) are defined as

$$l_{i,j} = \begin{cases} u_i, & \text{if } (m_{i,j} + 0.5u_i) > 1 \\ \max(0, m_{i,j} - 0.5u_i), & \text{otherwise} \end{cases}$$
$$r_{i,j} = \begin{cases} 1 - u_i, & \text{if } (m_{i,j} - 0.5u_i) < 0 \\ \min(1, m_{i,j} + 0.5u_i), & \text{otherwise} \end{cases} \quad (6)$$

where

$$u_i = \max\left(0, 1 - \frac{\text{freq}(w_i)}{\text{constantC}}\right) \quad (7)$$

and C is a constant used to determine whether the semantics are uncertain. Restated, no uncertainty exists in the semantics if the number of occurrences of a word exceeds the threshold C, and the spread of the fuzzy number is zero.

In (6) and (7), parameter u_i is employed to represent the uncertainty of the semantics of word w_i. For instance, for simplicity, the distance between all words is assumed to be unity and constant C is 100. The number of occurrences of words "cloud," "sky," and "campus" are 10, 50, and 80, respectively. Moreover, the number of co-occurrences of "cloud" and "sky" is assumed to be five, and that of "campus" and "sky" to be 40. Fig. 3(a) and (b) shows the fuzzy numbers for the semantics of word "cloud" and "campus" in the dimension associated with "sky." In Fig. 3(a), u_{cloud} shows the probability that the semantic degree of "cloud" in the dimension associated with "sky" might be 0.9 because of few occurrences of the word "cloud." On the other hand, since the word "campus" occurs many times, u_{campus} states that 0.9 is impossible in the domain of the semantic degree of the word "campus" in the "sky" dimension. The spread of a fuzzy number can be used easily and clearly to model the uncertainty for the semantics of the words.

B. Fuzzy Semantics of Sentence

Following the definition of fuzzy semantics of a word, the same representation is used to define the fuzzy semantics of a sentence. However, the computation of building the representation is more complex, since a sentence usually contains several words, and each of which often contributes different degrees of the semantics. Equation (9) shows the computation of the semantics of a sentence from various words in the sentence.

The fuzzy semantics $SV(s)$ of sentence s is defined as

$$SV(s) = \langle H_{s,1}, H_{s,2}, \ldots, H_{s,j}, \ldots, H_{s,n} \rangle, \quad 1 \leqslant j \leqslant n \quad (8)$$

where

$$H_{s,j} = \left(\frac{\sum_{w_i \in K(s)} (1 - u_i) \times l_{i,j}}{\sum_{w_i \in K(s)} (1 - u_i)}, \frac{\sum_{w_i \in K(s)} (1 - u_i) \times m_{i,j}}{\sum_{w_i \in K(s)} (1 - u_i)}, \frac{\sum_{w_i \in K(s)} (1 - u_i) \times r_{i,j}}{\sum_{w_i \in K(s)} (1 - u_i)} \right) \quad (9)$$

where $K(s)$ represents the set of the words formed by intersecting the set of words in sentence s and the set of n reference words in a corpus; u_i can be computed from (7).

The rationale that underlies (9) can be briefly stated as follows. The fuzzy semantics of a sentence is defined as the union of the fuzzy semantics of the words in the sentence. Greater uncertainty of a word clearly corresponds to a lower weighting of its contribution to the semantics of the sentence.

C. Fuzzy Similarity Between Sentences

Most subtopic segmentation approaches segment the text into various subtopics, based on similarities among sentences. These approaches apply cosine functions to the semantics vectors of the sentences to measure their similarities, but none of them is applicable to fuzzy semantics. Fig. 4 shows two fuzzy semantics that correspond to two sentences whose horizontal axis represents various dimensions and whose vertical axis represents the spread of the fuzzy semantics in each dimension. A longer rectangle corresponds to a larger spread of the fuzzy number. The darker area represents the larger value of the membership function. An approach is developed below to measure the similarity between two fuzzy semantics.

The similarity between two fuzzy semantics consists of the degree of similarity and the uncertainty in the degree, so the similarity between sentences will be defined as a triangular fuzzy number. It is constructed using the following steps. Assume the fuzzy semantics of sentences s_a is $\langle H_{a,1}, H_{a,2}, \ldots, H_{a,n} \rangle$, while the fuzzy semantics of sentences s_b is $\langle H_{b,1}, H_{b,2}, \ldots, H_{b,n} \rangle$, where $H_{i,j} = (l_{i,j}, m_{i,j}, r_{i,j})$. Initially, the highest k fuzzy numbers from s_a and s_b are selected, respectively. Let F be the union of the corresponding dimensions of these k fuzzy numbers. Let the number of

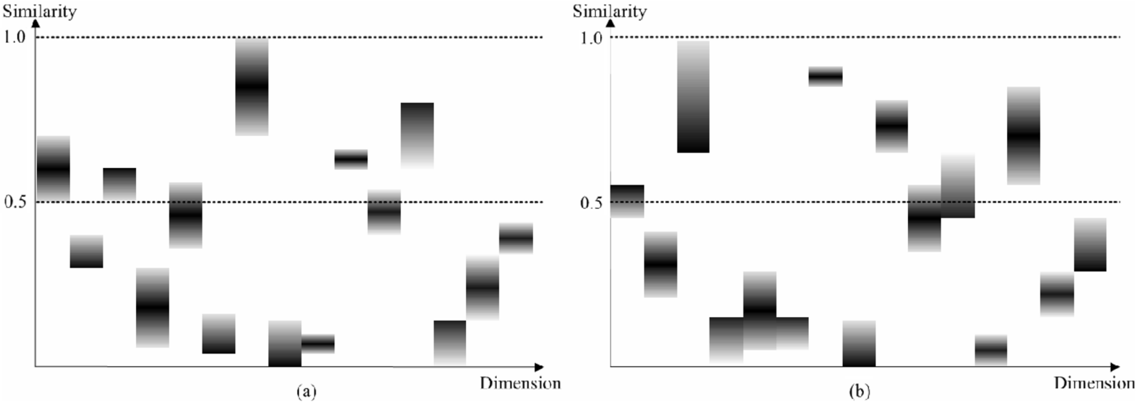

Fig. 4. Two fuzzy semantics corresponding to two sentences.

elements in F be $|F|$. Then, the fuzzy similarity $\mathrm{Sim}(s_a, s_b)$ between s_a and s_b is defined as follows:

$$\mathrm{Sim}(s_a, s_b) = (l_{\mathrm{sim}}, m_{\mathrm{sim}}, r_{\mathrm{sim}}) \tag{10}$$

where we have (11) as shown at the bottom of the page.

In (10), l_{sim} and r_{sim} specify the boundaries of the interval for similarity. Briefly, the similarity is represented by a fuzzy number and contains all possible similarities and their uncertainties. Using a fuzzy number to represent the similarity between sentences has two advantages. First, it can embrace uncertainty from measured similarity and improve the performance of subtopic segmentation. Secondly, the standard fuzzy arithmetic operation can be easily applied to triangular fuzzy numbers.

V. Fuzzy Model for Subtopic Segmentation

In this system, fuzzy similarities among sentences are computed using (10) and (11). Table I shows an example of the similarities among seven sentences in a document. There are various ways of grouping sentences into subtopics. For instance, $\{\mathrm{S1}, \mathrm{S2}, \mathrm{S3}\}, \{\mathrm{S4}, \mathrm{S5}\}$ and $\{\mathrm{S6}, \mathrm{S7}\}$ represents one segmentation, whereas $\{\mathrm{S1}, \mathrm{S2}, \mathrm{S3}, \mathrm{S4}\}$ and $\{\mathrm{S5}, \mathrm{S6}, \mathrm{S7}\}$ represents another segmentation. Such sets of sentences as $\{\mathrm{S1}, \mathrm{S2}, \mathrm{S3}, \mathrm{S4}\}$, and $\{\mathrm{S5}, \mathrm{S6}, \mathrm{S7}\}$ are called segments. A sentence is called a member of a segment if and only if the sentence is in the segment and called nonmembers if it is not a member. Additionally, if a set that comprises various segments can cover all of the sentences, then the set is called a candidate segmentation. For instance, the set that comprises segment $\{\mathrm{S1}, \mathrm{S2}, \mathrm{S3}\}, \{\mathrm{S4}, \mathrm{S5}\}$, and $\{\mathrm{S6}, \mathrm{S7}\}$ is regarded as a candidate segmentation.

The table of similarities among sentences is used in the proposed method to assign a score to each segment and to rank candidate segmentations by the scores of segments in candidate segmentation. In the scoring procedure, the following four fuzzy arithmetic operations are applied. Given triangular fuzzy numbers $A_1, A_2, \ldots, A_k, \ldots, A_n, A_k = (l_k, m_k, r_k), 1 \leqslant k \leqslant n$, the following simple fuzzy arithmetic operators are defined:

$$\text{Fuzzy addition}: A_i \oplus A_j = (l_i + l_j, m_i + m_j, r_i + r_j) \tag{12}$$

$$\text{Fuzzy subtraction}: A_i \ominus A_j = (l_i - r_j, m_i - m_j, r_i - l_j). \tag{13}$$

Dividing a fuzzy number by an integer

$$\frac{1}{n} A_i = \left(\frac{l_i}{n}, \frac{m_i}{n}, \frac{r_i}{n} \right). \tag{14}$$

The maximum of fuzzy numbers

$$\mathrm{fmax}(A_1, A_2, \ldots, A_n) = (\max(l_1, l_2, \ldots, l_n),$$

$$\begin{aligned}
l_{\mathrm{sim}} &= 1 - \frac{1}{|F|} \sum_{j \in F} \max(|r_{b,j} - l_{a,j}|, |r_{a,j} - l_{b,j})| \\
m_{\mathrm{sim}} &= 1 - \frac{1}{|F|} \sum_{j \in F} |m_{a,j} - m_{b,j}| \\
r_{\mathrm{sim}} &= 1 - \frac{1}{|F|} \sum_{j \in F} f(j) \\
f(j) &= \begin{cases} \min(|l_{a,j} - r_{b,j}|, |l_{b,j} - r_{a,j}|), & \text{if } l_{a,j} > r_{b,j} \text{ or } l_{b,j} > r_{a,j} \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{11}$$

TABLE I
EXAMPLE OF THE SIMILARITIES AMONG SEVEN SENTENCES

Sentence	S1	S2	S3	S4	S5	S6	S7
S1	-	(0.4,0.7,1.0)	(0.2,0.3,0.5)	(0.1,0.1,0.1)	(0.0,0.1,0.2)	(0.0,0.0,0.0)	(0.1,0.1,0.1)
S2	(0.4,0.7,1.0)	-	(0.1,0.4,0.5)	(0.0,0.2,0.2)	(0.1,0.2,0.3)	(0.0,0.0,0.0)	(0.2,0.3,0.3)
S3	(0.2,0.3,0.5)	(0.1,0.4,0.5)	-	(0.1,0.3,0.6)	(0.2,0.3,0.3)	(0.1,0.1,0.2)	(0.0,0.1,0.1)
S4	(0.1,0.1,0.1)	(0.0,0.2,0.2)	(0.1,0.3,0.6)	-	(0.5,0.7,1.0)	(0.2,0.4,0.4)	(0.1,0.1,0.2)
S5	(0.0,0.1,0.2)	(0.1,0.2,0.3)	(0.2,0.3,0.3)	(0.5,0.7,1.0)	-	(0.2,0.3,0.4)	(0.0,0.0,0.0)
S6	(0.0,0.0,0.0)	(0.0,0.0,0.0)	(0.1,0.1,0.2)	(0.2,0.4,0.4)	(0.2,0.3,0.4)	-	(0.6,0.8,1.0)
S7	(0.1,0.1,0.1)	(0.2,0.3,0.3)	(0.0,0.1,0.1)	(0.1,0.1,0.2)	(0.0,0.0,0.0)	(0.6,0.8,1.0)	-

$$\max(m_1, m_2, \ldots, m_n), \max(r_1, r_2, \ldots, r_n)). \quad (15)$$

The scores of the segments are fuzzy numbers, and so are defuzzified before the candidate segmentations are ranked. The following sections discuss the scoring procedure, the defuzzification, and the ranking algorithm.

A. Scoring Candidate Segmentation

The score of a segment is defined as the difference between internal and external scores. Segment T is assumed to comprise n member sentences, so the score of T is given by

$$\text{The score of segment } T = \text{internal score} \ominus \text{external score}. \quad (16)$$

The internal score is computed as

$$\text{internal score} = \frac{1}{n} \sum_{i \in T} \operatorname*{fmax}_{\substack{k \in T \\ k \neq i}} (\text{sim}(i,k)) \quad (17)$$

where i and k represent the member sentences of T. The external score of a segment is given by

$$\text{external score} = \frac{1}{2} \left(\operatorname*{fmax}_{\substack{i \in M_L \\ j \in E_L}} (\text{sim}(i,j)) \oplus \operatorname*{fmax}_{\substack{i \in M_R \\ j \in E_R}} (\text{sim}(i,j)) \right) \quad (18)$$

where M_L is the set of left-side members of the segment; M_R is the set of right-side members of the segment; E_L is the set of adjacent nonmembers of the segment on the left side, and E_R represents that on the right side. These sets must be all equally sized.

Fig. 5 shows an example of the computation used to score a segment. In Fig. 5, the segment that is currently being processed comprises five sentences, indicated by shaded circles, and the values of M_L, M_R, E_L, and E_R are two. The internal score of a segment represents the probability that the members describe the same subtopic, whereas the external score represents the probability that both members and nonmembers describe the same subtopic. Obviously, a large internal score shows that the current segment has a high probability for describing the same subtopic, whereas a large external score indicates that the boundary of the current segment is not a well-selected position.

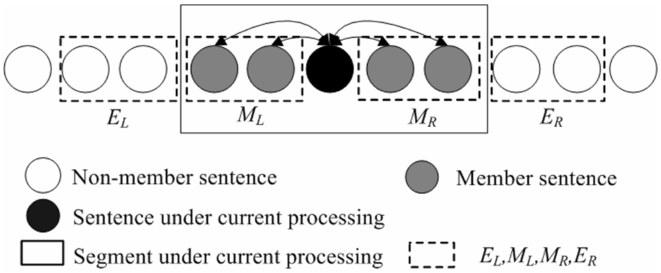

Fig. 5. Example of computation of score.

The scores of segments are fuzzy numbers, and so are first defuzzified by defuzzification in [30]. The defuzzified value A_e of $A = (l, m, r)$ can be computed by

$$A_e = \frac{l + 2m + r}{4}. \quad (19)$$

B. Ranking Candidate Segmentations

The score of a candidate segmentation is the sum of the scores of each segment, which are computed using the aforementioned scoring approach. The candidate segmentation with the highest score exhibits high cohesion and low repulsion among the segments, and vice versa. The candidate segmentation with the highest score is selected to segment the document.

Since there are many candidate segmentations in a document, dynamic programming and recursive function are used to determine the best one. The algorithm is as follows and will return the highest score and a set of the best shifts for a test document.

```
**FindingOptimalShiftsOfSegment** ($T$) **returns** the highest
score and a set of optimal cut points of segment $T$

    **local variables:** *ts*, a temporal score
                         *tf*, the position of a temporal cut point
                         *tb*, the set of optimal cut points for a
                         segment
                         *ns*, the number of the sentences in $T$
                         *c*, the cut point which divides $T$ into
                         two segments *TF* and *TB*
    function:
        SegScore(T'), the score of a segment $T'$ obtained
        from formula (16)
        ScoreOf(T'), the highest score of $T'$ obtained from the
        return of FindingOptimalShiftsOfSegment(T')
OptimalOf(T'), the set of optimal cut points of $T'$ obtained
from the return of FindingOptimalShiftsOfSegment(T')
main {
    ts ← SegScore(T)/* initialization of the highest score
    for $T$ */
        tf ← the end of $T$/* initialization of cut point for $T$ */
        if $ns > 2$ then {/* $T$ has more than two sentences */
            for $c := 2$ to ns {
                if SegScore(TF) + ScoreOf(TB) > ts then {
                    ts ← SegScore(TF) + ScoreOf(TB)/*
                    update current highest score for $T$ */
                    tf ← c/* update current optimal cut point
                    for $T$ */
                    tb ← OptimalOf(TB)
                }
            }
    }
        return ts and tf ∪ tb/* the highest score and a set of
        optimal cut points for segment $T*/$
}
```

VI. Experimental Results

Choi [17] pointed out that the performance of topic segmentation depends on the application domains. This paper focuses on domain-specific applications with small corpora, and two sets of data on short writings by students are used to evaluate and compare the proposed method with other approaches. The first data set comprises 1200 training essays and 100 test essays on the theme, "Recess at school," and the second data set comprises 805 training essays and 100 test essays on the theme "A day as a student on duty." Students in the eighth grade composed all of the writings in Chinese.

TABLE II
Semantics of Both Sentences 1 and 4 in the Example

	Sentence 1	Sentence 4

Keywords	Students(809)
	Look forward to(215)
		Horrible(31)
		Study(1560)

Dimensions	(0.84,0.84,0.84) Time	(0.89,1.00,1.00) Teacher
	(0.61,0.61,0.61) Teacher	(0.87,0.88,0.98) Time
	(0.35,0.35,0.35) Study	(0.51,0.58,0.63) Ring
	(0.26,0.26,0.26) Class	(0.51,0.56,0.62) Classroom
	(0.22,0.22,0.22) Classroom	(0.54,0.55,0.66) Classmate
	(0.22,0.22,0.22) School	(0.37,0.40,0.48) Play
	(0.18,0.18,0.18) Play	(0.35,0.38,0.46) Moment
	(0.17,0.17,0.17) Thing	(0.26,0.37,0.37) Thing
	(0.17,0.17,0.17) Rest	(0.29,0.31,0.40) Class
	(0.16,0.16,0.16) Classmate	(0.26,0.30,0.38) Sound
	(0.16,0.16,0.16) Use	(0.28,0.29,0.39) Student
	(0.15,0.15,0.15) Campus	(0.21,0.21,0.32) Worm
	(0.14,0.14,0.14) Look	(0.20,0.20,0.31) Attention
	(0.13,0.13,0.13) Happy	(0.17,0.20,0.29) Back
	(0.13,0.13,0.13) Ring	(0.17,0.20,0.29) Come
	(0.11,0.11,0.11) Moment	(0.18,0.19,0.29) Prepare
	(0.11,0.11,0.11) This way	(0.17,0.19,0.28) Feel
	(0.10,0.10,0.10) Like	(0.19,0.19,0.30) Material
	(0.09,0.09,0.09) Activity	(0.18,0.18,0.29) Rest
	(0.09,0.09,0.09) Question	(0.18,0.18,0.29) Sleep

A. *Example for Segmentations*

The example in Appendix shows a literal translation of an essay from the first dataset. We will use this example to illustrate various terms discussed in the proposed method. Table II displays the semantics of sentences 1 and 4 in the essay. In Table II, row 2 shows the keywords of the two sentences and the frequency of the keywords, whereas row 3 shows the dimensions of the semantics of the sentences and the entries of the dimensions with fuzzy numbers. Given that the constant C in (7) is 200, (4)–(7) yield the semantics of the words "study" and "horrible" in the "time" dimension, to (1.0, 1.0, 1.0) and (0, 0.08, 0.85). Using (9), the semantics of sentence 4 in the "time" dimension can be computed as follows:

$$\left(\frac{1\times 1+0.16\times 0}{1+0.16}, \frac{1\times 1+0.16\times 0.08}{1+0.16}, \frac{1\times 1+0.16\times 0.85}{1+0.16}\right) = (0.87, 0.88, 0.98).$$

Equations (10) and (11) and the result in Table II yield the semantic similarity between sentences 1 and 4, which is (0.65, 0.77, 0.81). Table III shows the semantic similarities among sentences 1–14 in the essay.

The scores of every possible segment can be computed from all of the similarities and (16)–(18). For instance, the scores of

$$\frac{1}{2}\left(\mathrm{fmax}\begin{pmatrix}(0.65,0.83,0.97)\\(0.79,0.90,0.96)\\(0.51,0.69,0.87)\\(0.61,0.73,0.80)\end{pmatrix} + \mathrm{fmax}\begin{pmatrix}(0.00,0.75,0.98)\\(0.00,0.69,1.00)\\(0.91,0.91,0.91)\\(0.69,0.77,0.88)\end{pmatrix}\right) = (0.88, 0.91, 0.99)$$

TABLE III
SEMANTIC SIMILARITIES AMONG SENTENCES 1 TO 14 IN APPENDIX

	1	2	3	4	5	6	7
1		(0.68,0.82,0.88)	(0.49,0.64,0.72)	(0.65,0.77,0.81)	(0.83,0.83,0.83)	(0.68,0.77,0.88)	(0.00,0.75,0.98)
2	(0.68,0.82,0.88)		(0.56,0.74,0.90)	(0.65,0.83,0.97)	(0.79,0.90,0.96)	(0.58,0.76,0.90)	(0.01,0.76,1.00)
3	(0.49,0.64,0.72)	(0.56,0.74,0.90)		(0.51,0.69,0.87)	(0.61,0.73,0.80)	(0.30,0.50,0.71)	(0.00,0.58,1.00)
4	(0.65,0.77,0.81)	(0.65,0.83,0.97)	(0.51,0.69,0.87)		(0.67,0.75,0.83)	(0.45,0.63,0.82)	(0.00,0.65,1.00)
5	(0.83,0.83,0.83)	(0.79,0.90,0.96)	(0.61,0.73,0.80)	(0.67,0.75,0.83)		(0.67,0.74,0.85)	(0.00,0.75,0.98)
6	(0.68,0.77,0.88)	(0.58,0.76,0.90)	(0.30,0.50,0.71)	(0.45,0.63,0.82)	(0.67,0.74,0.85)		(0.00,0.69,1.00)
7	(0.00,0.75,0.98)	(0.01,0.76,1.00)	(0.00,0.58,1.00)	(0.00,0.65,1.00)	(0.00,0.75,0.98)	(0.00,0.69,1.00)	
8	(0.82,0.82,0.82)	(0.81,0.92,0.97)	(0.59,0.71,0.79)	(0.68,0.77,0.84)	(0.91,0.91,0.91)	(0.69,0.77,0.88)	(0.01,0.76,0.98)
9	(0.77,0.84,0.88)	(0.72,0.88,0.97)	(0.50,0.66,0.82)	(0.65,0.77,0.88)	(0.83,0.87,0.92)	(0.65,0.79,0.89)	(0.00,0.75,1.00)
10	(0.62,0.81,0.89)	(0.56,0.82,0.99)	(0.42,0.68,0.92)	(0.55,0.77,0.96)	(0.64,0.81,0.92)	(0.37,0.65,0.87)	(0.00,0.66,1.00)
11	(0.00,0.72,1.00)	(0.00,0.67,1.00)	(0.00,0.48,1.00)	(0.00,0.61,1.00)	(0.00,0.68,1.00)	(0.00,0.65,1.00)	(0.00,0.61,1.00)
12	(0.77,0.77,0.77)	(0.66,0.75,0.85)	(0.47,0.61,0.69)	(0.59,0.68,0.75)	(0.75,0.75,0.75)	(0.67,0.76,0.88)	(0.00,0.73,0.97)
13	(0.59,0.81,0.90)	(0.65,0.89,1.00)	(0.52,0.75,0.93)	(0.68,0.87,1.00)	(0.68,0.84,0.94)	(0.40,0.68,0.88)	(0.00,0.70,1.00)
14	(0.66,0.72,0.79)	(0.56,0.72,0.87)	(0.37,0.54,0.70)	(0.45,0.60,0.77)	(0.69,0.74,0.81)	(0.59,0.75,0.89)	(0.00,0.66,0.99)

	8	9	10	11	12	13	14
1	(0.82,0.82,0.82)	(0.77,0.84,0.88)	(0.62,0.81,0.89)	(0.00,0.72,1.00)	(0.77,0.77,0.77)	(0.59,0.81,0.90)	(0.66,0.72,0.79)
2	(0.81,0.92,0.97)	(0.72,0.88,0.97)	(0.56,0.82,0.99)	(0.00,0.67,1.00)	(0.66,0.75,0.85)	(0.65,0.89,1.00)	(0.56,0.72,0.87)
3	(0.59,0.71,0.79)	(0.50,0.66,0.82)	(0.42,0.68,0.92)	(0.00,0.48,1.00)	(0.47,0.61,0.69)	(0.52,0.75,0.93)	(0.37,0.54,0.70)
4	(0.68,0.77,0.84)	(0.65,0.77,0.88)	(0.55,0.77,0.96)	(0.00,0.61,1.00)	(0.59,0.68,0.75)	(0.68,0.87,1.00)	(0.45,0.60,0.77)
5	(0.91,0.91,0.91)	(0.83,0.87,0.92)	(0.64,0.81,0.92)	(0.00,0.68,1.00)	(0.75,0.75,0.75)	(0.68,0.84,0.94)	(0.69,0.74,0.81)
6	(0.69,0.77,0.88)	(0.65,0.79,0.89)	(0.37,0.65,0.87)	(0.00,0.65,1.00)	(0.67,0.76,0.88)	(0.40,0.68,0.88)	(0.59,0.75,0.89)
7	(0.01,0.76,0.98)	(0.00,0.75,1.00)	(0.00,0.66,1.00)	(0.00,0.611.00)	(0.00,0.73,0.97)	(0.00,0.70,1.00)	(0.00,0.66,0.99)
8		(0.83,0.87,0.92)	(0.62,0.79,0.90)	(0.00,0.68,1.00)	(0.75,0.75,0.75)	(0.68,0.86,0.94)	(0.70,0.75,0.82)
9	(0.83,0.87,0.92)		(0.61,0.83,0.95)	(0.00,0.68,1.00)	(0.68,0.75,0.80)	(0.62,0.83,0.96)	(0.61,0.72,0.85)
10	(0.62,0.79,0.90)	(0.61,0.83,0.95)		(0.00,0.64,1.00)	(0.49,0.67,0.80)	(0.52,0.82,1.00)	(0.41,0.64,0.85)
11	(0.00,0.68,1.00)	(0.00,0.68,1.00)	(0.00,0.64,1.00)		(0.00,0.62,1.00)	(0.00,0.65,1.00)	(0.00,0.63,1.00)
12	(0.75,0.75,0.75)	(0.68,0.75,0.80)	(0.49,0.67,0.80)	(0.00,0.62,1.00)		(0.56,0.72,0.85)	(0.72,0.80,0.85)
13	(0.68,0.86,0.94)	(0.62,0.83,0.96)	(0.52,0.82,1.00)	(0.00,0.65,1.00)	(0.56,0.72,0.85)		(0.44,0.66,0.85)
14	(0.70,0.75,0.82)	(0.61,0.72,0.85)	(0.41,0.64,0.85)	(0.00,0.63,1.00)	(0.72,0.80,0.85)	(0.44,0.66,0.85)	

segment {5,6,7} can be derived by adding the internal score to the external score. The internal score of segment {5,6,7} in Table III is

$$\left(\frac{0.67+0.67+0.67}{3}, \frac{0.75+0.75+0.74}{3}, \frac{0.83+0.85+0.85}{3}\right) = (0.67, 0.75, 0.84).$$

The external score of the segment is shown in the equation at the bottom of the previous page. After defuzzification, the score of segment {5,6,7} is -0.17.

The original essay comprises five paragraphs, and therefore, four subtopic shifts. Applying the proposed method to the example, we will find four subtopic shifts and therefore segment it into five segments. The best subtopic shifts for the example occur at positions, respectively, between sentences 4 and 5, 11 and 12, 20 and 21, and 23 and 24 according to our method. The detected four subtopic shifts coincide with the subtopic shifts of the original writing at two locations between sentences 4 and 5, as well as 11 and 12. According to the teacher, the segmentation by the proposed method is quite reasonable.

Below we explain through the example how the fuzzy model is able to select one segmentation over the other even if their difference is very small. Because the documents in the training corpus all had the same theme, the difference between the degrees of semantic similarities between neighboring sentences is usually small. This fact is evidenced by the m values of the similarities in Table III and reduced the accuracy of the segmentations obtained by previous methods [4], [7]. However, the proposed fuzzy model overcomes this difficulty, as it incorporates the fuzzy number in the similarities.

To illustrate the above arguments, consider two segmentations A and B: the optimal segmentation A comprises segments {1,2,3,4} and {5,6,7,8,9,10,11}, and possible segmentation B comprises segment {1,2,3,4,5} and {6,7,8,9,10,11}. The m values of the fuzzy numbers alone yield scores of segmentations A and B of 0.47 and 0.49, respectively, and segmentation B

TABLE IV
SCORES OF TWO SEGMENTATIONS

segmentation	internal score	external score
A {1,2,3,4}{5,6,7,8,9,10,11}	(0.91,0.91,1.00)	(0.67,0.75,0.83)
B {1,2,3,4,5}{6,7,8,9,10,11}	(0.83,0.90,0.96)	(0.67,0.75,0.98)

TABLE V
EVALUATION OF METHODS [4], [7], INCLUDING PROPOSED METHOD, USING *WINDOWDIFF*

	WD	
	dataset 1	dataset 2
the proposed method	0.35	0.37
the method in [4]	0.43	0.41
the method in [7]	0.49	0.46

TABLE VI
RATIO OF NUMBER OF ACCURATE SHIFTS TO TOTAL NUMBER OF SHIFTS OBTAINED USING THE PROPOSED METHOD AND OTHER METHODS

	hits		moves	
	dataset 1	dataset 2	dataset 1	dataset 2
the proposed method	0.46	0.45	0.30	0.28
the method in [4]	0.33	0.34	0.29	0.31
the method in [7]	0.20	0.21	0.18	0.21

TABLE VII
RATIO OF NUMBER OF INACCURATE SHIFTS TO TOTAL NUMBER OF SHIFTS OBTAINED USING THE PROPOSED METHOD AND OTHER METHODS

	insertions		deletions	
	dataset 1	dataset 2	dataset 1	dataset 2
the proposed method	0.13	0.15	0.11	0.12
the method in [4]	0.26	0.25	0.12	0.10
the method in [7]	0.41	0.39	0.21	0.19

would be treated as the optimal segmentation. However, computing with fuzzy numbers would yield scores of segmentations A and B of 0.45 and 0.43, respectively, and segmentation A would be correctly treated as the optimal segmentation.

Table IV shows the internal and external scores of the two segmentations. The difference between the two segmentations lies in the classification of sentence 5. In Table IV, both m values of the external scores for segmentations A and B are 0.75, while the m values of the internal scores are close to 0.91. Therefore, approaches that use only crisp values cause difficulties in selecting optimal segmentation and hence the classification of sentence 5. However, the l values of the internal score for segmentation A and B in the second column are 0.91 and 0.83, respectively, indicating that sentence 5 should be grouped into segment {5,6,7,8,9,10,11}. Additionally, the r values of the external score for segmentations A and B in the last column are 0.83 and 0.98, respectively. This result strengthens the decision that sentence 5 should be included into segment {5,6,7,8,9,10,11}.

B. Evaluation

Test essays are segmented into subtopics using the methods developed in [4] and [7], as well as the method proposed herein, to yield hypothetical segmentations of these essays. Synthetic evaluation using *WindowDiff* [31] and analytic evaluation [7] are then used to estimate the performance of these approaches.

WindowDiff is used to determine the index WD of hypothetic segmentation, which is applied to measure the similarity between standard segmentation and the proposed segmentations. A lower WD of the proposed segmentation indicates that the segmentation is more similar to the standard segmentation. Table V shows the results of different methods evaluated by using *WindowDiff*. The WD field shows the mean of the WDs for the two data sets of essays mentioned above. The results demonstrate that the proposed method segments the document into subtopics more accurately and precisely than earlier approaches.

One method of evaluation [7] was used to interpret the differences among the performances in Table V. The evaluation classifies the subtopic shifts into *hits, moves, insertions,* and *deletions*. A *hit* occurs when a shift lines up with the real shift. A *move* occurs when a shift does not line up with the real shift in the same position. An *insertion* occurs when the method generates a shift that does not exist in the position. A *deletion* occurs when a real shift exists but the method does not generate one. Both hits and moves are regarded as accurate shifts, and both insertions and deletions are regarded as inaccurate shifts.

Two humans are asked to classify subtopic shifts into four categories by examining the segmentations of two datasets. Tables VI and VII show the ratio of the shifts in each category to all shifts in the examined segmentations. In Table VI, the hits in the proposed method are 11–13% more than those in [4] and 24–26% more than those in [7]. However, the *moves* detected by the proposed method are similar to that obtained in [4] but 7–12% higher than that obtained in [7]. In Table VII, the proportion of *insertions* obtained using the proposed method is 10–13% lower than that obtained in [4] and 24–28% lower than that obtained in [7]. The proportion of *deletions* obtained using the proposed method is similar to that obtained in [4] but 7–9% lower than that obtained in [7].

The experiment demonstrates that the proposed method can increase the accuracy of shifts and reduce the proportion of incorrect shifts. The incorrect shifts are typically associated with the uncertainty caused by the small training corpus and short texts. This result shows that the proposed method more accurately handles the uncertainty.

VII. CONCLUSION

The paper proposes a new method for segmenting subtopics based on the fuzzy modeling of the semantics of words and sentences. It focuses on uncertainty in both word semantics and sentence semantics. In particular, the semantics are computed from the co-occurrence of words and represented using a fuzzy number rather than a crisp value. Additionally, this paper developed a method for measuring the similarity between the fuzzy semantics of two sentences and to describe the uncertainty in the similarity by a fuzzy number. Based on the fuzzy similarities, the proposed method more accurately segments a document into subtopics. The experiments and examples in Section VI clearly demonstrate that the method can greatly and efficiently increase the accuracy of segmentation.

It is clear that semantic relations among the subtopics are often hierarchical even though a text is represented as a linear sequence of subtopics. Hence, one of the extensions along the proposed method is to construct the hierarchical semantics structure of subtopics for the text. Another extension is to develop a hybrid method that integrates the proposed method with such techniques as discourse structure and noun coreferent resolution. These techniques will need to use external resources such as a grammar parser and WordNet. This would increase the precision of semantics measure for words and sentences and the accuracy of subtopic segmentation.

Appendix

Literal translation of an essay in the first dataset.

1. Many students look forward to class recess.
2. During recess, some review the just-learned materials while others chat.
3. When the recess bell rings, everyone looks like they have just been released from custody,
4. and you can tell how horrible the class is.
5. Observing classmates' behaviors during recess is very interesting!
6. Some students hurriedly finish up their homework,
7. Students on duty slowly and reluctantly clean the blackboard,
8. Peers happily play games or chat,
9. The teacher prepares the teaching-aids for the next class.
10. Indeed, a ten-minute break is short,
11. but it satisfies the students!
12. Some students wake up when the recess bell rings.
13. They totally forget all of the tedium and boredom of the class.
14. The snack bar is crowded.
15. In front of the student center, several students are ready for PE classes.
16. On the basketball field, several students play basketball as quickly as monkeys.
17. Some students walk back to their classrooms following the PE class.
18. When the recess is about to end,
19. some students are in a hurry to return to the classrooms.
20. Some students prepare their equipment for the next course.
21. The entire campus becomes silent like a dead town, which wakes up every forty-five minutes.
22. When the town comes alive, it likes the hurly-burly of a market.
23. However, during a ten-minute-recess, the campus becomes the dead town again!
24. The ringing bell can both relax,
25. and repress us.
26. A ten-minute recess is precious,
27. for it allows us to take a rest, and a break from study.

References

[1] M. Bawa, G. S. Manku, and P. Raghavan, "SETS: Search enhanced by topic-segmentation," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Toronto, ON, Canada, 2003, pp. 306–313.

[2] R. Angheluta, R. De Busser, and M. F. Moens, "The use of topic segmentation for automatic summarization," in *Proc. Workshop Autom. Summarization*, Philadelphia, PA, 2002, pp. 66–70.

[3] M. J. Mana-Lopez, M. de Buenaga, and J. M. Gomez-Hidalgo, "Multidocument summarization: An added value to clustering in interactive retrieval," *ACM Trans. Inf. Syst.*, vol. 22, no. 2, pp. 215–241, 2004.

[4] T. H. Chang and C. H. Lee, "Topic segmentation for short text," in *Proc. 17th Pacific Asia Conf. Language, Inf. Comp.*, Singapore, 2003, pp. 159–165.

[5] F. Y. Y. Choi, "Advances in domain independent liner text segmentation," in *Proc. 1st Conf. North Amer. Chapter Assoc. Comp. Linguistics*, Seattle, Washington, 2000, pp. 26–33.

[6] O. Ferret, "Using collocations for topic segmentation and link detection," in *Proc. 19th Int. Conf. Comp. Linguistics*, Taipei, Taiwan, R.O.C., 2002, pp. 260–266.

[7] J. M. Ponte and W. B. Croft, "Text segmentation by topic," in *Proc. 1st Eur. Conf. Res. Adv. Technol. Digital Libraries*, Pisa, Italy, 1997, pp. 120–129.

[8] L. A. Zadeh, "Fuzzy sets," *Inf. Contr.*, vol. 8, pp. 338–353, 1965.

[9] G. Akrivas, M. Wallace, G. Andreou, G. Stamou, and S. Kollias, "Context-sensitive semantic query expansion," in *Proc. IEEE Int. Conf. Artif. Intell. Syst.*, Divnomorskoe, Russia, 2002, pp. 109–114.

[10] P. Subasic and A. Huetter, "Affect analysis of text using fuzzy semantic typing," *IEEE Trans. Fuzzy Syst.*, vol. 9, no. 4, pp. 483–496, 2001.

[11] J. P. Sun, K. Shaban, S. Podder, F. Karry, O. Basir, and M. Kamel, "Fuzzy semantic measurement for synonymy and its application in an automatic question-answering system," in *Proc. IEEE Int. Conf. Natural Language Process. Knowl. Eng.*, Beijing, China, 2003, pp. 263–268.

[12] T. Takagi and K. Kawase, "A trial of data retrieval using conceptual fuzzy sets," *IEEE Trans. Fuzzy Syst.*, vol. 9, no. 4, pp. 497–505, 2001.

[13] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, pp. 177–210, 1999.

[14] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Comp. Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[15] X. Ji and H. Zha, "Domain-independent text segmentation using anisotropic diffusion and dynamic programming," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Toronto, ON, Canada, 2003, pp. 322–329.

[16] A. C. Jobbins and L. J. Evett, "Text segmentation using reiteration and collocation," in *Proc. 17th Int. Conf. Comp. Linguistics*, Montreal, PQ, Canada, 1998, pp. 614–618.

[17] ——, "Segmenting documents using multiple lexical feature," in *Proc. 5th Int. Conf. Doc. Anal. Recognit.*, Banglore, India, 1999, pp. 721–724.

[18] M. F. Moens and R. de Busser, "Generic topic segmentation of document texts," in *Proc. 24th ACM SIGIR Annu. Int. Conf. Res. Develop. Inf. Retrieval*, New Orleans, LA, 2001, pp. 418–419.

[19] G. Salton, J. Allan, and A. Singhal, "Automatic text decomposition and structuring," *Inf. Process. Manage.*, vol. 32, no. 2, pp. 127–138, 1996.

[20] B. Bigi, R. De Mori, M. El-Beze, and T. Spriet, "Detecting topic shifts using a cache memory," in *Proc. 5th Int. Conf. Spoken Language Process.*, Sydney, Australia, 1998, pp. 2331–2334.

[21] D. M. Blei and P. J. Moreno, "Topic segmentation with an aspect hidden Markov model," in *Proc. 24th ACM SIGIR Annu. Int. Conf. Res. Develop. Inf. Retrieval*, New Orleans, LA, 2001, pp. 343–348.

[22] M. Franz, B. Ramabhadran, T. Ward, and M. Picheny, "Automated transcription and topic segmentation of large spoken archives," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 953–956.

[23] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc. 41st Annu. Meeting Assoc. Comp. Linguistics*, Sapporo, Japan, 2003, pp. 562–569.

[24] G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Comp. Linguistics*, vol. 7, no. 1, pp. 31–57, 2001.

[25] J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," in *Proc. 1998 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, Washington, 1998, vol. 1, pp. 333–336.

[26] M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," in *Proc. 39st Annu. Meeting Assoc. Comp. Linguistics*, Toulouse, France, 2001, pp. 491–498.

[27] W. Y. Ma and K. J. Chen, "Introduction to CKIP chinese word segmentation system for the first international Chinese word segmentation bakeoff," in *Proc. 2nd SIGHAN Workshop Chinese Language Process.*, Sapporo, Japan, 2003, pp. 168–171.

[28] Y. F. Tsai and K. J. Chen, “Reliable and cost-effective pos-tagging,” *Int. J. Comp. Linguistics Chinese Language Process.*, vol. 9, no. 1, pp. 83–96, 2004.

[29] T. H. Chang and C. H. Lee, “Automatic Chinese unknown word extraction using small-corpus-based method,” in *Proc. IEEE Int. Conf. Natural Language Process. Knowledge Eng.*, Beijing, China, 2003, pp. 459–464.

[30] A. Kaufmann and M. M. Gupta, *Fuzzy Mathematical Models in Engineering and Management Science*. Amsterdam, The Netherlands: Elsevier, 1988.

[31] L. Pevzner and M. A. Hearst, “A critique and improvement of an evaluation metric for text segmentation,” *Comp. Linguistics*, vol. 28, no. 1, pp. 19–36, 2002.

Tao-Hsing Chang received the B.S. degree in mathematics and science education from Taipei Municipal Teachers College, Taiwan, R.O.C., in 1996 and the M.S. degree in computer science from National Chiao Tung University, Taiwan, in 1998, where he is currently pursuing the Ph.D. degree in computer science.

He is a Research Fellow with the Research Center for Psychological and Educational Testing, National Taiwan Normal University, Taiwan. His research interests include fuzzy modeling, natural language processing, information retrieval, and automated Chinese essay scoring.

Chia-Hoang Lee received the Ph.D. degree in computer science from the University of Maryland, College Park, in 1983.

From 1984 to 1985, he was with the Department of Mathematics and Computer Science, University of Maryland. From 1985 to 1992, he was with the Department of Computer Science, Purdue University, West Lafayette, IN. He is currently a Professor in the Department of Computer Science and Deputy Director of the MediaTek Research Center, National Chiao Tung University, Taiwan. His current research interests include artificial intelligence, human machine interface systems, natural language processing, and automated Chinese essay scoring.