PACIS 2004 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

December 2004

Use of Text Summarization for Supporting Event Detection

Chih-Ping Wei
National Sun Yat-sen University

Pao-Feng Wu
Taiwan Semiconductor Manufacturing Company

Yen-Hsien Lee
National Sun Yat-sen University

Use of Text Summarization for Supporting Event Detection

Chih-Ping Wei
Dept. of Info. Management
National Sun Yat-sen Univ.
Kaohsiung, Taiwan, ROC
cwei@mis.nsysu.edu.tw

Pao-Feng Wu
Fab-2 Manufacturing Dept.
Taiwan Semiconductor
Manufacturing Company
Hsinchu, Taiwan, ROC
pfwud@tsmc.com

Yen-Hsien Lee
Dept. of Info. Management
National Sun Yat-sen Univ.
Kaohsiung, Taiwan, ROC
roso@mis.nsysu.edu.tw

Abstract

Event detection, an important task in organizational environmental scanning, is to identify the onset of new events from streams of news stories. Existing event detection techniques identify whether a news story contains an unseen event generally by comparing the similarity between features of a new news story and past news stories. However, for illustration and comparison purposes, a news story may contain sentences or paragraphs that are not highly relevant to defining its event. The inclusion of such sentences and paragraphs in the similarity comparison by a traditional event detection technique might significantly degrade its detection effectiveness. Therefore, in this study, we propose and develop a summary-based event detection (SED) technique that first filters less relative sentences or paragraphs from each news story before performing feature-based event detection. Using a traditional event detection technique (i.e., INCR) as a performance benchmark, our empirical evaluation results suggest that the proposed SED technique achieve comparable or even better detection effectiveness than its benchmark technique.

Keywords: Event detection, Text summarization, Environmental scanning

1. Introduction

Globalization and the emergence of E-commerce have made strategy management increasingly important to organizations. Strategy management is the set of managerial decisions and actions that determines the long-run performance of a corporation (Wheelen and Hunger 2002). Environmental scanning, an important process in strategy management, can provide an organization a comprehensive view or understanding of the current and future condition to its external environment, and that will become the foundation for basing strategic planning, making strategic decisions, or guiding product/service development (Maier et al. 1997; Jain 1984; Mason & Wilson 1987). Empirical research results suggest that environmental scanning is linked with improved organizational performance. For instance, Daft et al. (1988) found that chief executives of high-performing firms scanned their environments more frequently and more broadly than their counterparts in low-performing firms. Conversely, failure to scan has been associated with corporate decline and failure (Starbuck et al. 1978). Accordingly, scanning the external business environment for events, trends, and changes has become a critical information activity of chief executive officers for planning their firms.

One of the major sources for environmental scanning is online news websites (Choo 1998). With the rapid growth of the World Wide Web and electronic information services, the online news sources available on the Internet have grown tremendously in number and sheer volume. On the other hand, increases in scope and complexity of business environments make the interval between scanning efforts needed shorten. Consequently, environmental scanning becomes more difficult to manage and has been a burden to managers, demanding an efficient and effective technique to facilitate their detection of the onset of new events from news documents and track of subsequent news stories that discuss an event of interest. In this study, we mainly focus on event detection for supporting organizational environmental scanning.

Event detection is to identify the onset of new events from streams of news stories, where an event refers to something happening in a certain place at a certain time (Allan et al. 1998; Yang et al. 1998; Yang et al. 1999) and an event topic consists of a set of event instances of the same type. For example, Cisco Systems Inc. acquiring V-Bits Inc. is an event pertaining to the event topic of business merger. Traditional event detection techniques identify an unseen event generally by comparing the similarity of features between a new news story and past ones.

Nevertheless, being a feature-based approach, traditional event detection techniques incur several problems. First, for illustration and comparison purposes, a news story may contain sentences or paragraphs that are not highly relevant to defining its event. The inclusion of such sentences and paragraphs in the similarity comparison by a traditional feature-based event detection technique might significantly degrade its detection effectiveness. Secondly, vocabulary discrepancies between reporters even when they describe the same event may degrade the effectiveness of an existing event detection technique (Wei and Lee 2004). For example, some reporters may use “purchase” to describe a business merger event, while others may use “acquisition” for the same event. Thirdly, two news stories for different events may contain very similar feature sets since the events belong to the same event topic (Wei and Lee 2004).

Motivated by the significance of event detection in supporting organizational environmental scanning and the need for improving effectiveness of event detection, this study attempts to address the first problem inherent to traditional event detection techniques by filtering irrelevant sentences or paragraphs in a news story before performing feature-based event detection. Text summarization, a process of selecting from a full text document important sentences that serve as a summary of it, can be used to identify and remove from a news document irrelevant sentences or paragraphs. In this vein, the first problem of traditional feature-based event detection techniques can be minimized, potentially resulting higher event detection effectiveness.

Specifically, in this study, we propose the Summary-based Event-Detection (SED), which first selects important sentences as a summary for the news story and subsequently performs event detection based on the summary of the news story. The remainder of the paper is organized as follows. Section 2 reviews the literature relevant to this research, including traditional event detection techniques and text summarization ones. Section 3 details the proposed SED technique. In Section 4, we depict the experimental design and discuss our experimental results. In Section 5, we conclude with a summary and some future research directions.

2. Literature Review

This section reviews traditional feature-based event detection techniques and text summarization ones, which together serve as the foundation for the development of the SED technique.

2.1 Event Detection

Event detection aims at identifying stories in several continuous news streams that pertain to new or previously unidentified events (Yang et al. 1999). Event detection can be in a form of retrospective or online detection. The retrospective event-detection entails the discovery of previously unidentified events in a chronologically ordered accumulation of new stories, while the online event-detection identifies the onset of new events from live news fed in real-time.

Most of the proposed event detection algorithms, retrospective or online, are developed based on the document clustering approach. Yang et al. (1999) implemented two clustering methods for event detection: group-average clustering algorithm (GAC) and single-pass incremental clustering algorithm (INCR). GAC, designed for retrospective detection, performs agglomerative clustering for producing hierarchically organized document clusters. GAC employed the conventional vector space model to represent news documents and clusters. Specifically, each document is represented using a vector of weighted terms, based on the TF×IDF (term frequency×inverse document frequency) scheme. For cluster representation, the normalized vector of documents in a cluster is summed and the *k* most significant terms, called the prototype or centroid of the cluster, are selected to represent the cluster. To improve computation efficiency and to preserve the characteristics that events tend to appear in news bursts, GAC adopts a divide-and-conquer strategy that grows clusters iteratively. In each iteration, the current pool of clusters is divided according to their temporal order into evenly sized buckets. Subsequently, group-average clustering is applied to each bucket locally, merging smaller clusters into larger ones. Periodically, the news documents within each of the top-level clusters are reclustered. Reclustering is useful when events straddle the initial temporal-bucket boundaries or when the bucketing causes undesirable groupings of news stories about different events.

On the other hand, INCR, suitable to both retrospective and online detection, is a single-pass incremental clustering algorithm that produces nonhierarchical clusters incrementally (Yang et al. 1999). For retrospective detection, the TF×IDF scheme is adopted to represent documents or clusters. However, to deal with the problem of continuously incoming documents that might change term weighting and vector normalization during online detection, the incremental IDF is employed by INCR. That is, $IDF(f, p) = \log_2 \frac{n(p)}{n(f, p)}$, where p is the current time point, $n(p)$ is the number of documents accumulated up to time p (including the retrospective corpus if used), and $n(f, p)$ is the document frequency of term f at time p.

Because news stories discussing the same event tend to be temporally proximate, a combined measure of lexical similarity and temporal proximity as a criterion for event detection is often employed. Moreover, since a time gap between bursts of topically similar news stories is often an indication of different events, the incorporation of a time window for event scoping is also commonly adopted (Yang et al. 1998). Particularly, INCR incorporate a time penalty when calculating the similarity between a document x and any cluster c in the past. The time penalty can be a uniformly weighted time window (i.e., a time window of m documents before x is imposed) or a linear decaying-weight function (shown as below).

$$similarity(x, c) = \begin{cases} (1 - \frac{i}{m}) \times similarity(x, c) & \text{if } c \text{ has any member in the time window} \\ 0 & \text{otherwise} \end{cases}$$

where i is the number of news documents between x and the most recent member document in c, and m is the time window of documents before x.

For retrospective detection, INCR sequentially processes news documents. A document is absorbed by the most similar cluster in the past if the similarity between the document and the cluster is larger than a pre-selected clustering threshold (t_c); otherwise, the document becomes the seed of a new cluster. For online detection, the novelty threshold (t_n) is introduced. If the maximal similarity between the current document and a cluster in the past is no less than t_n, the document is flagged as containing an old event.

2.2 Text Summarization

Text summarization, defined as "a reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source" (Jones 1999), selects from a full text document important sentences to serve as a summary of it. Luhn (1958) initiated the research work on automated text summarization with a statistical approach based on term (i.e., keyword) frequency and term normalization. Since then, various text summarization techniques, differing in their criterion (referred to as features) for measuring significance of sentences in a document and in their underlying summarization methods, have been proposed. In the following, features and underlying methods for text summarization are summarized.

2.2.1 Features for Text Summarization

Commonly used features for measuring importance of sentences in a document include:

- Thematic word (or keyword): Thematic words of a document are most frequent words in the document, but occur rarely in the overall collection. Thematic words generally communicate the theme discussed in a document. Thus, if a sentence in a document contains more thematic words, the sentence is more likely to be important (Luhn 1958; Edmundson 1969; Kupiec et al. 1995; Teufel & Moens 1997; Mani & Bloedorn 1998; Myaeng & Jang 1999; Neto et al. 2000).
- Cue phrase: This feature is based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words. According to previous research, cue phrases can be categorized into positively relevant, negatively relevant, and irrelevant ones. Generally, the weight of a sentence with respect to the cue phrase feature is calculated according to the match of the words in the sentence with the cue dictionary (Edmundson 1969; Kupiec et al. 1995; Teufel & Moens 1997; Neto et al. 2000).
- Title and heading word: Title and headings of a document are usually good indicators of what the document is about. The weight of a sentence is determined based on the match of the words in the sentence with the title and heading words (Edmundson 1969; Teufel & Moens 1997; Myaeng & Jang 1999).
- Location: The location of a sentence may signify its significance (Baxendale 1958). Prior research studies have introduced their views on locations of important sentences within a document (Edmundson 1969; Kupiec et al. 1995; Teufel & Moens 1997; Mani & Bloedorn 1998; Myaeng & Jang 1999). For example, Edmondson (1969) suggests that sentences in the first and last paragraphs and the first and last sentences of each paragraph should be assigned higher weights than other sentences in a document, while Kupiec et al. (1995) give more weights to the first ten paragraphs and last five paragraphs in a document.
- Sentence length: This feature is proposed based on the hypothesis that short sentences tend to be excluded in summaries. Given a pre-specified threshold (e.g., 5 words), this

feature is true for all sentences whose length is larger than the threshold, and false otherwise (Kupiec et al. 1995; Teufel & Moens 1997).

- *Cohesion*: This feature considers how central each sentence is with respect to the source document and measures the similarity between a sentence and the rest of the document in which it appears (Mani & Bloedorn 1998; Myaeng & Jang 1999; Neto et al. 2000). Sentences not essential for a summary typically present low cohesion.
- *Proper noun or uppercase word*: This feature is proposed based on the hypothesis that occurrence of proper nouns (e.g., people, places, and organizations) or, more generally, uppercase words represent clues of positive relevance of a sentence for the summary, especially in news texts (Neto et al. 2000; Kupiec et al. 1995).
- *Anaphor*: Occurrence of anaphors usually indicates the presence of additional information, not essential for the contents of the summary (Neto et al. 2000).

2.2.2 Methods for Text Summarization

Existing text summarization techniques employ different methods for generating a summary from a full text document. They mainly include linear function, Naïve Bayes, decision tree, and decision rule.

1. *Linear function*: The overall weight of a sentence in a document is determined as a linear function of selected features. Assume that h features $f_1, f_2, \ldots, f_h$ be the selected variables for measuring significance of sentences in a document. The weight $w(s)$ of a sentence s is defined as $w(s) = w_1f_1 + w_2f_2 + \ldots w_hf_h$, where $w_1, w_2, \ldots, w_h$ are the weights for $f_1, f_2, \ldots, f_h$, respectively. In this method, sentences with an overall weight higher than a pre-specified threshold or the k sentences with highest weights will be chosen as a summary for the document. Edmundson (1969) and Mani & Bloedorn (1998) employ this method for text summarization.
2. *Naïve Bayes*: Given h features $f_1, f_2, \ldots, f_h$, the probability that a sentence s will be included in a summary S is computed based on the following Bayes rule:
$$p(s \in S \mid f_1, f_2, \ldots, f_h) = \frac{p(f_1, f_2, \ldots, f_h \mid s \in S) p(s \in S)}{p(f_1, f_2, \ldots, f_h)}$$
Assuming statistical independence of the features, the Bayes rule is transformed into:
$$p(s \in S \mid f_1, f_2, \ldots, f_h) = \frac{\left(\prod_{j=1}^{h} p(f_j \mid s \in S)\right) p(s \in S)}{\prod_{j=1}^{h} p(f_j)}$$
$p(s \in S)$ is a constant, while $p(f_j \mid s \in S)$ and $p(f_j)$ can be estimated directly from the training set by counting occurrences. Subsequently, the sentences with highest probabilities are selected as a summary for the target document. Techniques proposed by Kupiec et al. (1995), Teufel & Moens (1997), Myaeng & Jang (1999) and Neto et al. (2000) adopt the Naïve Bayes approach as their underlying method for text summarization.
3. *Decision tree induction*: A decision-tree-based classification method is a supervised learning method that constructs a decision tree from a set of training examples (Quinlan 1986; Quinlan 1993). The decision tree starts as a single node containing all training examples. If the training examples are all of the same class, the node becomes a leaf and is labeled with that class. Otherwise, the algorithm selects a "best" attribute according to some metric (e.g., information gain) and grows the decision tree accordingly. Neto et al. (2000) and Mani & Bloedorn (1998) employ this approach for learning text summarization model.
4. *Decision rule induction*: Rule induction methods attempt to find a compact "covering" rule set that completely partitions the training examples into their correct classes

(Michalski et al. 1986; Clark & Niblett 1989). The covering set is found by searching heuristically for a single best rule that covers training examples for only one class. Having found a best conjunctive rule for a class, the rule is added to the rule set, and the training examples satisfying it are removed from further consideration. The process is repeated until no training examples remain to be covered. The learning approach is adopted by the text summarization technique proposed by Mani & Bloedorn (1998).

3. Development of Summary-based Event Detection (SED) Technique

We detail in this section the proposed Summary-based Event Detection (SED) technique. As shown in Figure 1, the overall process of SED is composed of two phases including news summarization phase and event detection phase. In the news summarization phase, a text summarization technique is employed to select representative sentences as the summary of a new news story or each past (i.e., historical) news story, while in the event detection phase, the summary of the new news story together with those of all past news stories are used for event detection.

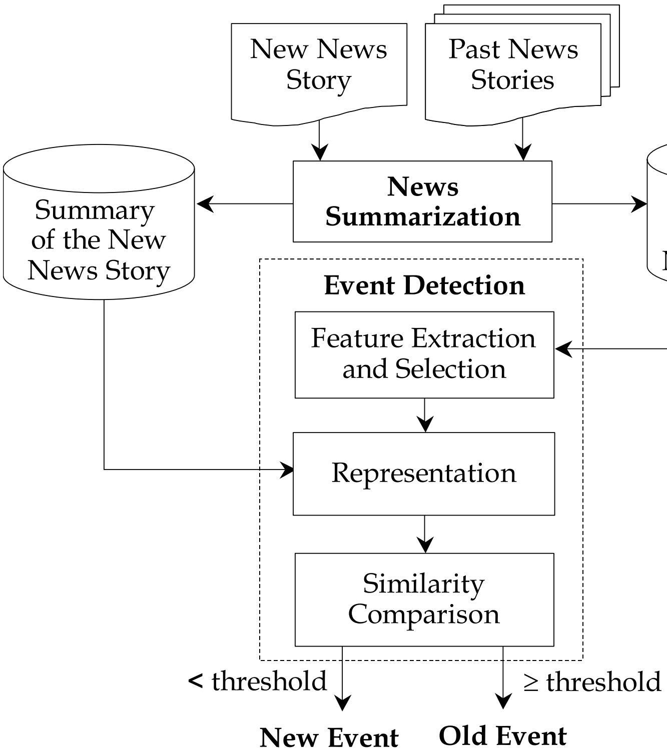

Figure 1: Overall Process of the SED Technique

3.1 News Summarization Phase

As mentioned, the goal of the news summarization phase is to select representative sentences from a news story as its summary. As shown in Figure 2, the process of the news summarization phase consists of two tasks: 1) news summarization learning that involves the induction of a summarization model from a set of training examples, and 2) news summary generation that actually generates a summary for a news story based on the summarization model induced in the previous task.

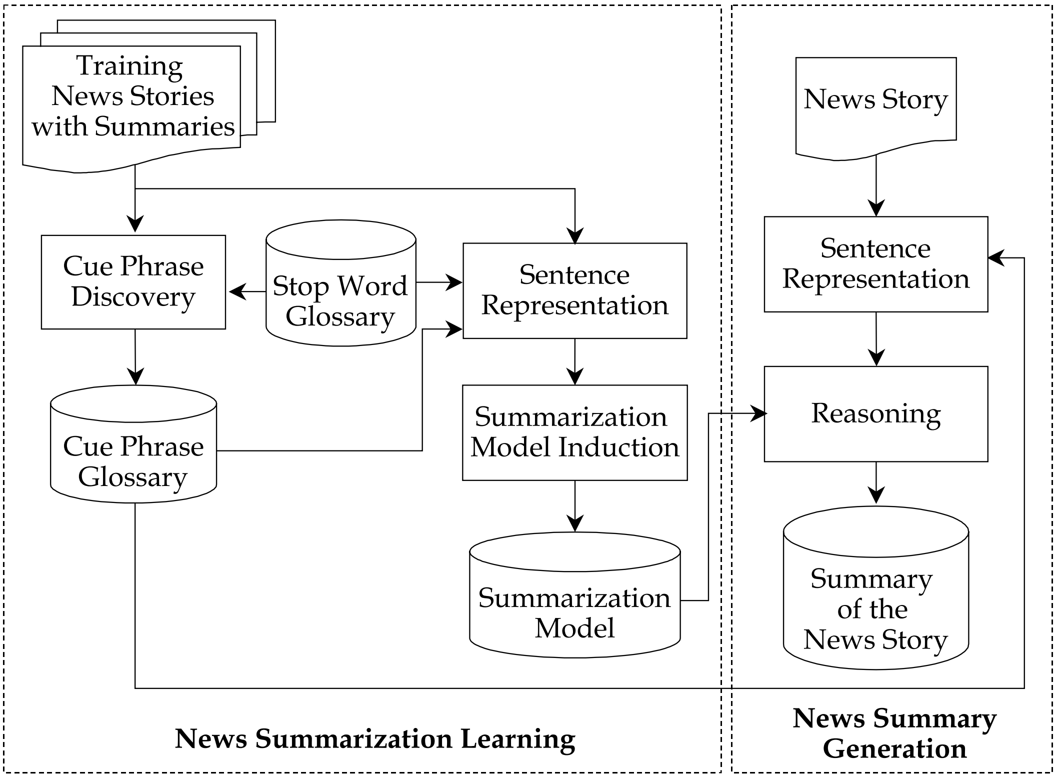

Figure 2: Process of News Summarization Phase

3.1.1 News Summarization Learning Task

A fundamental issue in the news summarization learning task is to select appropriate features for measuring significance of sentences. Based on the text summarization literature reviewed in the previous section, in this study, we identify and use for representing each sentence in the training news stories. Descriptions of these features are discussed in the following and their alternatives representation schemes are shown in Table 1.

<u>Cue phrase feature</u>: Cue phrases are classified into positive and negative ones in this study. We take a similar approach proposed by Edmundson (1969) for discovering a set of cue phrases from the training news stories. Since adjectives or adverbs seem to be more relevant to positive or negative cue phrases, all adjectives and adverbs in training new stories are parsed, using the rule-based part-of-speech tagger (Brill 1992), and include as candidate cue phrases. Subsequently, two statistics, *support ratio* and *selection ratio*, are computed for each candidate *cp*. The support ratio is defined as $\frac{ns(cp)}{ns}$, where *ns* is the number of sentences in the training corpus and *ns*(*cp*) is the number of sentences in the training corpus where *cp* appears. On the other hand, the selection ratio is defined as $\frac{nss(cp)}{ns(cp)}$ where *nss*(*cp*) is the number of sentences in the summaries (in the training corpus) where *cp* appears. For each candidate *cp* whose support ratio exceeds a pre-specified support-threshold ρ_{minsup}, *cp* is considered as 1) a positive cue phrase if its selection ratio is greater than a pre-specified upper-selection-threshold $\rho_{u\text{-}selection}$ and 2) a negative cue phrase if its selection ratio is no greater than a pre-specified lower-selection-threshold $\rho_{l\text{-}selection}$.

<u>Thematic word feature</u>: Thematic words of a document are typically nouns or noun phrases. Hence, we identify nouns and noun phrases for each training news story using the rule-based part-of-speech tagger. A standard TF×IDF method is then used to measure the weight of each

word in every training news story. Accordingly, the top n_{tw} words with highest TF×IDF weights in each training news story are selected as its thematic words.

Title word feature: As mentioned, words occurring in the title are usually considered as important indicators for measuring importance of sentences in a document. In this study, all stop words occurring in the title are removed first, and the remaining title words of a news document form the title glossary for the news story.

Location feature: Paragraphs that are closer to the beginning or ending of a news story tend to be more content-loaded and are useful for a summary. For this reason, a sentence in a news document is classified according to its appearance in the first, middle, or last third of paragraphs in the document.

Cohesion feature: Sentences not essential for summary present low cohesion with the document (Mani & Bloedorn 1998; Myaeng & Jang 1999; Neto et al. 2000). Hence, we employ this feature to measure the degree of connectivity between sentences. As with the measure proposed by Neto et al. (2000), the cohesion of a target sentence *s* in the news document *d* is measured as $COH_s = \sum_{s' \in d \text{ and } s' \neq s} \frac{sim(s, s')}{n-1}$, where *n* is the number of sentences in d.

Sentence length feature: Sentences with too many or too few words tend not to be included in summaries (Kupiec et al. 1995; Teufel & Moens 1997). The length of a sentence is the number of words appearing in the sentence.

Table 1: Representation Schemes Employed by the SED Technique

Features	Sentence Representation Schemes	Description
Positive Cue Phrase	Binary	1 if the sentence contains any positive cue phrases; 0 otherwise.
	Top-*N*	1 for top- N scoring sentences; 0 otherwise.
	P_c / L_s	Number of positive cue phrases in the sentence (P_c) divided by the length of the sentence (L_s).
Negative Cue Phrase	Binary	1 if the sentence contains any negative cue phrases; 0 otherwise.
	Top-*N*	1 for top- *N* scoring sentences; 0 otherwise.
	N_c / L_s	Number of negative cue phrases in the sentence (N_c) divided by the length of the sentence (L_s).
Thematic Word	Binary	1 if the sentence contains any thematic words; 0 otherwise.
	Top-*N*	1 for top- *N* scoring sentences; 0 otherwise.
	N_k / L_s	Number of thematic words in a sentence (N_k) divided by the length of the sentence (L_s).
Title Word	Binary	1 if the sentence contains any title words; 0 otherwise.
	Top-*N*	1 for top- *N* scoring sentences; 0 otherwise.
	N_t / L_s	Number of title words in the sentence (N_t) divided by the length of the sentence (L_s).
Sentence Location	{F, M, L}	The sentence occurs in first (F), middle (M), or last (L) third of paragraphs in the new story.
Cohesion	Binary	1 if the sentence COH_s is greater than average COH_s; 0 otherwise.
	COH_s	Sum of similarities with all other sentences divided by (number of sentences in the news story – 1).
Sentence Length	Binary	1 if the sentence length is greater than a pre-specified threshold; 0 otherwise.
	L_s / $MaxL_s$	Number of words occurring in the sentence (L_s) divided by the number of words occurring in the longest sentence of the document ($MaxL_s$).

3.1.2 News Summary Generation Task

The news summary generation task is to select relevant sentences to be included in the summary. Previous research studies suggested that C4.5 (a decision induction algorithm) has outperformed other techniques (e.g. Naïve Bayes, SCDF, and AQ15c) in text summarization (Mani & Bloedorn 1998; Neto et al. 2000). Hence, we adopt C4.5 as the underlying learning method in this study. However, the summarization model induced by C4.5 can only arrive at the dichotomous classification. That is, the prediction for a sentence can be either important (i.e., highly relevant in defining the event of the news story) or not important (i.e., less relevant or even irrelevant in defining the event of the news story). To have the capability of adjusting the length of a summary by specifying a desirable compression ratio, the prediction mechanism of C4.5 is extended. We apply the Laplacian accuracy (Clark and Boswell 1991) to estimate the accuracy of each decision path in the decision tree induced. The Laplacian accuracy of a decision path is defined as $\frac{n_c + 1}{n_{tot} + n_d}$, where n_d is the number of decision classes, n_c is the number of the training examples in the predicted class covered by the path, and n_{tot} is the total number of training examples covered by the path. When a sentence in a news story is covered by a decision path of the summarization model, the Laplacian accuracy of this decision path is used as the weight of the sentence. Accordingly, given a compression ratio, a desired number of sentences will be selected from a news story based on the their respective weights.

3.2 Event Detection Phase

As shown in Figure 1, given a summary for a newly arrived news story and the summaries of all past ones, a traditional event detection algorithm is employed for identifying whether the new news story discusses an old event or a new event. In this study, we adopt and implement the INCR technique for event detection purpose. In the feature extraction and selection step, a rule-based part of speech tagger (Brill 1992; Brill 1994), a noun phrase parser (Voutilainen 1993), and then the TF×IDF feature selection method are employed for extracting and selecting features from all summaries of the past news stories. After a set of representative features is selected, the new news summary and all past news summaries are represented using the incremental version of the TF×IDF representation scheme adopted in INCR. Subsequently, the feature vector for the new news summary is compared with all past news summaries, using the INCR's similarity measure that combines lexical similarity and temporal proximity. Without loss of generality, in this study, the INCR's similarity function is modified by changing the time window from the number of news stories to the number of days (i.e. set i as the number of days between x and the most recent document in c, and m as the time window measured in number of days before x). Finally, if the maximal similarity between the new news story and any past news stories is no less than t_n, the target news story is flagged as discussing an old event, otherwise a new event.

4. Empirical Evaluation

4.1 Evaluation Design

We evaluate the effectiveness of the proposed SED technique, using that of a traditional event detection technique (i.e., INCR) as a performance benchmark. For the evaluation purpose, news stories from November 1999 to December 1999 were collected from a news website, excite.com. Six event topics and 506 news stories pertaining to these event topics were manually identified. Events included in each news story were also coded manually. For obtaining correct news summaries, a senior researcher who is familiar with this data corpus participated in summarizing each news story in November 1999 into a couple of highly relevant and highly irrelevant sentences. The profile of the data corpus and that of the news summary are provided in Table 2 and in Table 3, respectively.

Table 2: Profile of Data Corpus

Event Topic	Number of Events	Number of News Stories	Average Number of Words per News Story
Adjustment of Interest Rate	12	26 (16/10)*	548
Initial Public Offering	7	7 (6/1)	444
Business Merger	169	238 (110/128)	502
New Product Announcement	72	83 (31/52)	523
Businesses Partnership	77	84 (32/52)	522
Computer Virus	9	68 (25/43)	428
Total	346	506 (220/286)	494

*: (16/10) denotes 16 news stories from November 1999 and 10 from December 1999.

Table 3: Summary of News Summarization

Event Topic	Average Number of Sentence per News Story	Average Number of Highly Relevant Sentences per News Story	Average Number of Highly Irrelevant Sentences per News Story
Adjustment of Interest Rate	21.50	2.19	7.63
Initial Public Offering	16.67	1.83	9.83
Business Merger	20.25	2.81	8.97
New Product Announcement	21.68	3.68	7.97
Business Partnership	21.91	2.94	11.50
Computer Virus	21.64	4.80	9.08
Total	20.61	3.04	9.16

The effectiveness of an event detection technique is measured by miss and false alarm rates. The miss rate is defined as the percentage of that an event detection technique fails to detect a new event, while the false alarm rate is defined as the percentage of that an event detection technique fails to detect an old event. In the context of supporting environmental scanning, a low miss rate may improve an organization's responsiveness to the changes of its external environment and therefore can enhance the organization's adaptability to its environment. On the other hand, an improvement in the false alarm rate reduces an organization's load in filtering news stories containing known events. Because of ever-increasing complexity and dynamics of an organization's environment, responsiveness and adaptability of the organization clearly are more essential than efficiency of environmental scanning. In this light, event detection should aim at achieving the lowest attainable miss rate while maintaining false alarm rate at a satisfactory level (Wei and Lee 2004).

4.2 Parameter Tuning

In this subsection, we report the parameter tuning experiments and results for INCR and SED. For the INCR technique, three parameters are involved, including the number of features k, time window w, and novelty threshold t_n. Specifically, k ranging from 50 to infinite (i.e., k = 50, 100, 150, 200 and infinite), w ranging from 7 to 60 (w = 7, 14, 30 and 60 days), and the novelty threshold t_n ranging from 0.01 to 0.5 at increments of 0.01 were examined. Among all tuning experiments conducted, when setting w as 60 and k as 150, INCR achieved the best performance at the novelty threshold of 0.18 (where the minimal Euclidean distance to the origin was attained). Hence, we adopted these values for w and k for the INCR technique in subsequent experiments.

For the parameter tuning of news summarization phase of the proposed SED technique, we set ρ_{minsup} as 0.02, $\rho_{u\text{-}selection}$ as 0.3, and $\rho_{l\text{-}selection}$ as 0.25 when producing the positive and negative cue phrase glossaries. We investigated the number of thematic words n_{tw} ranging from 10 to 20 at increments of 5, different representation schemes for each feature (as shown in Table 1). Among all tuning experiments for SED, the best learning effectiveness of news summarization phase was achieved when n_{tw} as 20, the top-3 representation scheme for the positive and negative cue phrase features, the standardized score representation scheme for the thematic word and title word features, the binary representation scheme for the cohesion feature, and the binary representation scheme with the sentence-length threshold of 5 for the sentence length feature (due to space limitation, the detailed tuning results are not shown here). Therefore, these parameter values or representation schemes were employed for subsequent experiments.

The event detection phase of the SED technique involves four parameters, including the number of features k, time window w, novelty threshold t_n and compassion ratio *cp-ratio*. We first set *cp-ratio* as 25% and investigated effects of w ranging from 7 to 60 (w = 7, 14, 30, to 60 days), t_n ranging from 0.01 to 1.0 at increments of 0.01, and k ranging from 25 to infinite (i.e., k = 25, 50, 75, 100, and infinite). The SED technique arrived at the best performance when w = 60, k = infinite, and t_n = 0.15.

4.3 Comparative Evaluation Results

The detection effectiveness of INCR and that of SED were compared using the parameter values and representation schemes (in the case of SED) determined in the previous tuning experiments. The data corpus was divided into two sets: historical (including news stories in November1999) and testing (including news stories in December 1999). Since the SED technique requires inducing a summarization model, all of historical data set was also used for the news summarization learning purpose. To expand the number of trials, 70% of news events were randomly selected from the historical and the testing data set, respectively, and that was repeated 30 times. The overall detection effectiveness was then estimated by averaging the performance across all trials.

We investigated the novelty threshold t_n for INCR and SED, ranging from 0.01 to 1 at increments of 0.01. As shown in Figure 3, as the compression ratio increased from 25% to 75%, the detection error tradeoff curves of SED generally shifted toward the origin. Moreover, at any level of false alarm rate that was lower than 10%, the proposed SED technique across the three different compression ratios achieved lower miss rate than INCR did. Furthermore, the false alarm rate attained by SED was comparable to that by INCR when the miss rate was lower than 10% and the compression ratio was 50% or 75%. These comparative evaluation results suggested that the proposed SED technique was comparable to or even outperformed its counterpart (i.e., when full text documents rather than summaries were used for event detection).

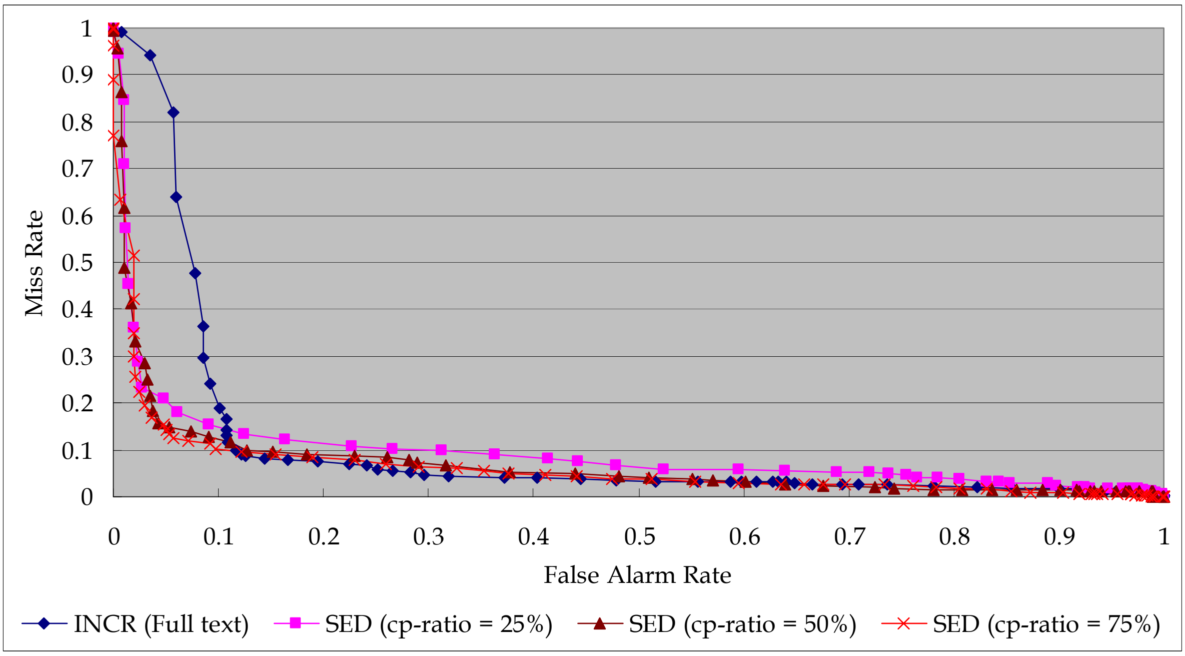

Figure 3: Detection Error Tradeoff Curves of Different Event Detection Techniques

5. Conclusions and Future Research Directions

In this study, we developed a summary-based event detection (SED) technique that filtered less relevant sentences or paragraphs in a news story before performing feature-based event detection. Using the data corpus collected, our empirical evaluation results showed that our proposed SED technique could achieve comparable or even better detection effectiveness than the benchmark. Some future research directions related to this study should be continued. First, in this study, the news summarization learning of the SED technique employs a decision tree induction approach (i.e., C4.5). Adoption of other induction techniques (e.g., Naïve Bayes or backpropagation neural network) by the news summarization learning of SED should be conducted and empirically evaluated. Moreover, an empirical evaluation that involves a larger data set with more news stories and event topics is one of our future research directions. Finally, in this study, we focus only on event detection for supporting environmental scanning. To supporting another challenging task in organizational scanning of external environments—event tracking, the development of an appropriate event tracking method based on the proposed SED technique would be desirable.

Acknowledgment

This work was supported by National Science Council of the Republic of China under the grant NSC 91-2416-H-110-014.

References

Allan, J., Papka, R. and Lavrenko, V., "On-line New Event Detection and Tracking," *Proceedings of SIGIR '98: 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 37-45.

Baxendale, P.B., "Machine-made Index for Technical Literature–An Experiment," *IBM Journal of Research and Development* (2:4), 1958, pp. 354-361.

Brill, E., "A Simple Rule-Based Part of Speech Tagger," *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992.

Brill, E., "Some Advances in Rule-Based Part of Speech Tagging," *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1994.

Choo, C. W., "Information Management for the Intelligent Organization: The Art of Scanning the Environment, 2ed.," *Information Today*, Medford, NJ, 1998.

Clark, P. and Boswell, R., "Rule Induction with CN2: Some Recent Improvements," *Proceedings of the Fifth European Conference on Machine Learning*, 1991, pp. 151-163.

Clark, P. and Niblett, T., "The CN2 Induction Algorithm," *Machine Learning* (3:4), 1989, pp. 261-283.

Daft, R.L., Sormunen, J., and Parks, D., "Chief Executive Scanning, Environmental Characteristics and Firm Performance: An Empirical Study," *Strategic Management Journal* (9), 1988, pp. 123-139.

Edmundson, H. P., "New Method in Automatic Extraction," *Journal of the ACM* (16:2), 1969, pp. 264-285.

Jain, S. C., "Environmental Scanning–How the Best Companies Do It," *Long Range Planning*, 1984, pp. 117-128.

Jones, K. S., "Automatic Summarizing: Factors and Directions," in *Advances in Automatic Text Summarization*, I. Mani and M. Maybury (eds.), MIT Press, Cambridge, MA, 1999.

Kupiec, J., Pedersen, J., and Chen, F., "A Trainable Document Summarizer," *Proceedings of the 18th ACM-SIGIR Conference*, 1995, pp. 68-73.

Luhn, H. P., "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, 1958, pp. 159-165.

Mani, I. and Bloedorn, E., "Machine Learning of Generic and User-Focused Summarization," *Working Notes of the AAAI'98 Spring Symposium on Intelligent Text Summarization,* Stanford, CA, 1998, pp. 69-76.

Myaeng, S. H. and Jang, D. H., "Development and Evaluation of a Statistically-Based Document Summarization System," in *Advances in Automatic Text Summarization*, I. Mani and M. Maybury (eds.), MIT Press, Cambridge, MA, 1999.

Michalski, R. S., Mozetic, I., Hong, J. and Lavrac, N., "The Multipurpose Incremental Learning System AQ15 and Its Testing Application to Three Medical Domains," *Proceedings of AAAI-86*, 1986, pp. 1041-1045.

Maier, J. L., Rainer, R. K., Snyder, C. A., "Environmental Scanning for Information Technology: An Empirical Investigation," *Journal of Management Information Systems* (14:2), 1997, pp. 177-200.

Mason, D. H. and Wilson, R. G., "Future Mapping: A New Approach to Managing Strategic Uncertainty," *Planning Review*, 1987, pp. 20-29.

Neto, J. L., Santos, A. D., Kaestner, C. A. A., Freitas, A. A., Nievola, J. C., "A Trainable Algorithm for Summarization News Stories," *Proceeding PKDD'2000 Workshop on Machine Learning and Textual Information Access*, Lyon, France, September, 2000.

Quinlan, J. R., "Induction of Decision Trees," *Machine Learning* (1), 1986, pp. 81-106.

Quinlan, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

Starbuck, W., Greven, A., and Hedberg, B. L., "Responding to Crises," *Journal of Business Administration* (9), 1978, pp. 111-137.

Teufel, S. and Moens, M., "Sentence Extraction as a Classification Task," *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, July 1997, pp. 58-65.

Voutilainen, A., "Nptool: A Detector of English Noun Phrases," *Proceedings of Workshop on Very Large Corpora*, Ohio, June 1993.

Wheelen, T. and Hunger, J., *Concept in Strategic Management and Business Policy*, 8th ed., Pearson Education, Inc., 2002.

Wei, C. and Lee, Y. H., "Event Detection from Online News Documents for Supporting Environmental Scanning," *Decision Support Systems* (36:4), March 2004, pp. 385-401.

Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T., and Liu, X., "Learning Approaches for Detecting and Tracking News Events," *IEEE Intelligent Systems and Their Applications*, 1999, pp. 32-43.

Yang, Y., Pierce, T. and Carbonell, J., "A Study on Retrospective and Online Event Detection," Proceedings of SIGIR '98: 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, 1998, pp. 28-36.