Use of Text Summarization for Supporting Event Detection

研究生：吳寶鳳 撰
指導教授：魏志平 博士

中華民國 九十二 年 七 月
Abstract

Environmental scanning, which acquires and use the information about event, trends, and changes in an organization’s external environment, is an important process in the strategic management of an organization and permits the organization to quickly adapt to the changes of its external environment. Event detection that detects the onset of new events from news documents is essential to facilitating an organization’s environmental scanning activity. However, traditional feature-based event detection techniques detect events by comparing the similarity between features of news stories and incur several problems. For example, for illustration and comparison purpose, a news story may contain sentences or paragraphs that are not highly relevant to defining its event. Without removing such less relevant sentences or paragraphs before detection, the effectiveness of traditional event detection techniques may suffer. In this study, we developed a summary-based event detection (SED) technique that filters less relevant sentences or paragraphs in a news story before performing feature-based event detection. Using a traditional feature-based event detection technique (i.e., INCR) as benchmark, the empirical evaluation results showed that the proposed SED technique could achieve comparable or even better detection effectiveness (measured by miss and false alarm rates) than the INCR technique, for data corpora where the percentage of news stories discussing old events is high.

Keywords: event detection, text summarization, environmental scanning
中文提要

在資訊爆炸以及資訊流通快速的時代，組織所面臨之外部環境也隨之越趨複雜且變化快速，使得組織必須不斷地偵測其面臨的外部環境、及時反應和掌握環境的變化及趨勢。隨著網際網路和線上電子新聞的崛起，有關組織外部環境的資訊量也隨之增加，因此利用資訊科技來輔助組織進行環境掃描已成為組織策略管理中重要的一環。事件偵測技術為協助組織環境掃描的技術之一，其藉由比較新產生的新聞文件與過去的新聞文件之間的文字相似度，判定新產生的新聞文件所描述之新聞事件為已發生過或未發生過之事件。然而，一般的新聞文件中，記者為了使報導更加完整，會做額外的補充報導。然而，這些補充報導通常跟該新聞文件所欲描述的主題沒有高度相關性，且容易降低事件偵測的準確性。因此，本論文提出以文件摘要技術作為基礎的事件偵測技術，其結合了文件摘要技術，對於每一新聞文件萃取出跟主題有高度相關性的句子來代表每一篇新聞，在以傳統比較文字相似度的方法來判斷該新聞是否為未發生之新聞事件。以實際的新聞資料來做實驗評估本論文提出的事件偵測技術時，此技術能夠達到與傳統事件偵測技術相似或較好的準確率。

關鍵字: 事件偵測、文件摘要、環境掃描
# TABLE OF CONTENTS

Chapter 1 Introduction ...........................................................................................................1
  1.1 Background ....................................................................................................................1
  1.2 Research Motivation and Objective ...........................................................................2
  1.3 Organization of the Thesis .........................................................................................4

Chapter 2 Literature Review ..............................................................................................5
  2.1 Event Detection ..........................................................................................................5
  2.2 Text Summarization ...................................................................................................9
    2.2.1 Edmundson’s Approach ......................................................................................9
    2.2.2 Kupiec et al’s Approach .................................................................................11
    2.2.3 Teufel and Moens’ Approach ..........................................................................13
    2.2.4 Mani and Bloedorn’s Approach .......................................................................14
    2.2.5 Neto et al’s Approach .....................................................................................16
    2.2.6 Myaeng and Jang’s Approach .........................................................................19
    2.2.7 Summary of Text Summarization Approaches .................................................20

Chapter 3 Development of Summary-based Event Detection (SED) Technique ..................22
  3.1 Process of Summary-based Event-Detection (SED) Technique .................................24
  3.2 News Summarization Phase ......................................................................................25
    3.2.1 News Summarization Learning Task ...............................................................26
    3.2.2 News Summary Generation Task .....................................................................32
  3.3 Event Detection Phase ..............................................................................................33

Chapter 4 Empirical Evaluation .........................................................................................35
  4.1 Evaluation Design .......................................................................................................35
    4.1.1 Data Collection and Summary Preparation ......................................................35
II
LIST OF FIGURES

Figure 2.1: Algorithm of Agglomerative Clustering ..................................................... 17
Figure 3.1: Example of News Story ............................................................................. 23
Figure 3.2: Overall Process of the SED Technique ..................................................... 25
Figure 3.3: Process of News Summarization Phase .................................................... 26
Figure 4.1: Detection Error Tradeoff Curves of INCR (k=150) ................................. 40
Figure 4.2: Detection Error Tradeoff Curves of SED (k = Infinite, c = 25%) .......... 43
Figure 4.3: Detection Error Tradeoff Curves of Different Event Detection

Techniques .................................................................................................................. 44
Figure 4.4: Detection Error Tradeoff Curves of Different Event Detection

Techniques based on Three Event Topics ................................................................. 46
LIST OF TABLES

Table 2.1: Features Employed in Mani and Bloedorn’s Approach ........................................... 15
Table 2.2: Summary of Text Summarization Approaches ......................................................... 21
Table 3.1: Definitions of Variables .......................................................................................... 30
Table 3.2: Representation Schemes Employed by the SED Technique .................................... 31
Table 4.1: Summary of Data Corpus ....................................................................................... 36
Table 4.2: Summary of News Summarization ......................................................................... 37
Table 4.3: Sentence Representation Schemes for Summarization Features
  Investigated ......................................................................................................................... 42
Chapter 1
Introduction

1.1 Background
As an organization’s environment becomes more diverse, dynamic and complex, uncertainty faced by the organization increases. An organization’s competitive position, financial success, and even bottom-line survival depend on its ability to scan, understand, and ultimately adapt to its environmental conditions [L81]. Environmental scanning, a major vehicle for an organization to adapt to external environmental changes, is defined as a process of acquiring and using information about events, trends, and changes in an organization’s external environment in order to guide the organization’s future course of action [A67, FK77, C99]. Empirical research results suggest that environmental scanning is linked with improved organizational performance. For instance, Daft et al. [DSP88] found that chief executives of high-performing firms scanned the environment more frequently and more broadly than their counterparts in low-performing firms. Conversely, failure to scan has been associated with corporate decline and failure [SGH78]. Another empirical study conducted by Ahituv et al. [AZM98] indicated significant differences in the level of environmental scanning and in the use of information systems between firms that were more successful in introducing new products into the market and firms that were less successful. Accordingly, scanning the external business environment for events, trends, and changes from news services and websites 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 [C98].
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. As a result, the amount of information pertaining to an organization’s environment is exploded. 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 handle and has been a burden to managers. Thus, an information system that facilitates organizational scanning of external environments is essential. Specifically, the system needs to support 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 will mainly focus on event detection for supporting organizational environmental scanning.

1.2 Research Motivation and Objective

Event detection is to identify the onset of new events from streams of news stories [APL98, YPC98, YCB99]. An event is referred to something happening in a certain place at a certain time [APL98, YPC98, YCB99], while an event topic consists of a set of event instances of the same type. For example, SPSS Inc. acquiring NetGenesis Corp. is an event pertaining to the event topic of business merger. Traditional event detection techniques usually adopted the feature co-occurrence approach. It identifies whether a news story contains an unseen event by comparing the similarity of features between the news story and past news stories. 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 was often employed [YPC98]. Moreover, since a time gap between bursts of topically similar stories is often an indication of different events, the
incorporation of a time window for event scoping was also commonly adopted [YPC98].

Nevertheless, traditional feature-based event detection techniques incur several problems. First, for illustration and comparison purpose, 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. For example, some reporters may use “merger” or “purchase” to describe a particular business-merger event, while others may use “acquisition” for the same event. Moreover, two news stories discussing the same event may be oriented from different angles, resulting in differences in features. On the other hand, two news stories for different events may contain very similar feature sets since the events belong to the same event topic [WL01].

Motivated by the significance of event detection in supporting organizational environmental scanning and the need for improving effectiveness of event detection, this thesis 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 is the process of selecting from a full text document important sentences that serve as a summary of this full text. In a different viewpoint, text summarization can be used to identify and remove sentences or paragraphs irrelevant to the main theme of a target document and, hence, is an appealing approach for supporting event detection.
Specifically, we will propose an event detection technique, referred to as 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. In this vein, the first problem of traditional feature-based event detection techniques can be minimized, potentially resulting higher event-detection effectiveness. The proposed SED technique will empirically be evaluated, using a traditional event detection technique as benchmarks.

1.3 Organization of the Thesis
The remainder of the thesis is organized as follows. Chapter 2 reviews the literature related to this research, including existing event detection techniques and text summarization approaches. Chapter 3 details the proposed Summary-based Event-Detection (SED) technique. Chapter 4 describes an empirical evaluation, including evaluation design, evaluation criteria and discussions of empirical evaluation results. Finally, contributions of this research as well as some future research directions will be concluded in Chapter 5.
Chapter 2

Literature Review

In this chapter, we will review literature related to the development of the proposed Summary-based Event Detection (SED) technique. Specifically, existing event detection and text summarization will be depicted.

2.1 Event Detection

The objective of event detection is to identify stories in several continuous news streams that pertain to new or previously unidentified events [YPC98]. Existing event detection techniques can be classified into two forms: namely retrospective detection and online detection. The former entails the discovery of previously unidentified events in a chronologically ordered accumulation of news stories, and the latter strives to identify the onset of new events from live news feeds in real-time. Both forms of detection intentionally lack prior knowledge of novel events, but do have access to unlabelled historical news stories for use as contrast sets.

Most of existing event detection algorithms, retrospective or online, were developed based on the document clustering approach. Yang et al. [YPC98, YCB99] proposed two clustering methods for event detection: GAC and INCR. GAC, operating in a strict retrospective detection setting, performs agglomerative clustering for producing hierarchically organized document clusters. GAC employed the conventional vector space model to represent documents and clusters. Each document is represented using a vector of weighted terms, based on the TF×IDF (within-document term frequency ×
inverse document frequency) scheme:

\[
    w(t, d) = \left( \frac{1 + \log_2 tf(t, d)}{\|d\|} \right) \times idf(t)
\]

where \( w(t, d) \) is the weight of term \( t \) in document \( d \),
\( tf(t, d) \) is the within-document term frequency (TF),
\( idf(t) = \log_2 \left( \frac{N}{n_t} \right) \) is the inverse document frequency (IDF) of term \( t \),
\( N \) is the number of the training documents used to compute the IDF,
\( n_t \) is the number of training documents where \( t \) appears, and
\[
    \|d\| = \sqrt{\sum_t w(t, d)^2}
\]
is the 2-norm of vector \( \vec{d} \).

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. GAC is a divide-and-conquer version of a group-average clustering algorithm. Group-average clustering maximizes the average similarity between document pairs in the resulting clusters by merging clusters in a greedy, bottom-up fashion. To improve the computation efficiency and to preserve the characteristics that events tend to appear in news bursts, GAC adopted a divide-and-conquer strategy that grows clusters iteratively in a bottom-up fashion. 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 stories 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 stories about different events.

On the other hand, INCR, designed for both retrospective and online detection, is a
single-pass incremental clustering algorithm that produces nonhierarchica clusters incrementally [YPC98, YCB99]. For retrospective detection, the TF×IDF scheme was adopted to represent documents or clusters. However, for online detection, to deal with the problem of continuously incoming documents that might affect term weighting and vector normalization during online detection, the incremental IDF was employed by INCR:

\[ idf(t, p) = \log_2 \left( \frac{N(p)}{n(t, p)} \right) \]

where \( p \) is the current time point,
\( N(p) \) is the number of documents accumulated up to the current point (including the retrospective corpus if used), and
\( n(t, p) \) is the document frequency of term \( t \) at time \( p \).

Moreover, INCR incorporated a time penalty when determining the similarity between a news 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 \( m \) news documents before \( x \) is imposed) or a linear decaying-weight function shown as follows:

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

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 news document is absorbed by the most similar cluster in the past if the similarity between the target news document and the cluster is larger than a pre-selected clustering threshold \( (t_c) \); otherwise, the news document becomes the seed of a new cluster. For online detection, the novelty threshold \( (t_n) \) was introduced. If the maximal similarity between a new news document and any cluster in the past is no less than \( t_n \), the target
news document is flagged as discussing an old event.

Allan et al. [APL98] proposed another online event detection technique that is a modification of single-pass clustering algorithm. Each news document is represented as a query. The similarity between a news document \( d \) and an earlier query (i.e., a previous news document) \( q \) is determined as follows:

\[
eval(q,d) = \frac{\sum_{i=1}^{N} w_i \times d_i}{\sum_{i=1}^{N} w_i}
\]

where \( w_i \) is the relative weight of a query feature \( q_i \), and \( d_i \) is the belief that the feature’s appearance in \( d \) indicates relevance to \( q \).

The above-mentioned belief function is composed of a TF component and an IDF component. For any document \( d \) and a collection \( c \):

\[
d_i = \text{belief}(q_i, d, c) = 0.4 + 0.6 \times TF \times IDF
\]

where \( c \) is an auxiliary collection, independent of the news stream, for estimating \( IDF \),

\[
TF = t \left( t + 0.5 + 1.5 \times \frac{dl}{\text{avg}_d dl} \right),
\]

\[
IDF = \log \left( \frac{|c| + 0.5}{df} \right) / \log(|c| + 1),
\]

\( t \) is the number of times the feature \( q_i \) occurs in \( d \),
\( df \) is the number of documents where \( q_i \) appears in the collection \( c \),
\( dl \) is the length of \( d \), and
\( \text{avg}_d dl \) is the average document length in \( c \).

To determine whether a news story contains an unseen event, the news story is compared against all of earlier queries. If the news story triggers an existing query by exceeding a pre-specified threshold, the story is assumed to discuss the event
represented in the query; otherwise it contains a new event.

2.2 Text Summarization

With a lot of information pouring in everyday, text summaries are essential for users to alleviating from information overloading. Instead of having to go through the entire text, a concise summary allows users to understand the text quickly and easily. Jones [J99] defines text summarization 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.” In other words, the task of text summarization is to select from a full text document important sentences that will serve as a summary of this full text.

Research on automated text summarization has a long history, beginning with the work by Luhn [L58] dating back to the fifties. His work describes a statistical approach to text summarization based on term (i.e., keyword) frequency and term normalization and has had a considerable influence on this research field. Since then, many text summarization techniques have been proposed. These text summarization techniques differ in their criterion (referred to as features) in measuring significance of sentences in a document and their underlying summarization methods. In the next subsections, reviews of some representative text summarization techniques are provided.

2.2.1 Edmundson’s Approach

In addition to the key word feature used by Luhn’s work [L58], Edmundson [E69] proposed the use of three additional features for text summarization, including cue phrases, title and heading word, and sentence location. Specifically,
• **Key word feature:** This feature is based on the hypothesis that high frequency words are positive relevant. The words comprising a key glossary were selected by listing all words that are not in the cue dictionary (see the next discussion). The weight of a sentence with respect to the key word feature is estimated as the frequency of key words in the sentence.

• **Cue phrase feature:** This feature is proposed based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words such as “significant,” “impossible,” and “hardly.” A cue dictionary used to identify cue phrases in sentences comprises of three sub-dictionaries: (1) *bonus* words, which are positively relevant; (2) *stigma* words, which are negatively relevant; and (3) *null* words, which are irrelevant. 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.

• **Title and heading word feature:** The title feature is proposed based on the hypothesis that the author conceives the title as circumscribing the subject matter of a document. On the other hand, when the author partitions the document into major sections, he/she summarizes each section by choosing an appropriate heading [E69]. A title glossary is created and consists of all words in the title, sub-title and headings of the document. Subsequently, positive weights are assigned to the title glossary, where the title words will be assigned a weight relatively prime to the heading words. The weight of a sentence with respect to the title and heading feature is determined based on the match of the words in the sentence with the title and heading words.

• **Location feature:** The significance of a sentence may be indicated by its location [B58] and topic sentences tend to occur at the beginning or end of documents or paragraphs [E69]. Thus, in the Edmondson approach [E69], the
sentences in the first and last paragraphs and in the first and last sentences of
each paragraph are assigned higher weights than other sentences in a document.

The overall weight of a sentence in a document is determined as a linear function of
the four component weights described above:

\[ a_1C + a_2K + a_3T + a_4L \]

where \(a_1, a_2, a_3,\) and \(a_4\) are the parameters for the cue phrase, key word, title
and heading word, and location weights, respectively.

The sentences with an overall weight higher than a pre-specified threshold are
selected as a part of the summary for the document.

2.2.2 Kupiec et al’s Approach

Kupiec et al’s approach [KPC95] extracts sentences using a Naïve Bayes classifier.
Given \(k\) features \(F_1, F_2, \ldots, F_k\), the probability of a sentence \(s\) will be included in a
summary \(S\) is computed based on the following Bayes rule:

\[
P(s \in S \mid F_1, F_2, \ldots, F_k) = \frac{P(F_1, F_2, \ldots, F_k \mid s \in S)P(s \in S)}{P(F_1, F_2, \ldots, F_k)}
\]

Assuming statistical independence of the features, the Bayes rule is transformed
into:

\[
P(s \in S \mid F_1, F_2, \ldots, F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S)P(s \in S)}{\prod_{j=1}^{k} P(F_j)}
\]

\(P(s \in S)\) 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.
The Kupiec et al’s approach employs the following five features for determining the probability of a sentence to be included in a summary.

- **Sentence length feature**: This feature is proposed based on the hypothesis that short sentences tend not to be included 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.

- **Cue phrase feature**: Sentences containing any of a list of fixed phrase, mostly two words long such as “In conclusion”, or occurring immediately after a section heading containing a keyword such as “conclusions”, “results”, “summary”, and “discussion” are more likely to be in summaries. This feature is true for those sentences that contain any of pre-determined cue phrases (a total of 26 cue phrases were included), or that follow section heads that contain specific keywords.

- **Location feature**: The first ten paragraphs and last five paragraphs in a document are given more weights. Furthermore, sentences in a paragraph are distinguished according to whether they are paragraph-initial, paragraph-final and paragraph-medial.

- **Thematic word feature**: Thematic words of a document are most frequent words in the document. Any pair of words having the same frequency is resolved on the basis of word length. A small number of thematic words are then selected and each sentence is scored as a function of frequency of these thematic words. This feature is binary, depending on whether a sentence is presented in the set of highest scoring sentences.

- **Uppercase word feature**: This feature is computed similarly to the previous one, with the constraints that an uppercase thematic word is neither sentence-initial nor common abbreviations and begins with a capital letter. Sentences in which
such words appear first score twice as much as later occurrences.

Their empirical results, using 188 documents with summary sampled from 21 publications in the scientific/technical domain, suggested that a subset of features containing location, cue phrase and sentence length features could achieve the best results.

2.2.3 Teufel and Moens’ Approach

Extending Kupiec et al’s approach [KPC95], Teufel and Moens [TM97] employed a Naïve Bayes classifier for constructing a text summarization technique based on the following five features.

- **Cue phrase feature**: Teufel and Moens used a manually created list of 1670 negative and positive cues and indicator phrases or formulaic expressions. These cue phrases were then manually classified into 5 classes: a score of –1 means “very unlikely” and +3 means “very likely” to be included in a summary.

- **Location feature**: Paragraphs at the start and end of a document are likely to be useful for a summary. Those sentences in document peripheral sections receive non-zero scores, while the sentences in the middle of the document receive a 0 score.

- **Sentence length feature**: All sentences under 15 tokens including punctuation receive a 0 score, all sentences above the threshold a 1 score. This method is useful filtering out captions, titles, and headings.

- **Thematic word feature**: This feature concentrates on words (excluding stop-list words) that occur frequently in the document, but rarely in the overall collection. Therefore, a standard TF\times IDF method used to find thematic words. The 10 top-scoring words are chosen as thematic words. Sentence scores are then
computed as a weighted count of thematic word in sentence, averaged by sentence length. The 40 top-rated sentences are then assigned a score of 1, all others 0.

- **Title feature**: The score of a sentence with respect to the title feature is the mean frequency of title word occurrences (excluding stop-list words). The 18 top-scoring sentences receive the value of 1, all other sentences 0.

### 2.2.4 Mani and Bloedorn’s Approach

Mani and Bloedorn [MB98] employed three different machine learning algorithms, including Standardized Canonical Discriminant Function (SCDF) analysis [S97], C4.5-Rules [Q92], and AQ15c [WBM95], to produce generic and user-specific summaries. Features used in their proposed text summarization techniques are classified into three groups, including location, thematic word and cohesion features.

- **Location features**: Location features exploit the structure of the text at different levels of analysis. The exploited analysis levels for the location features are shown in Table 2.1.

- **Thematic word features**: Three metrics are used to measure the thematic weights of a term, namely TF, TF×IDF, and $G^2$ statistic (i.e., indicating the likelihood that the frequency of a term in a document is greater than what would be expected from its frequency in the corpus, given the relative sizes of the document and the corpus).

- **Cohesion features**: Text cohesion [HH96] involves relations between words or referring expressions, which determine how tightly connected the text is. Two cohesion-based features are adopted in the Mani and Bloedorn’s approach [MB98], including *synonymy* and *co-occurrence* based on bigram statistics. Co-occurrence scores between contentful words up to 40 words apart are
computed using a standard mutual information metric [F61]. The mutual
information (MI) between terms $j$ and $k$ in document $i$ is determined as:

$$\text{MI}(j, k, i) = \ln \left( \frac{N_i \times TF_{jk,i}}{TF_{ji,i} \times TF_{ki,i}} \right)$$

where $TF_{jk,i}$ is the maximum frequency of bigram $jk$ in the document $i$,
$TF_{ji,i}$ is the frequency of term $j$ in the document $i$, and $N_i$ is the total
number of terms in the document $i$.

Table 2.1: Features Employed in Mani and Bloedorn’s Approach

<table>
<thead>
<tr>
<th>Feature</th>
<th>Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sent-loc-para</td>
<td>{1, 2, 3}</td>
<td>sentence occurs in first, middle or last third of paragraph</td>
</tr>
<tr>
<td>para-loc-section</td>
<td>{1, 2, 3}</td>
<td>sentence occurs in first, middle or last third of section</td>
</tr>
<tr>
<td>Sent-special-section</td>
<td>{1, 2, 3}</td>
<td>1 if sentence occurs in introduction, 2 if in conclusion, 3 if in other</td>
</tr>
<tr>
<td>depth-sent-section</td>
<td>{1, 2, 3, 4}</td>
<td>1 if sentence is a top-level section, 4 if sentence is a subsubsubsection</td>
</tr>
<tr>
<td>Thematic Words Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sent-in-highest-tf</td>
<td>{1, 0}</td>
<td>average tf score (Filter 1)</td>
</tr>
<tr>
<td>Sent-in-highest-tf.idf</td>
<td>{1, 0}</td>
<td>average tf.idf score (Filter 1)</td>
</tr>
<tr>
<td>Sent-in-highest-G2</td>
<td>{1, 0}</td>
<td>average G2 score (Filter 1)</td>
</tr>
<tr>
<td>Sent-in-highest-title</td>
<td>{1, 0}</td>
<td>number of section heading or title term mentions (Filter 1)</td>
</tr>
<tr>
<td>Sent-in-highest-pname</td>
<td>{1, 0}</td>
<td>number of name mentions (Filter 1)</td>
</tr>
<tr>
<td>Cohesion Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sent-in-highest-syn</td>
<td>{1, 0}</td>
<td>number of unique sentences with a synonym link to sentence (Filter 1)</td>
</tr>
<tr>
<td>Sent-in-highest-co-occ</td>
<td>{1, 0}</td>
<td>number of unique sentences with a co-occurrence like to sentence (Filter 1)</td>
</tr>
</tbody>
</table>

Filter 1: sorts all the sentences in the document by the feature in discussion. It assigns
1 to the current sentence iff it belongs in top $c$ of the scored sentences, where $c =$
compression rate. As it turns out, removing this discretization filter completely, to use
raw scores for each feature, merely increases the complexity of learnt rules without
improving performance.
The three different machine learning techniques for text summarization were empirically evaluated. The empirical results suggested the C4.5 technique achieved the best summarization effectiveness.

2.2.5 Neto et al’s Approach

Rino [RS94] and Weissberg [WB90] observed that certain types of texts, for instance, news articles, technical reports, research papers, etc., conform to a set of style and organization constraints, called the Discourse Macro Structure (DMS), which help authors to achieve a desired communication effect. The components in news reports generally can be classified into one of the two categories: the “what's-the-news” category and the optional “background” category. The background, if present, covers previous event and supplies the context necessary to understand the central story or to make a follow up story self-contained. When the background is a common knowledge or is implied in the main news section, it usually can be omitted. On the other hand, the what's-the-news covers the new developments and the news facts that make the news [SSW99]. Based on this observation, the text summarization approach proposed by Neto et al. [NSK00] employs a text segmentation method [Y97] for detecting sentences that are either relevant (i.e., belonging to the what's-the-news category) or background (i.e., belonging to the background category). Specifically, the text segmentation method (see Figure 2.1), extended from a hierarchical agglomerative clustering algorithm, successively grows “coherent” segments by appending lexically related paragraphs.
(1) Partition the text to elementary segments
(2) While more than one segment left do
    Apply a proximity test to find the two most similar consecutive segments, $s_i$ and $s_{i+1}$.
    Merge $s_i$, $s_{i+1}$ into one segment
End While

Figure 2.1: Algorithm of Agglomerative Clustering

Subsequently, their text summarization approach employs the following features for representing and evaluating the significance of sentences in a document. These features include thematic word, proper noun, anaphor, discourse marker in the beginning of sentence, cohesion (i.e., connectivity of sentences), sentence depth in the tree, and position in the tree.

- **Thematic word feature:** TF and TF×IDF are the two metrics used to measure the thematic weights of a term (typically nouns). Fifteen terms with largest metric scores are selected as thematic words for the target document.

- **Proper noun feature:** This feature is proposed based on the hypothesis that occurrence of proper nouns (e.g., people, places, and organizations) represent clues of positive relevance of a sentence for the summary, especially in news texts.

- **Anaphor feature:** Occurrence of anaphors usually indicates the presence of additional information, not essential for the contents of the summary. Specifically, certain nouns and expressions that occur in the beginning of sentences (e.g., in the first six words) are used for detecting anaphors.

- **Discourse marker feature:** Some markers that frequently occur in the beginning of discourse such as “Because”, “Furthermore”, and “Additionally” are considered indicators of the presence of additional information, not essential
for the summary. About 150 common markers are listed and used in the text summarization system proposed by Neto et al. [NSK00].

- **Cohesion Feature:** There is evidence that sentences not essential for the summary present low cohesion. The method of computing cohesion is similar to Text Relationship Maps proposed by [MSB97]. It is to compute the similarity between any pair of sentences. Total similarity value of each sentence is the sum of the individual similarity between that sentence and each of other sentences, and is normalized by dividing the value computed in the previous step by the largest total similarity value among all sentences.

- **Sentence depth in the tree feature:** This feature indicates the depth of the sentence in the tree produced by the text segmentation method, normalized by the entire tree depth. Sentences located at shallow levels (close to the root) of the tree are associated with a low degree of cohesion with the rest of the document and, hence, probably represent non-essential information.

- **Position in the tree feature:** Instead of relying on structural information available only in semi-structured texts, Neto et al. [NSK00] decided to obtain locational information from the tree produced by the text segmentation method. The path from the root of the tree to the target sentence is considered, and only the first four levels of the tree are taken into account. For each level 4 nodes. Possible values for each level of the path are left, right, and none.

Neto et al. [NSK00] implemented two learning algorithms, Naïve Bayes and C4.5, for text summarization based the features described above. Their empirical results show that C4.5 outperformed Naïve Bayes. In addition, the empirical results also suggest that three out of the seven features have a predictive accuracy, namely proper nouns, cohesion, and thematic words.
2.2.6 Myaeng and Jang’s Approach

The approach developed by Myaeng and Jang [MJ98] constructs a summary by using a Naïve Bayes classifier. Each sentence is considered as a part of summary by measuring the probability of cue word, negative word, position, thematic word, title, and centrality.

- **Cue word feature:** Given a cue word $CW_i$, the probability that a sentence $s$ will be included in a summary $S$ is calculated using Bayes’ rule as follows:
  \[
  P(s \in S \mid CW_i) = \frac{P(CW_i \mid s \in S)P(s \in S)}{P(CW_i)}
  \]
  where $P(CW_i \mid s \in S)$ is the probability that a cue word appears in a summary sentence and calculated by counting the number of the cue word in all the summary sentences in the training corpus.

- **Negative word feature:** The most frequent words included in the non-summary sentences are defined as negative words. A phrase such as “For example” belongs to this category and serves as negative evidence. Given a negative word phrase $NW_i$, the probability that a sentence $s$ will not be included in a summary $S$ is computed as follows:
  \[
  P(s \not\in S \mid NW_i) = \frac{P(NW_i \mid s \not\in S)P(s \not\in S)}{P(NW_i)}
  \]

- **Position feature:** Myaeng and Jang [MJ98] observed that sentences in the final part of an introduction section or in the first part of a conclusion section are more likely to be included in a summary than those in other parts of the sections. Thus, six different positions can be assigned to individual sentences that are in the introduction and conclusion sections. The probability that a sentence appearing in one of the six regions $P_i$ is expressed as:
  \[
  P(s \in S \mid P_i) = \frac{P(P_i \mid s \in S)P(s \in S)}{P(P_i)}
  \]
  where $P(P_i \mid s \in S)$ and $P(P_i)$ are probability of observing a summary
sentence in position $P_i$ and of observing a summary sentence, respectively.

- **Thematic word feature:** It was intuitively appealing to consider only those sentences with strong keywords. The more important keywords are included in a sentence, the higher score it receives.

- **Title feature:** This feature is concerned about how similar a sentence is to the title of the source document. Because the title and heading are usually good indicators of concept of what is in the document. A similarity measure between a source sentence and title is calculated by the amount of overlap between the two in terms of the keywords.

- **Centrality Feature:** 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. The centrality value of a sentence can be calculated by comparing the vector representation of the sentence and that of the document.

The experimental results conducted by Myaeng and Jang [MJ98] suggested that cue word, position, and title resemblance features are most useful to filter irrelevant sentences.

### 2.2.7 Summary of Text Summarization Approaches

In this section, a summary of the text summarization approaches discussed previously is provided, based on such dimensions as features employed, underlying methods for producing summaries, and predictive feature subset suggested empirically.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Features</th>
<th>Underlying Method</th>
<th>Predictive Feature subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edmundson [E69]</td>
<td>• Key word</td>
<td>Linear weighted function</td>
<td>• Cue phrase</td>
</tr>
<tr>
<td></td>
<td>• Cue phrase</td>
<td></td>
<td>• Title and heading word</td>
</tr>
<tr>
<td></td>
<td>• Title and heading word</td>
<td></td>
<td>• Location</td>
</tr>
<tr>
<td></td>
<td>• Location</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kupiec et al. [KPC95]</td>
<td>• Sentence length</td>
<td>Naïve Bayes classifier</td>
<td>• Location</td>
</tr>
<tr>
<td></td>
<td>• Cue phrase</td>
<td></td>
<td>• Cue phrase</td>
</tr>
<tr>
<td></td>
<td>• Location</td>
<td></td>
<td>• Sentence length</td>
</tr>
<tr>
<td></td>
<td>• Thematic word</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Uppercase word</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teufel and Moens [TM97]</td>
<td>• Cue phrase</td>
<td>Naïve Bayes classifier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Location</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Sentence length</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Thematic word</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Title</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mani and Bloedorn [MB98]</td>
<td>• Location</td>
<td>Standardized canonical discriminant function (SCDF), C4.5-Rules, and AQ15c</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Thematic word</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Cohesion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neto et al. [NSK00]</td>
<td>• Thematic word</td>
<td>Naïve Bayes classifier and C4.5</td>
<td>• Proper noun</td>
</tr>
<tr>
<td></td>
<td>• Proper noun</td>
<td></td>
<td>• Cohesion</td>
</tr>
<tr>
<td></td>
<td>• Anaphor</td>
<td></td>
<td>• Thematic word</td>
</tr>
<tr>
<td></td>
<td>• Discourse maker in the beginning of sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Cohesion</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Sentence depth in the tree</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Position in the tree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myaeng and Jang [MJ98]</td>
<td>• Cue word</td>
<td>Naïve Bayes classifier</td>
<td>• Cue word</td>
</tr>
<tr>
<td></td>
<td>• Negative word</td>
<td></td>
<td>• Position</td>
</tr>
<tr>
<td></td>
<td>• Position</td>
<td></td>
<td>• Title</td>
</tr>
<tr>
<td></td>
<td>• Thematic word</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Title</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Centrality</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3

Development of Summary-based Event Detection (SED) Technique

Traditional event detection techniques utilize the full text of news stories for detecting whether a newly arrived news story belongs to a new or an old event. However, as mentioned, for illustration and comparison purpose, a news story may contain sentences or paragraphs that are not highly relevant to defining its event. Inclusion of such sentences and paragraphs in the similarity comparison by a traditional feature-based event detection technique might significantly degrade its detection effectiveness. Figure 3.1 shows a sample news story that reports new product releases (i.e., BCA series sector antennas) by Andrew Corporation. The paragraphs highlighted by gray areas contain the company background and contacting information and appear to be less relevant or even irrelevant ones in defining the event discussed in the news story. Thus, by identifying and filtering these less relevant or irrelevant sentences or paragraphs in a new story before performing feature-based event detection would have the potential in improving event detection effectiveness. Accordingly, in this study, we propose a Summary-based Event-Detection (SED) technique, whose process and detailed design are depicted in the following sections.
New BCA Series Sector Antennas Provide Optimized Coverage Solutions for Broadband Wireless Access

Updated 12:47 PM ET December 6, 1999

*****

Andrew Corporation (NASDAQ:ANDW) announced three new models in its Broadband Communications Antenna (BCA) series. In the sphere of broadband services, including high-speed Internet access, voice, and multimedia, these new antennas give effective broadband wireless coverage to enable new competitive broadband service providers to be successful.

The new models are: the BCA Series High Gain Sector Antenna for broadband wireless systems of 24.25-26.5 GHz; and BCA Series Mark II enhanced performance high gain sector antennas available in 27.35-28.5 GHz and 37.0-40.0 GHz frequency ranges. All three antennas offer versatile solutions for dealing with the complex and difficult issue of providing optimum, cost-effective coverage for broadband communications.

The High Gain Sector Antenna covers the 24.25-26.5 GHz frequency band, delivering 20.5 dBi of gain performance at 90 degrees azimuth coverage. The antenna provides high gain in an efficient size (18 inches high, 3.5 inches wide, and 9 inches deep) and weight (7.5 pounds). Available azimuth sizes include 45 and 90 degrees in either vertical or horizontal polarization. The elevation pattern is contoured to give the required signal strength inside the coverage area, free of nulls.

For more information, call the Andrew Customer Support Center at 1-800-255-1479 and request packet number 483 (bulletin numbers 10340, 10341, and 10342.)

Andrew Corporation is a global supplier of communications systems equipment and services. Major markets are wireless communications -- which includes cellular, personal communications services, and land mobile radio -- broadcast, and common carrier. Andrew is an S&P 500 company whose common stock trades on The Nasdaq Stock Market(R) under the symbol: ANDW.

For further information on Andrew products and services, including the text of this release, visit our Web site at http://www.andrew.com

Figure 3.1: Example of News Story
3.1 Process of Summary-based Event-Detection (SED) Technique

The overall process of the proposed Summary-based Event-Detection (SED) technique is shown in Figure 3.2. SED employs the text summarization method as a pre-processing task for selecting important sentences considered relevant to the event discussed in the news. That is, the news summarization phase shown in Figure 3.2 is to generate summaries for past news stories as well as for the new news story. Subsequently, the summary of the new news story together with the summaries of all past (i.e., historical) news documents will be used for event detection using a traditional event detection technique. Specifically, the event detection phase in the SED technique includes the following tasks, namely, feature extraction and selection step, document representation, and similarity comparison.
3.2 News Summarization Phase

The goal of the news summarization phase is to select relevant and representative sentences or paragraphs from a news story as its summary. It consists of two tasks: (1) news summarization learning and (2) news summary generation. The news summarization learning task involves the learning of a summarization model that will be used by the subsequent news summary generation task. By its inductive nature, the news summarization learning task requires manual summaries of news stories as training examples. The processes of the news summarization learning task and the news summary generation are shown in Figure 3.2, whose designs will be detailed in the following subsections.
3.2.1 News Summarization Learning Task

Selecting appropriate features in measuring significance of sentences in a news document in the most fundamental issue in the news summarization learning task. Based on the text summarization literature reviewed previously, this study employs six features, including cue phrase, thematic word, title word, location, cohesion, and sentence length, to represent each sentence in news documents for summarization model induction. In the following, we first explain these features whose alternative representation schemes are summarized in Table 3.2.

1. **Cue Phrase Feature**

Cue phrases are classified into two types: positive and negative cue phases. Sentences in a document containing positive cue phrases are more likely to be in the summary for the document. On the other hand, sentences including negative cue
phases are more likely to be excluded from the document’s summary.

To obtain the cue phase glossary, the process of **cue phrase discovery** is needed (as shown in Figure 3.3). We take a similar approach to that described in [E69] for discovering cue phrases from a set of training news stories with summaries. Since adjectives or adverbs seem to be more relevant as being positive or negative cue phrases, all adjectives and adverbs are parsed as candidate cue phrases from each training news story by the rule-based part-of-speech tagger developed by Brill [B92] in this study. Subsequently, the following two statistics are computed for each candidate cue phrase \( w \):

- Support rate = \( \frac{n_s(w)}{n_s} \)
- Selection ratio = \( \frac{n_{ss}(w)}{n_s(w)} \)

where \( n_s \) is the number of sentences in the training corpus,
\( n_s(w) \) is the number of sentences appearing \( w \) in the training corpus,
\( n_{ss}(w) \) is the number of sentences that contain \( w \) and is included in the summary.

According to these two statistics, some candidate cue phrases are selected as positive cue phrases, while some others are identified as negative cue phrases:

- Positive cue phrases: A candidate cue phrase is determined as a positive cue phrase if its selection ratio is greater than a pre-specified upper-selection-threshold and its support rate is greater than a pre-specified support-threshold. For example, some positive cue phrases discovered from a training news corpus in this study (see detailed discussions in Chapter 4) include “only”, “even”, “next”, “central”, “common”, “available”, and so on.
• Negative cue phrases: A candidate cue phrase is determined as a negative cue phrase if its selection ratio is no greater than a pre-specified lower-selection-threshold and its support rate is greater than the pre-specified support-threshold. Some positive cue phrases discovered from a training news corpus in this study “still”, “approximately”, “other”, “also”, “not”, “major”, and so on.

2. Thematic Word Feature

Words (typically nouns and noun phrases) that are not stop-list words and frequently occur in a news story are likely to be relevant to the topic of news, especially if such words don’t occur often elsewhere. Appendix A lists the set of stop words used in this study. As with the cue phrase discovery described above, nouns and noun phrases are identified for each training news document by the rule-based part-of-speech tagger. Subsequently, a standard TF\times IDF (term-frequency \times inverse-document-frequency) method is used to measure the weight of each word in every training document:

\[ w(i, k) = tf_{ik} \times \left( \frac{n}{df_k} + 1 \right) \]

where \( tf_{ik} \) denotes the frequency of term \( k \) in document \( i \)

\( df_k \) denotes the number of documents that contain term \( k \), and

\( n \) is the number of documents in collection

Accordingly, the top \( t_w \) words with highest TF\times IDF weights in each training news document are selected as the thematic words for the document.

3. Title Word Feature
Title and heading of a news story is often strongly related to its content. Hence, words occurring in the title and heading are usually considered as important indicators for measuring importance of sentences in a document. All stop words occurring in the title and heading are removed first, and the remaining title words of a news document form the title glossary for the news story.

4. 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 whether it appears in the first, middle, or last third of paragraphs in the document.

5. Cohesion Feature
Sentences are not essential for summary present low cohesion [MB98, NSK00]. Hence, the SED technique employs this feature that measures the degree of connectivity between sentences. As with the measure proposed by [NSK00], the cohesion of a sentence $s$ with other sentences in the same news story $N$ is measured as:

$$C_s = \sum_{s' \in N} \frac{\text{sim}(s, s')} {n - 1}$$

where $s$ is the target sentence, and $s \in N$, and $n$ is the number of sentences in $N$.

7. Sentence Length Feature
Too short or long sentences tend not to be included in summaries [KPC95, TM97]. For example, some sentences that belong to subtitles for next paragraphs are usually
short (e.g., the number of words being less than 10 terms) and, thus, are not important.

The above-chosen features are then used for representing each sentence in the training news documents. For each feature, representation schemes employed by some existing text summarization approaches are listed in Appendix B. In this study, alternative sentence representation schemes for each feature are shown in Table 3.2. Table 3.1 provides definitions of variables used in Table 3.2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_c$</td>
<td>Number of positive cue phrases appearing in a sentence.</td>
</tr>
<tr>
<td>$N_c$</td>
<td>Number of negative cue phrases appearing in a sentence.</td>
</tr>
<tr>
<td>$N_k$</td>
<td>Number of thematic words appearing in a sentence.</td>
</tr>
<tr>
<td>$N_t$</td>
<td>Number of title words appearing in a sentence.</td>
</tr>
<tr>
<td>$L_s$</td>
<td>Length of a sentence.</td>
</tr>
<tr>
<td>$MaxL_s$</td>
<td>Maximum sentence length in a news story.</td>
</tr>
<tr>
<td>$C_s$</td>
<td>Cohesion of a sentence.</td>
</tr>
</tbody>
</table>
### Table 3.2: Representation Schemes Employed by the SED Technique

<table>
<thead>
<tr>
<th>Feature</th>
<th>Sentence Representation Schemes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Cue Phrase</td>
<td>Binary</td>
<td>1 if containing any positive cue phrases, otherwise 0.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- ( P_c / L_s ) Number of positive cue phrases in the sentence divided by the length of the sentence.</td>
</tr>
<tr>
<td>Negative Cue Phrase</td>
<td>Binary</td>
<td>1 if containing any negative cue phrases, otherwise 0.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- ( N_c / L_s ) Number of negative cue phrases in the sentence divided by the length of the sentence.</td>
</tr>
<tr>
<td>Thematic Word</td>
<td>Binary</td>
<td>1 if containing any thematic words, otherwise 0.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- ( N_k / L_s ) Number of thematic words in the sentence divided by the length of the sentence.</td>
</tr>
<tr>
<td>Title Word</td>
<td>Binary</td>
<td>1 if containing any title words, otherwise 0.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- ( N_t / L_s ) Number of title words in the sentence divided by the length of the sentence.</td>
</tr>
<tr>
<td>Sentence Location</td>
<td>{F, M, L}</td>
<td>Sentence occurs in first (F), middle (M), or last (L) third of paragraphs.</td>
</tr>
<tr>
<td>Cohesion</td>
<td>( C_s )</td>
<td>Sum of similarities with all other sentences divided by (# of sentences in the news story – 1).</td>
</tr>
<tr>
<td></td>
<td>Binary</td>
<td>1 if greater than average ( C_s ), otherwise 0.</td>
</tr>
<tr>
<td>Sentence Length</td>
<td>Binary</td>
<td>1 if the sentence length is greater than a pre-specified threshold, otherwise 0.</td>
</tr>
<tr>
<td></td>
<td>( L_s / \text{Max} L_s )</td>
<td>Number of words occurring in the sentence divided by the number of words occurring in the longest sentence of the document.</td>
</tr>
</tbody>
</table>

Upon the completion of sentence representation for all sentences in the training news documents, the **summarization model induction** process is initiated. In this study, we adopt C4.5 (a decision induction algorithm) as the inductive learning
method. The output for the summarization model induction process is a summarization model that will be used by the news summary generation for predicting whether a sentence in a news document would be included in its summary.

3.2.2 News Summary Generation Task

When a news story arrives, the news summary generation task is to select relevant sentences to be included in the summary. As with the news summarization learning task, each sentence in the target news story is represented using the representation schemes shown in Table 3.2. Subsequently, the reasoning process predicts whether a sentence would be an important sentence as a summary, based on the summarization model induced previously.

Using the summarization model induced by C4.5, only the dichotomous classification can be made. That is, the prediction for a sentence is 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). Since we would like to adjust the length of a summary by specifying a desirable compression ratio (defined as the number of sentences in the summary divided by the total number of sentences in the original document), the prediction mechanism needs to be revised. In this study, we estimate the accuracy of each decision path (or called decision rule) in a decision tree (produced by C4.5) by the Laplacian accuracy [CB91]. The Laplacian accuracy of a decision rule is generally defined as:

\[
\text{LaplaceAccuracy} = \frac{(n_c + 1)}{n_{\text{tot}} + k}
\]

where \( k \) is the number of decision classes.
\( n_c \) is the number of the training instances in the predicted class covered by the rule, and

\( n_{tot} \) is the total number of training instances covered by the rule.

In our news summarization application, the number of decision classes is 2 (i.e., either important or not important). Thus, \( k = 2 \). Furthermore, when a sentence in a news story is covered by a decision path in the summarization model, the Laplacian accuracy of this decision path is used as the weight of the sentence. Accordingly, given a compression ratio, we select a desired number of sentences from a news story based on the weights of the sentences. Specifically, if \( R \) sentences need to be selected, we first select in the descending weight order from those sentences that are predicted as important. If the number of sentences predicted as being important is less than \( R \), we then select in the ascending weight order from those sentences that are predicted as not important until a total of \( R \) sentences is selected.

### 3.3 Event Detection Phase

As shown in Figure 3.2, given a summary for a newly arrived news story and the summaries of all past news stories, 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 employ INCR [YPC98, YCB99] for such event detection purpose. A rule-based part of speech tagging proposed by Brill [B92, B94] is adopted for syntactically tagging each word in the summaries of all past news stories. Subsequently, if only noun phrases in a news story is to be preserved, a noun phrase parser proposed by Voutilainen [V93] is employed and implemented for extracting noun phrases from syntactically tagged documents. Subsequently, \( k \) representative features are selected for the summaries of past news stories, using the
popular TF×IDF feature selection method.

After a set of representative features is selected, the summary of the new news story and the summaries of all past news stories are represented using the incremental version of the TF×IDF representation scheme proposed by [YPC98, YCB99]. Subsequently, the feature vector for the new news story is compared with those of all past news stories, using the INCR’s similarity measure that combines lexical similarity and temporal proximity. 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.
Chapter 4
Empirical Evaluation

4.1 Evaluation Design
This chapter reports the empirical evaluation of the proposed SED technique, using a traditional feature-based event detection technique as performance benchmarks. In the following, the design of the empirical experiments will be described first, including data collection, evaluation criteria, and evaluation procedure. Finally, the empirical evaluation results will be discussed.

4.1.1 Data Collection and Summary Preparation
For the evaluation of the proposed SED technique, news stories from November 1999 to December 1999 were collected from a news web-site, excite.com. Six event topics were identified and selected, including adjustment of interest rate, initial public offering, business merger, business new product, business partnership, and computer virus. 506 news stories (where 220 news from November 1999 and 286 news from December 1999) pertaining to the six event topics were manually identified. The event contained in each news story was also coded manually. The summary of the data corpus is provided in Table 4.1.
Table 4.1: Summary of Data Corpus

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

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

For obtaining correct news summaries, a senior researcher who is familiar with this data corpus participated in the process of news summarization. Specifically, the senior researcher summarized for each news story highly relevant and highly irrelevant sentences. Because such manual summarization is time-consuming, the summaries on the news stories of November 1999 in the data corpus were obtained. The profile of the news summary is shown in Table 4.2.
Table 4.2: Summary of News Summarization

<table>
<thead>
<tr>
<th>Event Topic</th>
<th>Average Number of Sentences per News Story</th>
<th>Average Number of Highly Relevant Sentences per News Story</th>
<th>Average Number of Highly Irrelevant Sentences per News Story</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjustment of Interest Rate</td>
<td>22</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Initial Public Offering</td>
<td>20</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Business Merger</td>
<td>22</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>New Product Announcement</td>
<td>23</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Business Partnership</td>
<td>22</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Computer Virus</td>
<td>21</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>21.64</td>
<td>3.04</td>
<td>9.19</td>
</tr>
</tbody>
</table>

### 4.1.2 Evaluation Criteria for Event Detection

In this study, 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. To address the inevitable tradeoffs between miss and false alarm rates, Detection Error Tradeoff (DET) curves were employed [APL98, YCB99, YPC98]. An event detection technique with its DET curve closer to the origin would be more desirable. 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.

4.1.3 Performance Benchmarks for Event Detection

A traditional feature-based event detection technique was used to provide the desired effectiveness benchmarks. Specifically, the single-pass incremental clustering (INCR) for event detection proposed by Yang et al. [YPC98, YCB98] was employed. Without loss of generality, we modified its linear decaying-weight similarity function by changing the time window from the number of news stories to the number of days, as follows:

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

where \(i\) is the number of days between \(x\) and the most recent document in \(c\), and \(m\) is the time window measured in number of days before \(x\).

4.2 Evaluation Result

4.2.1 Parameter Tuning

Parameter turning experiments were conducted for selecting appropriate parameter values that would result the best event detection performance measured by the evaluation criteria described above for each event detection technique. The design and results of the parameter tuning experiments are described as follows.

Parameter Turning of INCR:

The INCR technique involves three parameters: the number of features \(k\), time
window \( w \), novelty threshold \( t_n \). In this study, the news stories of November 1999 were employed as the data set for parameter tuning for the INCR technique. Specifically, the news stories from the first 15 days in the tuning set were used as historical news stories, and the rest of news stories in the tuning set were included as the testing set. To detect whether a news story in the testing set contained a new event by using the INCR technique, the news story was compared to all of its prior news stories (including historical ones). To expand the number of trials, 70\% of the news stories were randomly selected from the historical and testing sets respectively, and the random selection-and-detection process was repeated 10 times. Thus, the overall detection effectiveness was estimated by averaging the performance across all iterations.

We investigated the number of features \( k \) ranging from 50 to infinite (i.e., \( k = 50, 100, 150, 200 \) and infinite), the time window \( w \) ranging from 7 to 60 (\( w = 7, 14, 30 \) and 60 days), the novelty threshold \( t_n \) ranging from 0.01 to 0.5 at increments of 0.01. Evidently, the increase of the novelty threshold resulted in the decrease of miss rate at the cost of false alarm rate. At any level of \( w \) investigated, the Detection Error Tradeoff (DET) curve, in general, was getting closer to the origin as \( k \) increased from 50 to 150. However, further increases of \( k \) from 150 to infinite (INF) would not improve the detection effectiveness. Thus, in this study, we set \( k \) as 150 for subsequent experiments. On the other hand, as shown in Figure 4.1, when \( k = 150 \), setting \( w \) as 60 generally resulted in the best detection effectiveness measured in miss and false alarm rates. Specifically, when \( w \) was 60 and \( k \) was 150, the best performance was achieved at the novelty threshold of 0.18 (where the minimal Euclidean distance to the origin was attained). Hence, we selected the time window of 60 and number of features as 150 for the INCR technique.
Parameter Tuning of the Summarization Process of SED:

The summarization process involved several decisions, including 1) the upper-selection threshold, lower-selection threshold and support ratio for determining positive and negative cue phrases, 2) the number of words $t_w$ selected as the thematic words for each news document, and 3) appropriate sentence representation schemes for each feature for summarization model induction and reasoning (as shown in Table 3.2). In this study, we set the upper-selection threshold as 0.3, the lower-selection threshold as 0.25 and the support ratio as 0.02 when determining positive and negative cue phrases. The parameter tuning experiments for the summarization process of the SED technique were designed for the rest of the decisions.

As mentioned, the manual summaries were available for the 220 news stories in November 1999 in the data corpus. Thus, only the 220 news stories with these manual summaries were used as the target dataset for tuning purpose for the summarization process of the SED technique. Since the summarization process
involves the learning and reasoning tasks, we randomly selected 70% of the news stories in the target dataset for training and the remaining 30% of the news stories for testing. Ten iterations of learning-and-testing were performed. For evaluating the learning effectiveness of the summarization process in the SED technique, the precision rate and recall rate were used in this study. The precision rate is defined as the percentage of highly relevant sentences identified by the news summarization process that are highly relevant ones as suggested in manual summaries. On the other hand, the recall rate is defined as the percentage of highly relevant sentences as appeared in manual summaries that are identified by the news summarization process. Thus, the overall learning effectiveness of the summarization process was estimated by averaging the precision rate and recall rate across the ten iterations respectively.

We investigated the number of thematic words $t_w$ ranging from 10 to 20 at increments of 5 (i.e., $t_w=10, 15, \text{ and } 20$) for each news story in this study. Different sentence representation schemes for each feature were examined, as shown in Table 4.3. Among all experiments, $t_w = 20$, the binary representation scheme (with 3 top-scoring sentences being 1 and the rest sentences in a new story being 0) for the positive and negative cue phrase features, the standardized score representation scheme for the thematic word and title word features, the categorical representation scheme for the sentence location feature, 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 resulted in the best learning effectiveness of the summarization process of the SED technique. Thus, they were adopted for subsequent experiments.
Table 4.3: Sentence Representation Schemes for Summarization Features Investigated

<table>
<thead>
<tr>
<th>Feature</th>
<th>Sentence Representation Schemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cue Phrase (Positive/Negative)</td>
<td>Binary, top 3 (i.e., top 3-scoring sentences get 1; all others 0), top 4, top 5, and standardized score</td>
</tr>
<tr>
<td>Thematic Word</td>
<td>Binary, top 3 (i.e., top 3-scoring sentences get 1; all others 0), top 4, top 5, and standardized score</td>
</tr>
<tr>
<td>Title Word</td>
<td>Binary, top 3 (i.e., 3 top-scoring sentences get 1; all others 0), top 4, top 5, and standardized score</td>
</tr>
<tr>
<td>Sentence Location</td>
<td>Categorical (i.e., appearing in either first (F), middle (M), or last (L) third of paragraphs)</td>
</tr>
<tr>
<td>Cohesion</td>
<td>Binary or standardized score</td>
</tr>
<tr>
<td>Sentence Length</td>
<td>Binary, length threshold of 5, length threshold of 10, length threshold of 15, or standardized score</td>
</tr>
</tbody>
</table>

**Parameter Tuning of the Event Detection Process of SED:**

The SED technique involves four parameters, including: the number of features $k$, time window $w$, novelty threshold $t_n$ and compassion ratio $c$. The first three parameters are the same as those required in the INCR technique. The parameter tuning experiment for the SED technique was designed in the same manner as that for the INCR technique, and the tuning experiment was performed ten times.

We investigated the time window $w$ ranging from 7 to 60 ($w = 7, 14, 30, \text{to} 60$ days), the novelty threshold $t_n$ ranging from 0.01 to 1.0 at increments of 0.01, and the number of features $k$ ranging from 25 to infinite (i.e., $k = 25, 50, 75, 100, \text{and} \infty$). We set the compression ratio $c$ as 25% in this parameter tuning experiment. Across all values of $w$ examined, when $k$ was set to infinite, the SED technique would result in better detection effectiveness. On the other hand, as shown in Figure 4.2, when $k = \infty$, setting $w$ as 60 appeared to obtain the best detection.
effectiveness. As shown, at any level of false alarm rate that was lower than 18%, the DET curves of the SED technique generally shifted toward the origin as the time window increased from 7 to 60. On the other hand, when the false alarm rate was higher than 18%, the DET curves of the SED technique were comparable over different time windows examined. The SED technique arrived at the best performance when $w=60$, $k = \infty$, and the novelty threshold was 0.15. Thus, we decided on the time window of 60 and $k$ was infinite for the SED technique.

![Figure 4.2: Detection Error Tradeoff Curves of SED ($k = \infty$, $c = 25\%$)](image)

**4.2.2 Comparative Evaluation of Event Detection Techniques**

The traditional feature-based event detection (INCR) and the proposed summary-based event detection (SED) technique were compared using the parameter values determined in the previous subsection. Similar to previous tuning experiments, the data corpus was divided into two sets: historical (including news stories in November 1999) 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 stories were randomly selected from the historical
and the testing set, respectively, and the random selection process 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 ranging from 0.01 to 0.5 and that for SED ranging from 0.01 to 1 at 0.01 increments. As shown in Figure 4.3, as the compression ratio increased from 25\% to 75\%, the DET curves of the SED technique generally shifted toward the origin. However, the INCR technique generally outperformed the SED technique.

![Figure 4.3: Detection Error Tradeoff Curves of Different Event Detection Techniques](image)

The above evaluation result is not encouraging. As shown in Table 4.1, in some event topics in the data corpus used for this evaluation, most of their news stories were new events (i.e., initial public offering, new product announcement, and business partnership). Thus, inclusion of such event topics and their news stories in
the evaluation may not truly reflect the effectiveness of the SED and INCR techniques in a real world setting where news stories discussing old events may often encounter. Thus, we conducted another comparative evaluation using the other three event topics (i.e., adjustment of interest rate, business merge, and computer virus).

With this new dataset, the SED technique could achieve comparable or even better detection effectiveness than the INCR technique. As shown in Figure 4.4, at any level of miss rate that was higher than 18%, the INCR technique achieved lower false alarm rate than the SED technique across different compression rates examined. However, when the miss rate was lower than 18%, the detection effectiveness achieved by the SED technique at the compression ratio of 55% or 75% was comparable to or even higher than that of the INCR technique. We expect that given a data corpus where the percentage of news stories discussing old events is higher, the SED technique would achieve comparable or even better detection effectiveness than the INCR technique.
Figure 4.4: Detection Error Tradeoff Curves of Different Event Detection Techniques based on Three Event Topics
Chapter 5

Conclusions and Future Research Directions

As an organization’s environment becomes more diverse, dynamic and complex, uncertainty faced by the organization increase. Environmental scanning is an important process of strategic management that permits an organization to adapt to its environment and subsequently to develop effective responses to secure or improve its position in the future. Event detection that detects the onset of new events from news documents has become a critical activity of chief executive officers for planning their firms. Traditional feature-based event detection techniques detect events by comparing the similarity between features of news stories and incur several problems. For example, for illustration and comparison purpose, a news story may contain sentences or paragraphs that are not highly relevant to defining its event. Without removing such less relevant sentences or paragraphs before detection, the effectiveness of traditional event detection techniques may suffer. In this study, we developed a summary-based event detection (SED) technique that filters less relevant sentences or paragraphs in a news story before performing feature-based event detection. Using a traditional feature-based event detection technique (i.e., INCR) as benchmark, the empirical evaluation results showed that the proposed SED technique could achieve comparable or even better detection effectiveness (measured by miss and false alarm rates) than the INCR technique, for data corpora where the percentage of news stories discussing old events is high.

Some future research works related to this study should be continued, including:
1. In this study, manual summaries were obtained from a senior researcher. Involving more human experts for generating manual summaries that are used for inducing a summarization model would have the potential to improving the effectiveness of the SED technique.

2. The news summarization induction in the SED technique employed a decision tree induction approach, namely C4.5 technique. Other induction techniques (Naïve-Bayes induction method, backpropagation neural network) could be adopted by the SED technique.

3. On the other hand, the proposed SED technique detects whether a news story contains an unseen event by comparing the similarity of features between news stories; i.e., still based on the traditional feature co-occurrence approach. Thus, word mismatch was not addressed in the SED technique. Further research should be directed to incorporate a mechanism for solving word mismatch problem in the SED technique.

4. Furthermore, in this study, the experimental data set used for evaluating the SED technique only comprised news stories of six event topics across two months. A larger data set with more news stories and event topics for empirical evaluation of the proposed technique would be desirable.

5. Finally, in this study, we focused 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.
Appendix A: List of Stop Words

- a
- about
- all
- an
- and
- anyone
- anything
- anywhere
- are
- as
- at
- away
- back
- be
- been
- being
- by
- can
- cannot
- can't
- could
- do
- did
- done
- does
- doesn't
- don't
- didn't
- every
- everybody
- everyone
- everything
- everywhere
- for
- from

- her
- herself
- here
- have
- having
- has
- he
- him
- himself
- his
- how
- I
- in
- itself
- is
- isn't
- its
- it
- it's
- lets
- let
- may
- might
- mightn't
- man
- me
- men
- mr
- mrs
- my
- myself
- nobody
- nothing
- nowhere
- on

- one
- our
- she
- should
- shouldn't
- somebody
- some
- someone
- something
- such
- somewhere
- their
- that
- the
- them
- they
- to
- us
- very
- was
- we
- which
- with
- would
- wouldn't
- won't
- you
- you're
- your
- yours
Appendix B: Sentence Representation Schemes Employed by Existing Text Summarization Approaches

<table>
<thead>
<tr>
<th>Positive Cue Phrase Feature</th>
<th>Negative Cue Phrase Feature</th>
<th>Thematic Word Feature</th>
<th>Title Word Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature</td>
<td>Representation</td>
<td>Description</td>
<td>Literature</td>
</tr>
<tr>
<td>[KPC95]</td>
<td>Binary</td>
<td>This feature is true for sentences that contain any of 26 indicator phrases.</td>
<td>[MJ98]</td>
</tr>
<tr>
<td>[TF97]</td>
<td>Score (5 classes)</td>
<td>Manually created</td>
<td>[TM97]</td>
</tr>
<tr>
<td>[MJ98]</td>
<td>Probability</td>
<td>(P(NW_i))</td>
<td>[MJ98]</td>
</tr>
<tr>
<td>[NKS00]</td>
<td>High / Low</td>
<td>TF×IDF, select 15 terms with high value</td>
<td>[MB98]</td>
</tr>
<tr>
<td>[MB98]</td>
<td>Binary</td>
<td>If TF×IDF&gt;average TF×IDF score gets 1, otherwise 0.</td>
<td></td>
</tr>
<tr>
<td>[TM97]</td>
<td>Binary</td>
<td># of Title words/ sentence length. 18 top-scoring sentences get 1, all others 0.</td>
<td>[MJ98]</td>
</tr>
<tr>
<td>[MB98]</td>
<td>Binary</td>
<td>Number of section heading or title term mentions.</td>
<td></td>
</tr>
<tr>
<td><strong>Sentence Location Feature</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[KPC95]</td>
<td>{1, 2, 3} According to whether sentences are paragraph-initial, paragraph-final, and paragraph-medial.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[TM97]</td>
<td>Binary Peripheral sections get non-zero values, middle of document get 0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[MJ98]</td>
<td>Probability P(P_i), six different position values can be assigned to individual sentences.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[MB98]</td>
<td>{1, 2, 3} Sentence occurs in first, middle or last third of paragraph.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>{1, 2, 3} Sentence occurs in first, middle or last third of section.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>{1, 2, 3} 1 if sentence occurs in introduction, 2 if in conclusion, 3 if in other.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>{1, 2, 3, 4} 1 if sentence is a top-level section, 4 if sentence is a subsubsubsection.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Cohesion Feature</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[MJ98]</td>
<td>Value The centrality value of a sentence can be calculated by comparing the vector representation of the sentence and that of the document.</td>
</tr>
</tbody>
</table>
| [NKS00]              | High / Low Total similarity value of each sentence is obtained through the sum of the individual similarity values between that sentence and each of the other sentences.  
Normalized= total similarity value of each sentence / largest total similarity value. |
| [MB98]               | Binary To compute synonyms the algorithm use WordNet, comparing contentful nouns as to whether they have a synset in common.  
Number of unique sentences with a synonym link to sentence.  
Co-occurrence scores between contentful words up to 40 words apart are computed using a mutual information metric.  
Number of unique sentences with a co-occurrence link to sentence. |
A sentence.

<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>[KPC95]</td>
<td>Binary</td>
<td>Threshold= 5 words</td>
</tr>
<tr>
<td>[TM97]</td>
<td>Binary</td>
<td>Threshold= 15 words</td>
</tr>
</tbody>
</table>
References


[BK97] Boguraev, B. and Kennedy, C., “Salience-Based Content


[L58] Luhn, H. P., “The Automatic Creation of Literature Abstracts,” *IBM*


