Automatic Bilingual Text Document Summarization

羅秀嫻
Lo Sau-Han Silvia

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Philosophy
in
Systems Engineering and Engineering Management

August 2002


香港中文大學
圖書館系統
28 MAR 2003
UNIVERSITY
LIBRARY SYSTEM
THE CHINESE UNIVERSITY OF HONG KONG

Abstract

Our research goal is to investigate approaches for automatic summarization that captures the main theme and events covered by a set of documents. In the first part of our research, a domain independent, single document summarization system was developed to generate informative extracts (sentences) based on a thematic term approach. Thematic terms are automatically discovered from a corpus as well as from the corresponding documents via information retrieval techniques. Next, we examined the feasibility of the thematic term summarization approach by applying it to a Chinese corpus of a different domain. For English summaries, the performance evaluation was conducted by the content-based similarity method. However, content-based evaluation could not be used for Chinese summaries since the handwritten summaries were not available. Consequently, we designed an alternative evaluation scheme, namely Average Inverse Rank (AIR), which makes use of an information retrieval model. This evaluation method attempts to measure

the representative power of a summary for its original text. The results suggested that 20% of the full-length document is sufficient for a good summary.

In the second part of our research, we developed a bilingual summarization technique over news documents based on an event-driven approach. First, we employed dictionary-based term translation in two steps, i.e., phrase translation and term disambiguation for handling English news. Next, unsupervised learning was used to discover events and to generate coherent event clusters. Afterwards, heuristic criteria were developed to select relevant and cohesive clusters in building the event list and the content for the summary. Finally, we adopted the recall and precision metrics in assessing the quality of the event-driven summaries. The results showed that our summaries performed better than the baseline method (summaries generated by randomly selecting sentences) and obtained precision scores around 70% at 10-20% in length of the original documents. To further demonstrate the effectiveness of our bilingual event-driven approach summarization technique, we conducted additional experiments using a parallel corpus comprised of Chinese and English newswire texts. Around 60% precision has been in all runs.

摘要

此論文探討不同全文摘要方法，用來自動地捕獲文件中覆蓋的主題及主要事件。在第一部分的研究中，我們基于主題詞方法，開發了一個領域無關的、單文件摘要系統，去產生具有資料性的抽取物（句子）。主題詞是通過信息提取技術自動地在語料庫及相應的文章中發現的。然後，我們用另一不同領域的中文語料庫來檢查主題詞摘要方法的可行性。在英文摘要效率方面，我們用了內容的相似性作爲評估。然而，因缺乏作者親自編寫的摘要，基于內容的評估方法並不能有效地用。因此，我們採用了信息檢索模型，設計了另一評估方法，名命爲『平均反排序法』（AIR）。此評估方法量度摘要對原文的代表能力。結果顯示好的摘要只須要文章的20%。

在第二部份的研究中，我們基于事件驅使方法，開發了一個雙語新聞摘要系統。首先，對於英文新聞，我們採用了基于字典的詞語翻譯；這方法包括了兩個步驟：短語翻譯及翻譯詞歧义排除。然後，利用非監督式的學習發現事件及生成凝聚的事件群去建立事件列及摘要

內容。 最後， 我們採用了精確及召回率來評估事件驅使的摘要質素。 實驗結果顯示我們的摘要比基準方法（隨機揀選文中句子生成摘要）表現得好。 並且在10-20%的原文長度時， 獲得高達70%的精確率。 我們利用了一個中、英新聞的平衡語料庫從而顯示雙語事件驅使摘要技術的有效性。 在所有實驗組合中， 其精確度可達60%左右。

Acknowledgments

Through out my M.Phil programme, in addition to engineering research, I have learned how to write and present in English fluently. I believe that this will be very valuable and important for my future. In this aspect, I would like to thank my supervisors, professors Wai Lam and Helen Meng, teachers in Independent Language Center (ILC), professor Kam-Fai Wong and Jackie Wong for correcting and withstanding my poor English.

Moreover, during these two years, I received plenty of useful advice and help from my supervisors. I would like to express my gratitude to them. I would also thank my thesis committee, professors Kam-Fai Wong and Jeffrey Xu Yu, both of the Department of Systems Engineering and Engineering Management CUHK; and professor Kui Lam Kwok from Queen's College, City University of New York, for their useful comments and feedback.

My grandma and sister have been very supportive. They have never prevented me from doing research all day and continuously enduring my bad

temper. Furthermore, I am in debt to my best friend, Mill, who always pushes me to concentrate on my work, cooks for me and stands by my side. He also helps feed my lovely puppy, Dang Dang, which gives me a lot of fun in my difficult period.

Finally, a lot of friends in the Department of Systems Engineering and Engineering Management, CUHK, have helped and encouraged me, and even did exercise with me in the past two years. I would like to give thank all. They are Brenda, Fiona, Ida, Kin, Mandy, Michael, Paul, Qiuyue, and Timothy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent decades, with the availability of large data storage and rapid growth of the World Wide Web, enormous amounts of electronic information are accessible. These changes lead to information explosion with massive amount of information are available in our daily life. For example, information are easily available from news articles, business reports, government documents, etc. over the World Wide Web. Many people from different sectors and industries could benefit from such sources if the information could be presented concisely and in real-time. As a result, various summarization techniques have emerged and play an important role for the future.

Our objective is to develop approaches for automatic summarization that captures the main theme and events covered by a set of documents. Also, we target to provide cross-lingual summarization from multiple sources. In

this way, users could choose the target language for the summary. In our research, we first developed a single document summarization framework based on thematic term approach. We then investigated multi-document bilingual summarization based on an event-driven approach.

The rest of this chapter is organized as follows: The definitions of a summary and text summarization are given in the next section. Previous effort in summarization and widely adopted evaluation metrics will then be presented. Lastly, our major research contributions are presented.

1.1 Definition of a summary

A summary contains important information and concepts presented in the original document which can help end-users achieve certain tasks. For instance, summaries of large corpora can serve as short index to retrieve the original documents. News headlines are a kind of summary that businessmen screen through to obtain the general outline of what has happened. Basically, the length of a summary is usually significantly shorter than the original document. Also, a summary can be produced from a single document or multiple documents. In general, a summary can be classified as an extract or an abstract. Extracts are portions of content in the document. These portions can be key words, sentences, clauses, or even paragraphs. Abstracts contain

the central idea information of the original document in a compressed form. They are expressed in different words, phrases, and sentence structures than those used in the source.

Sometimes, users prefer to generate a summary as an extract rather than an abstract. It is because extracts not only retain the main objective of the document, but also are generated easily. In general, not all of the extracts in a document, such as sentences, are relevant to the subject matter. Some sentences may be redundant or carry less informative items. On the contrary, an abstract requires more effort, not only focusing on the selection of the main theme of the text, but also requiring extra engines to compose the abstract text into coherent and cohesive passage.

1.2 Definition of text summarization

Text summarization is a process to generate a summary that has significant reduction in length compared to the original text. The objective of the process is to deliver information that is relevant to the main theme of the subject matter covered in a single document or different related documents. Text summarization is a complex task involving three elements: input, purpose and output.

The input of summarizers can be classified along four dimensions, in-

cluding source, language, specificity, and genre. The source can be a single document or multiple documents. Language can be monolingual or multilingual. Most of the summarization techniques are language sensitive, and hence the technique for different languages may be different. A summarizer can focus on specific format and content in a single domain. Less consideration may be given in idiosyncratic words. However, in general domains, we should consider some methods to deal with this problem.

The purpose of text summarization depends on what the summary is used for and who the audience will be. Some summaries should be tailor-made for particular situations. Therefore, where the subject matter happened, what for, and when this summary is used can be given in advance so that the summarizer can generate suitable content.

Lastly, the output of summarization can be an extract or an abstract. Coherence of the output should be considered as well. A fluent summary is written in full, grammatically correct sentences and all sentences are related to each other.

1.3 Previous work

Summarization methodologies have been investigated for over a decade. Information retrieval techniques are usually employed to extract relevant ma-

terials from a document [9, 23, 30, 33, 35, 43]. Recently, some efforts aim at performing summarization across related texts in multiple documents. This development is realistic under the availability of large amount of data.

In Sections 1.3.1 to 1.3.3, we describe previous work on text summarization under three categories, namely, extract-based text summarization, abstract-based text summarization, and sophisticated text summarization.

1.3.1 Extract-based text summarization

Extract-based summarization focuses on extracting important segments from documents. These segments can be key phrases, sentences, paragraphs, etc. The summarizer retains the appropriate texts of the original document. A common approach is to calculate a score for each segment to indicate the degree of importance. Those segments with high scores are collected to generate the summary.

Different researchers employ different criteria to compute the score for each text segment. Those criteria are related to features such as stochastic measurements for the significance of key terms in the sentence [23, 27, 44], sentence location in the source text [11, 16, 44], the presence of cue or indicator phrases [13, 37] or title words [16], and professional names [44]. Furthermore, statistical features derived by information retrieval techniques are

also used in the weighting metrics for text segments [18].

Apart from using features explored from origin texts, heuristic rules are developed to contribute to scores. Teufel and Moens considered the sentence length for generating extractive summaries [44]. Kupiec et al. utilized the uppercase word feature, with the belief that proper names are important in the extracts [23].

Many approaches make use of positive indicators to locate important text segments. On the contrary, Goldstein et al. tried to obtain negative indicators by analyzing newswire summaries [18]. Those negative indicators were anaphoric references, honorific, negation words, auxiliary verbs, integers, evaluative and vague words as well as conjunctions and prepositions.

Kan and McKeown [21] suggested applying the information extraction (IE) technique in a summarization system that was different from existing template-based methods. The system recognized four major entities, namely, people, organizations, places, and multi-word terms. After the weights of entities were computed, top-ranking ones were selected. Sentences containing those selected entities were extracted. A template filled with questions was prepared. The final summary was produced by selecting sentences containing answers for the questions.

Interestingly, Nakao [34] suggested detecting the thematic hierarchy of a text to generate a one-page summary. The system first decomposed a

text into an appropriate number of textual units by their subtopics. It then identified boundary sentences as theme information for a summary.

Some researchers considered machine learning-based text summarization [9, 13, 23, 35, 43, 44, 45]. They used statistical or decision tree models to build classifiers based on heuristic clues or thematic words with the help of training data (handwritten summary sentences). Such classifiers helped judge whether each sentence belonged to the summary.

Recently, a few researchers explored the event-driven summarization approach. Allan et al. [6] defined temporal summaries of news stories as extracting on-event sentences. "On-event" implies that the sentences under consideration mention one of the events that the topic covers. However, their work did not focus on the aspect of bilingual summarization.

Radev et al. used cluster centroids produced by a topic detection and tracking system for sentence extraction [40]. Stein et al. produced summaries for each document, then grouped all summaries together in cluster. For each cluster, the summarizer found the most representative passages to be included in the final summary [42].

Gong and Liu generated summaries by extracting sentences using latent semantic analysis and relevance measure separately [19]. The first method calculated the inner product between sentences and the document itself. This score was then used to rank sentences. The top-ranked sentences were in-

cluded in the summary. After that, all terms, which have already been found in the summary were ignored and the inner product between the remaining sentences and the document was computed. The selection process was stopped when the predefined summary length was reached. In addition, the singular value decomposition method was applied as the relevance measure to extract sentences.

Nomoto and Matsumoto investigated the diversity of concepts in text for the summarizing task [36]. Their hypothesis was that sentences should be both relevant to the query and have the least similarity to sentences selected previously. As a result, their work not only extracted appropriate sentences, but also minimized redundancy in the summary.

1.3.2 Abstract-based text summarization

This kind of text summarization is more complicated than an extract-based one. Besides, applying various methods in locating key fragments, fusion engines or more knowledgeable tools are used to reformulate or regenerate extracts in the summary generation process. In addition, evaluation metrics are difficult to develop. Therefore, less researchers attempt abstract-based text summarization.

Knight and Marcu focused on compressing extracted long sentences by

two algorithms [22]. They were the statistical noisy-channel approach and the decision-based deterministic model. As a result, coherence and readability of the summary were improved. Barzilay et al. used language generation to reformulate the wordings of a summary [10].

1.3.3 Sophisticated text summarization

The effectiveness of a summarizer is judged by the quality of the generated summary. Also, clear representation of a summary can improve readability. Ando et al. considered not only the quality of the summary, but also the way of presenting the summary items [7]. They provided an interface to show the thematic elements that came from the subsets of the document collection. Teufel and Moens tried to extract and display summary sentences according to their rhetorical units such as *introduction, purpose, experimental, design, results, discussion and conclusion* [43].

With the mature development in Internet services, people often browse Web pages to acquire the desired information. In view of this, Berger and Mittal suggested a prototype system to produce a summary of a Web page in a gist [32]. This summarization methodology has to deal with complicated structure of the text content. This is because the content of different Web pages can vary significantly in structure and context. They proposed a

probabilistic model in selecting and ordering words in a gist (summary).

1.4 Summarization evaluation methods

Even though methodologies in developing an extract-based summarization system is important, the quality of extracts needs to be assessed. To date, there is no golden rule to evaluate summaries. Generally, the summary quality can be evaluated by two categories: intrinsic and extrinsic [28]. Besides, research projects such as TIPSTER Text Summarization Evaluation (SUMMAC) [29] and Text Summarization Challenge (TSC) [1] have defined their own evaluation methods for text summarization.

1.4.1 Intrinsic evaluation

Intrinsic evaluation helps assess performance of document summarization by directly comparing the system generated summary with a standard human summary. For a single document, possible evaluation methods are recall and precision, utility figures [40] or content-based measures [15]. For multiple documents, performance can still be assessed by using the above metrics over the union of the important sentences in all documents. These evaluation methods can be used for an extract-based summary.

1.4.2 Extrinsic evaluation

Extrinsic evaluation is a task-based evaluation method, where the assessment of quality is based on the task specified. For example, Eduard suggested the question-answering approach for this task [17]. A group of analysts prepare a set of questions that can be answered by reading important texts in document beforehand. Upon the generation of extracts, human analysts then attempt to answer the questions from three different perspectives:

- Before reading any summary
- After reading the summary produced by a system
- After reading the whole document

The more questions answered, the better the system. Such a method can be applied to single document or multi-document summarization methods.

1.4.3 The TIPSTER SUMMAC text summarization evaluation

SUMMAC was the first large-scale evaluation of automatic text summarization system in 1998. The objective is to judge the performance of summarizers by different extrinsic evaluation tasks. The corpora involve newswire texts in English. This project proposed three evaluation tasks: (i) the adhoc

task; (ii) the categorization task; and (iii) the question-answering task. For each task, participants submit two kinds of summaries. One is 10% in length of the source and the other is unlimited length.

(i) The adhoc task

This task focused on summaries, which were tailored to a particular topic with human judgment. Given a summary and a topic description, professional analysts were asked to decide whether a summary was relevant to the topic. The full-text was fed into the retrieval system to obtain the ground-truth relevance. The summary was accurate if it had the same relevance with the corresponding full-text document.

(ii) The categorization task

The categorization task aimed to judge whether a summary contained sufficient information to enable analysts to categorize it to the appropriate topic quickly and correctly. First, a summary and five topics with topic descriptions were given. Analysts then chose one of the categories for the summary. Finally, the recall and precision of relevant documents were counted to generate f-scores as the final metric.

(iii) The question-answering task

This evaluation measured the degree to which a summary contained information in answering a set of topic-related questions. Human analysts based on common guidelines prepared a set of topic-related questions and marked the corresponding text segments as a key that might answer those questions. Next, comparing the summary against a set of key answers manually, analysts gave three kinds of judgments: correct, partially correct or missing. Accuracy metrics, Answer Recall Lenient (ARL) and Answer Recall Strict (ARS), were used to measure summarization performance.

In conclusion, the evaluation methods proposed by SUMMAC required much human efforts. Furthermore, they were designed for mono-lingual (English) summarization evaluation only.

1.4.4 Text Summarization Challenge (TSC)

Text Summarization Challenge (TSC) intends to investigate text summarization techniques. TSC adopted intrinsic and extrinsic evaluation methods. The summarizing task and evaluation are based on Japanese texts only.

For intrinsic evaluation of extract-based summaries, key sentences were firstly marked by human annotators. The number of extracted sentences marked as important was then computed. For abstract-based summaries,

subjective evaluation by humans was used. They gave assessment in terms of readability, content, and acceptability of the generated summaries.

For extrinsic evaluation, TSC made use of the question-answering evaluation method similar to what SUMMAC did.

1.5 Research contributions

1.5.1 Text summarization based on thematic term approach

In the first stage of our research, we developed a domain independent, single document summarization system which can generate informative extracts (sentences) by considering thematic terms discovered from the corpus and within an article [26]. Important sentences governed by thematic terms are extracted. Next, we demonstrated the feasibility of our thematic term approach using a Chinese corpus covering different domains.

Furthermore, we investigated summarization evaluation using a content-based method. A traditional content-based evaluation method could not be used since no handwritten summary was available for references. In view of this, we designed an evaluation scheme which made use of an information retrieval model. This evaluation method attempts to measure the represen-

tative power of a summary for its original text.

1.5.2 Bilingual news summarization based on an event-driven approach

Our next research contribution targeted summarization of news articles based on an event-driven approach. Bilingual summaries related to events under a particular topic can be generated. To deal with both Chinese and English news stories, dictionary-based term translation is employed. Specifically, it consists of two steps. The first step is the phrase translation method and the other is the translation term disambiguation method.

Next, unsupervised learning is used to discover events and generate coherent event clusters. After that, heuristic criteria are used to select relevant and cohesive clusters in building an event list and the content for the summary.

To demonstrate the feasibility of our bilingual event-driven summarization technique, a corpus over a different domain was used in the evaluation. This corpus is the press release from the Hong Kong SAR government. It is a parallel corpus comprised of English and Chinese documents. We investigated the effectiveness of our summarizer in retrieving the on-event and parallel sentences.

1.6 Thesis organization

Chapter 2 presents the architecture of single document summarization using thematic term approach. Chapter 3 describes two evaluation methods for summarization: the content-based similarity measure and the average inverse rank (AIR) method. Chapter 4 presents our work on an event-driven approach to generate bilingual summaries. Chapter 5 demonstrates the feasibility of our event-driven summarization approach over a parallel Chinese / English corpus. Chapter 6 gives the conclusions and future work.

Chapter 2

Text Summarization based on a Thematic Term Approach

In this chapter, we describe a text summarization approach based on the notion of thematic terms. It is a language independent summarization framework which extracts representative sentences from the original text article based on automatically identified thematic terms. We consider two kinds of thematic terms, namely, corpus-based thematic terms and article-based thematic terms. These thematic terms are determined using information retrieval techniques. Unlike many existing summarization methods, our approach considers not only the information contained in a single article, but also the information produced by the whole corpus. We attempt to apply our summarization framework on both English and Chinese documents.

2.1 System overview

The overall architecture of our thematic term approach is shown in Figure 2.1. It is composed of four major modules, namely, the document preprocessor, the corpus thematic term extractor, the article thematic term extractor, and the sentence score generator. Raw texts, in various formats such as XML, are first passed into the document preprocessor. It converts the raw texts to a suitable representation for subsequent processing. The purpose of the corpus thematic term extractor and the article thematic term extractor is to extract corpus-based and article-based thematic terms respectively. Corpus-based thematic terms capture the main coverage of the whole corpus; and article-based thematic terms capture the main theme of a single document. Next, representative sentences are determined by considering both kinds of thematic terms. The sentence score generator is responsible for generating the score for each sentence. A summary at a specific compression rate is finally obtained by selecting those sentences with high scores.

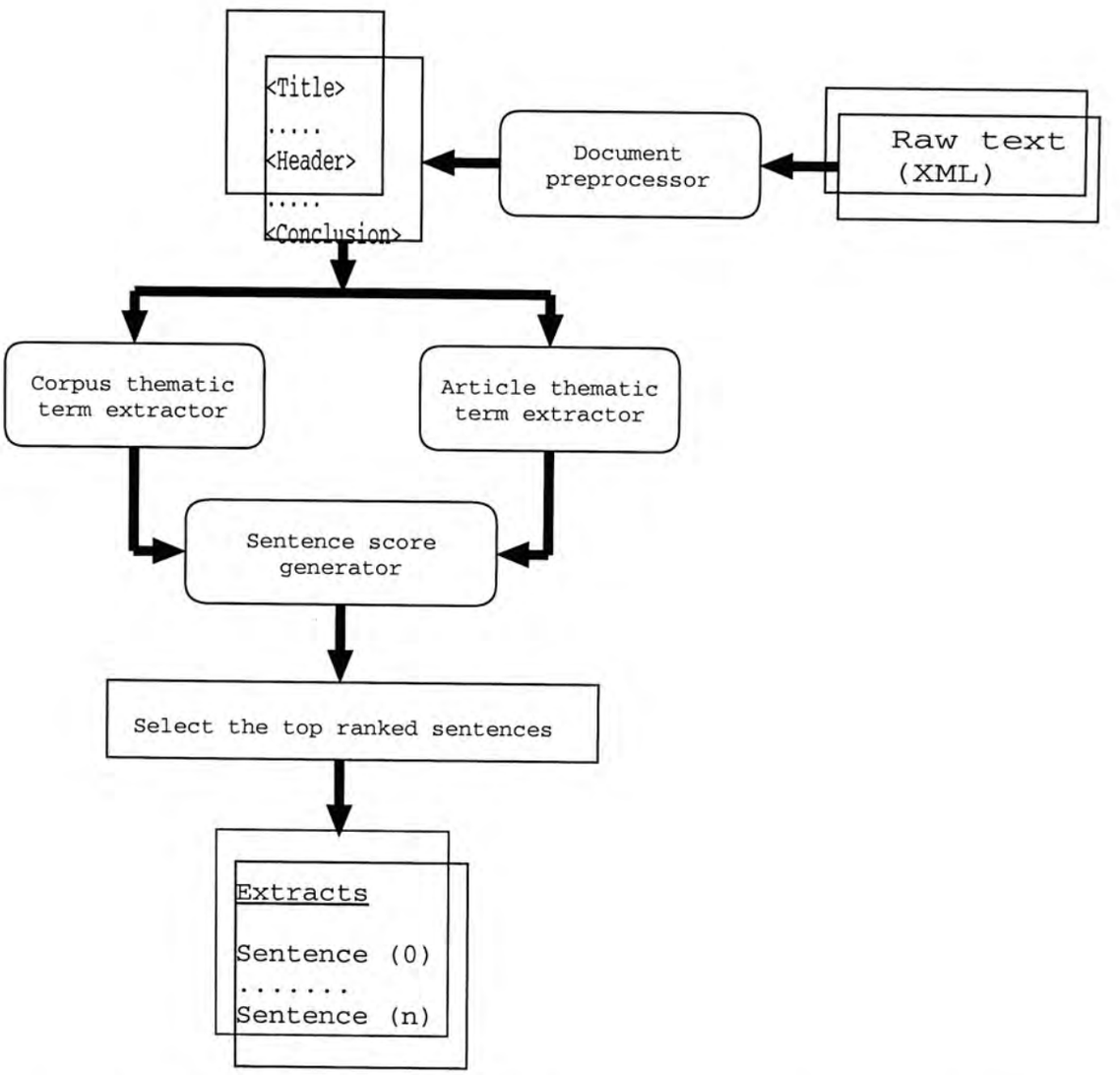


Figure 2.1: *The overall architecture of thematic term approach for text summarization.*

2.2 Document preprocessor

2.2.1 English corpus

The English corpus used in our experiment is the Computation and Language (cmp-lg) corpus[1]. It was prepared by MITRE Corporation based on extensions to some initial work carried out at the University of Edinburgh. There are 183 scientific papers in XML format and they are made available as a general resource to the information retrieval, extraction, and summarization communities. Mark-up tags specify information such as title as well as basic structure such as abstract, body, sections, lists, etc. Figures, tables, equations, cross-references, and references are all replaced by placeholder tags. Table 2.1 shows an excerpt of a sample document (9504018.xml).

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/cmp_lg.html

```
<?xml version='1.0'?>
<!DOCTYPE MINIMAL-DOC SYSTEM "mini.dtd">
<MINIMAL-DOC>
<TITLE>An Implemented Formalism for Computing
Linguistic Presuppositions and Existential Commitments</TITLE>
<ABSTRACT>
<P>
We rely on the strength of linguistic and philosophical perspectives
in constructing a framework that offers a unified explanation for
...
<P>
</ABSTRACT>
<BODY>
<DIV ID="1" DEPTH="1" R-NO="1"><HEADER> Introduction
</HEADER>
<P>
It is common knowledge that a rational agent is inclined to
presuppose defeated by some common sense knowledge
...
strength of both perspectives.  We achieve this using the following:
<ITEMIZE>
<ITEM>a set of methodological principles that unify
...
</ITEM>
<ITEM>an extension of stratified logic <REF/>
where the quantifiers are read under
Lejewski's <REF/> ``unrestricted interpretation'',
which provides us the formal tool for expressing the above layers.
</ITEM>
</ITEMIZE>
</P>
...
<!-- MATH: $(\forall x)( \neg
dragon(x))$ -->
<EQN/>
...
<IMAGE TYPE="TABLE"/>
...
<DIV ID="6.1" DEPTH="2" R-NO="1"><HEADER>Bibliography
</HEADER>
<P>
J.D. Atlas.
What are negative existence statements about?
Linguistics and Philosophy, 11:373-394, 1988.
...
</DIV>
</BODY>
</MINIMAL-DOC>
```

Table 2.1: *An excerpt of a sample document (9504018.xml) in the English corpus.*

2.2.2 English corpus preprocessor

XML formatted texts are separated into "body" and "abstract" texts. Sentences in both texts are ended with a "." or "?". Abstract text refers to handwritten abstracts. Body text refers to texts excluding abstracts. The body and the abstract are processed by the basic document preprocessing like lemmatization, stemming, case folding, as well as removal of all XML tags, figures, equations, tables, stop words, and punctuation.

The lemmatization process converts an English term into its lemma form. Morphological information such as tense is handled. WordNet [31] was used in our system. Given an English term or phrase as query, WordNet returns the lexicalized form under each syntactic category, namely, noun, verb, adjective, and adverb. We take the first available lexicalized term as the output for the lemma form. For example, if "leaves" is given, the first lexicalized form, i.e., "leaf" in the noun category will be obtained.

Stemming is for suffix-stripping. The stemming algorithm based on Porter [39] was adopted in our system. Various suffixes such as -ed, -ing, -ion, -ions are catered for. After discarding them, a single term in stemmed form is obtained. For example: "connect", "connected", "connecting", "connection", and "connections" share the same stem: "connect". Stemming can reduce the number of terms substantially. The distinction between stemming and

lemmatization is that sometimes stemming results the wrong spelling English terms such as "provid" and "relat". However, lemmatization must result in correct spelling terms.

Case folding is simply done by converting all characters into lower case. Besides, in our corpus, tags and figures do not carry informative text data. Equations are used to represent mathematical expressions. Tables are mostly used to depict numeric results or examples. Since all these elements do not contribute to the summarizing tasks, we remove them in data preprocessing.

We collected 629 words in our stop word list, including English terms and punctuation (see Appendix A). Stop words are discarded from the documents. This can usually filter unmeaningful terms and reduce subsequent processing time.

2.2.3 Chinese corpus

The Chinese corpus used in our experiment comes from the Topic Detection and Tracking Evaluation Project (TDT) organized by DARPA and NIST. The corpus contains news data collected daily from 3 news sources in two languages (American English and Mandarin Chinese). Chinese sources are Xinhua News Agency, Zaobao News Agency, and Voice of America. The texts are encoded in GB. We select news articles related to economic crisis in

our experiments for there are many Chinese news articles under this topic.

2.2.4 Chinese corpus preprocessor

We use "。" (Chinese period) and "？" to indicate sentence boundary. Each article is subjected to word segmentation. The word segmentation module obtained from the Linguistic Data Consortium (LDC) is used. It makes use of dynamic programming to find the path which has the highest multiple of word probabilities. In general, this works well in most Chinese texts except for some unknown proper names.

After segmentation, we then conduct the removal of stop words. There are 357 common Chinese terms and punctuation in our stop word list (see Appendix B).

2.3 Corpus thematic term extractor

The objective of the corpus thematic term extractor is to determine a set of corpus-based thematic terms. We use the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme to find the salient terms from the corpus. The idea of inverse document frequency is that a term appeared in a few documents is likely to be a better discriminator than a term appeared in almost all documents. The inverse document frequency of a term j is given

by $log_2(N_c/D_j)$ where D_j is the total number of documents containing term j and N_c refers to the total number of documents in a corpus. In addition, we consider term frequency F_j defined as the total number of occurrence of term j in the whole corpus and is further normalized by maximum term frequency found within all distinct terms. The weight, C_j, of term j is given by

$$C_j = F_j \cdot log_2(\frac{N_c}{D_j}) \tag{2.1}$$

After calculating the weight for all the distinct terms in the corpus, the terms will be ranked according to the weight in descending order. We use the threshold, H_c, to select the top-ranking terms and form the set of corpus-based thematic terms. The corpus-based thematic terms are capable of reflecting the important topics within a corpus. Examples of corpus-based thematic terms are shown in Table 2.2 and Table 2.3. These terms can indicate the nature of the corpus to some extent. For Table 2.2, the range of final weight is between 0.004 (terms like "alike" and "ahead") to 0.6 (term likes "tag"). For Table 2.3, the range of final weight is between 0.007 (terms like "心情" and "心动") to 0.359 (term likes "印尼").

Corpus-based thematic terms	Total number of documents containing this term	Total number of occurrences of the term in entire corpus	Final weight (maximum term frequency) is 1204)
tag	27	606	0.600
grammar	50	987	0.473
synchronous	1	59	0.220
temporal	10	102	0.185
anaphor	4	66	0.170
bigram	7	65	0.137
part-of-speech	14	85	0.131

Table 2.2: *Samples of corpus-based thematic terms from our English corpus.*

Corpus-based thematic terms	Total number of documents containing this term	Total number of occurrences of the term in entire corpus	Final weight (maximum term frequency) is 649)
股票	16	82	0.217
国际货币基金	42	126	0.146
指数	16	44	0.116
风暴	39	79	0.100
下跌	33	58	0.089

Table 2.3: *Samples of corpus-based thematic terms from our Chinese corpus.*

2.4 Article thematic term extractor

If we want to generate a representative summary for a document, we should also consider the main theme in the particular document. Intuitively, key terms found from a document are also important for summarization. Inspired by the idea in [35], we consider the inverse sentence frequency defined as $log_2(N_s/S_j)$ where S_j refers to the total number of sentences containing term

j and N_s refers to the total number of sentences within a document. In addition, we also consider T_j which is the total number of occurrences of term j in a document normalized by the maximum term frequency of all distinct terms found within a document. As a result, the weight, A_j, denoting the article thematic term weight, is given as follows:

$$A_j = T_j \cdot log_2(\frac{N_s}{S_j}) \tag{2.2}$$

After calculating the article thematic term weights of all distinct words found in a document, we rank them in descending order. Then we select those terms with weight larger than a threshold H_a, to form the article-based thematic terms. As a result, if a term is so popular that it occurs in almost every sentence in a document, then its inverse sentence frequency diminishes. If this term appears frequently in an article, it may be an important term as reflected by T_j. Examples of article thematic terms are shown in Table 2.4 and Table 2.5. There are 168 sentences in 9504008.xml and 27 sentences in 19980521_2000_2356_XIN_MAN_0100.sent.txt. H_a is set to 5% (i.e. 5% of top-ranking terms among article will be selected as article-based thematic terms) in both cases. For Table 2.4, the range of final weight is between 0.107 (terms like "ai" and "aj") to 1.690 (term likes "language"). For Table 2.5, the range of final weight is between 0.194 (terms like "其次" and "连连")

to 0.876 (term likes "美元").

Article-based thematic terms	**Total number of sentences containing this term**	**Total number of occurrences of the term in a document**	**Final weight (maximum term frequency) is 48)**
language	31	48	1.690
speech	31	39	1.373
processing	27	34	1.295
korean	28	31	1.157

Table 2.4: *Samples of article thematic terms from an English document (9504008.xml with title "SKOPE: A connectionist/ symbolic architecture of spoken Korean processing").*

Article-based thematic terms	**Total number of sentences containing this term**	**Total number of occurrences of the term in a document**	**Final weight (maximum term frequency) is 17)**
美元	10	15	0.876
经济	6	9	0.796
金融	6	6	0.531
汇率	3	4	0.517
外汇储备	4	4	0.449

Table 2.5: *Samples of article thematic terms from a Chinese document (19980521_2000_2356_XIN_MAN_0100.sent.txt).*

2.5 Sentence score generator

After corpus-based thematic terms are extracted as discussed in Section 2.3, we process each document to generate a summary. For each document, we obtain article-based thematic terms as discussed in Section 2.4. Based on the selected corpus-based and article-based thematic terms, we calculate the score for each sentence. Consider a term j appeared in a particular sentence. We first retrieve its inverse document frequency defined as $log_2(N_c/D_j)$, where N_c is the total number of documents in the collection and D_j is the total number of documents containing term j. We then normalize it by the maximum value obtained among all the inverse document frequencies of different terms across the corpus. Suppose the normalized inverse document frequency is I_j. We retrieve the inverse sentence frequency defined as $log_2(N_s/S_j)$, where N_s refers to the total number of sentences in a document and S_j is the total number of sentences containing term j. Let the normalized inverse sentence frequency be L_j. It is computed by dividing $log_2(N_s/S_j)$ with the maximum value found among distinct terms across a document. The score, s_k, for each sentence, k, is given as follows:

$$s_k = \beta \sum_{j=0}^{B_k} (T_j \cdot I_j) + (1 - \beta) \sum_{j=0}^{B_k} (T_j \cdot L_j) \tag{2.3}$$

where B_k is the total number of unique terms in sentence k and T_j is the

total number of occurrence of term j found within sentence k. We introduce β as the balancing factor between the weight of term j found from the whole corpus and found in an article. If terms found in the sentence are salient for the corpus as well as the article, then this sentence gains a high score. It implies that this sentence is likely representative of the article and of the corpus as well. After the score of each sentence is computed, we rank all sentences in descending order.

For summary retrieval, a user typically specifies a compression rate indicating the desired amount of text. Compression rate is defined as the ratio of the summary length to the full-length documents measured in sentences. At a given compression rate, the top-ranking sentences are selected. The sentences are then arranged according to the chronological order in the original article to form a summary.

2.6 Chapter summary

In this chapter, we have described the thematic term approach on single document text summarization. A new approach on sentence extractive summary is introduced. Our method considers both thematic terms found from the entire corpus and in the article as well. Terms capturing the main theme of the corpus are extracted automatically as our corpus-based thematic terms.

On top of that, key terms discovered within an article form the article-based thematic terms. We assign each sentence with a score generated by considering both kinds of thematic terms. Sentences with high scores will be collected to form the extractive summary. This method is language independent and we apply it on both Chinese and English corpora separately. Evaluation on our thematic term approach are discussed in Chapter 3.

Chapter 3

Evaluation for Summarization using the Thematic Term Approach

In this chapter, we present two evaluation methods for summarization based on thematic term approach proposed in the previous chapter. The first evaluation method uses a content-based measure. This measure belongs to the intrinsic type of summary evaluation method and is widely adopted by many researchers [14, 34]. The objective is to evaluate a summary based on the number of overlapping terms between the automatic generated summary and its handwritten counterpart. This method is suitable for our English corpus

since manual abstracts for all documents are available. However, it could have been infeasible otherwise. Furthermore, if a writer uses different vocabularies to construct abstracts, there may not be any overlapping terms between the generated summary and the handwritten one. This would create another infeasible situation. To avoid these cases, we propose a new extrinsic evaluation method, called the Average Inverse Rank (AIR). AIR aims at evaluating the degree that the extracted summary can serve as a surrogate for its original document in an entire collection. In this thesis, we use AIR as the second method to evaluate thematic term based summarization.

3.1 Content-based similarity measure

Content-based similarity measure is a popular evaluation method for extractive summaries. It evaluates in terms of the degree of term overlapping between an extract and a "standard" summary such as a handwritten abstract. Many researchers [14, 34] have applied this method in assessing the quality between system extracted summaries and handwritten abstracts. We illustrate the content-based similarity method in Figure 3.1.

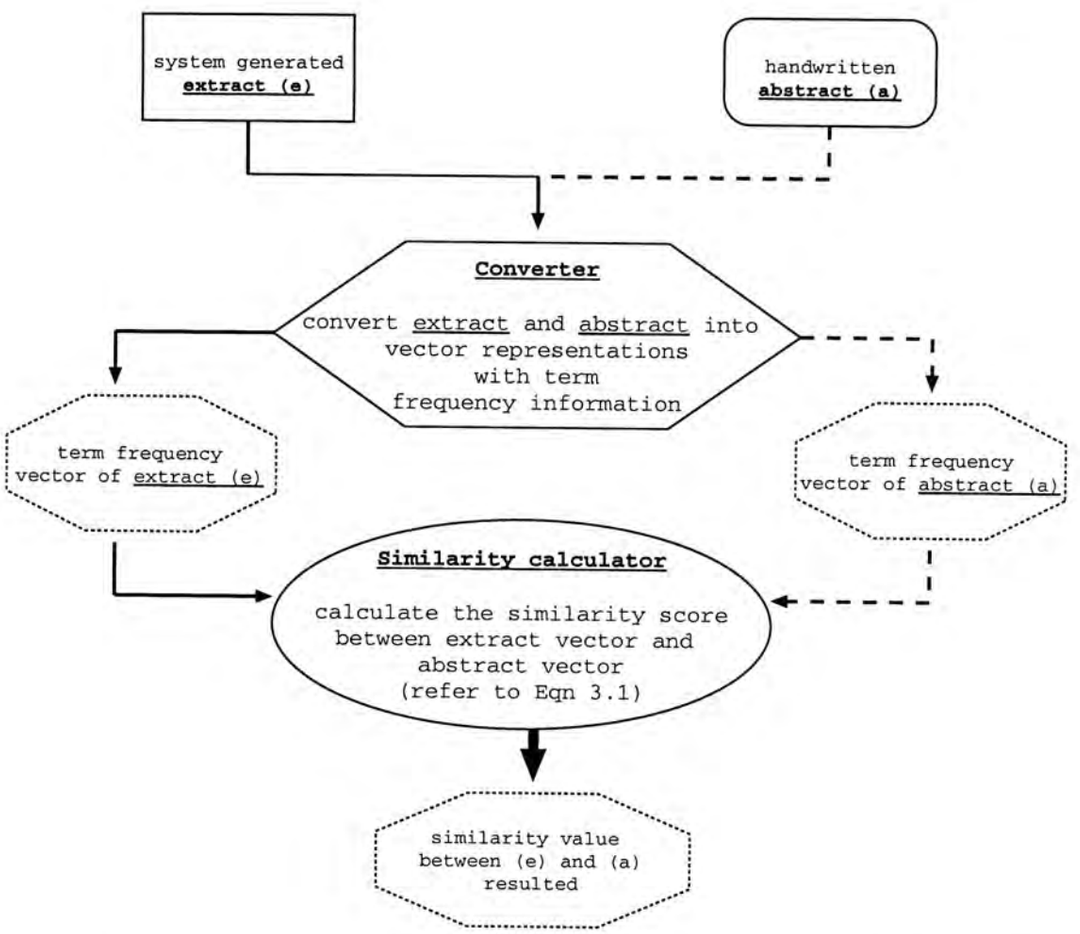


Figure 3.1: *Content-based similarity method for evaluating thematic term summarization.*

System extracts and handwritten abstracts are first converted to vector representations. Both vectors carry distinct terms and weights obtained from their texts. Next, we compute the cosine similarity measure to get the similarity score. If the system extract and the handwritten abstract share more common terms which are important, then the cosine similarity score tends to 1.

More precisely, the weight of term j, $w_{j,e}$, in the vector representation of extract, e, is denoted as:

$$w_{j,e} = f_{j,e} \cdot (IDF_{j,c})$$

where $f_{j,e}$ represents the term frequency of the term j found in extract e. The inverse document frequency of term j, $IDF_{j,c}$, represents the importance of this term considering the entire data collection c. Similarly, the weight of term j in the vector representation of abstract, a, is represented as:

$$w_{j,a} = f_{j,a} \cdot (IDF_{j,c})$$

where $f_{j,a}$ is the total number of frequency of term j appeared in the handwritten abstract a. The final similarity score $\Delta(e, a)$ is defined as follows:

$$\Delta(e, a) = \frac{\sum_j (w_{j,e} \cdot w_{j,a})}{\sqrt{\sum_j (w_{j,e})^2 \cdot \sum_j (w_{j,a})^2}} \tag{3.1}$$

3.2 Experiments using content-based similarity measure

3.2.1 English corpus and parameter training

As mentioned above, content-based similarity measure requires handwritten summaries in the evaluation process. Our English corpus contains a handwritten abstract for each document. Therefore, we could carry out the evaluation experiment.

We divided the entire English corpus into two portions, namely, the training and testing portions. The training portion consisted of half of the whole corpus and it was used for parameter training. The parameter to be tuned was the balancing factor, β, in generating the sentence weights (see Eqn 2.3). The testing portion was composed of the remaining half of the corpus which was used for evaluation purpose. Some statistics of these two portions are shown in Table 3.1.

Furthermore, we investigated the results of our summaries when different fractions of training and testing sets were used. The size of training portion varied from 10% to 90% in 10% interval. Correspondingly the testing portion varied from 90% to 10% in 10% interval. As a result, nine sets of experiments were conducted.

	Training set (50%)		Testing set (50%)	
	Body	**Abstract**	**Body**	**Abstract**
Total number of documents	89	89	89	89
Total number of sentences	14,876	536	16,945	495
Average number of sentences	167.1	6.0	190.4	5.6

Table 3.1: *Statistics of the training set and the testing set of the English corpus.*

In order to give more accurate result using content-based similarity evaluation method on our thematic term approach summarization, we performed three-fold cross-validation experiments. The entire corpus was divided into three portions which were approximately equal in the total number of articles. We then tuned the balancing factor, β, on first two portions and performed test on the remaining one. Three runs were conducted by alternately replacing different two portions for training and one portion for testing. After averaging the similarity results of three runs, the final average similarity score with the setting of two folds (2/3) in training and one fold (1/3) for testing was computed. We also generated another averaged result by performing three runs which based on one fold (1/3) in training and two folds (2/3) in testing.

At a given compression rate, we varied β in generating summary using our summarization approach. (Recall that β is the balancing factor between

the weight of terms found from the whole corpus and found in an article.) The average similarity scores of all trials was recorded. We then selected the β value that produced the best similarity score. The chosen β value would then be used in the open test (see Section 3.2.2). We repeated the above parameter training and testing for different threshold values H_c for corpus-based thematic terms and H_a for article-based thematic terms, at 5%, 25%, and 45%. This allowed us to investigate the effect of these threshold values. Finally, we also investigated different compression rates of 1%, 5%, 10%, 20%, 30%, 40%, and 50%.

To conduct a comparative study, we introduced a simple summarization method to act as the baseline. Given a specified compression rate, the baseline summary was generated by randomly selecting sentences from the original document. We conducted ten runs to give the averaged baseline result for fair comparison in each experiment setting.

3.2.2 Experimental results using content-based similarity measure

In Figures 3.2, 3.3 and 3.4, we plot three graphs depicting the average similarity score for different thresholds of corpus-based thematic terms and article-based thematic terms under different compression rates. From the results

shown in all figures, summaries generated by our thematic term approach always outperform the baseline result, showing that the extracts contain useful texts. This confirms that extracts generated by considering both corpus thematic terms and article thematic terms are able to provide more representative information. In Figure 3.4, the curve with $H_a = 45\%$ and $H_c = 5\%$ can still maintain a high average similarity score. However, the curve, representing $H_a = 45\%$ and $H_c = 45\%$, gives a relatively inferior performance compared to the others. The difference in the average similarity score varies from 0.02 to 0.07 for the compression rates below 10%. This result shows that a larger threshold for including too many thematic terms may not be good for summarizing task. Some of these terms may not be the thematic terms, which in term affect the quality of the summaries. Moreover, we have also conducted an experiment by setting the compression rate to 100% (i.e. no compression). In this setting, the summary is in fact the entire original article. The average similarity score for the original article is 0.46. Therefore, 0.46 is the upper bound of the similarity score for the summaries.

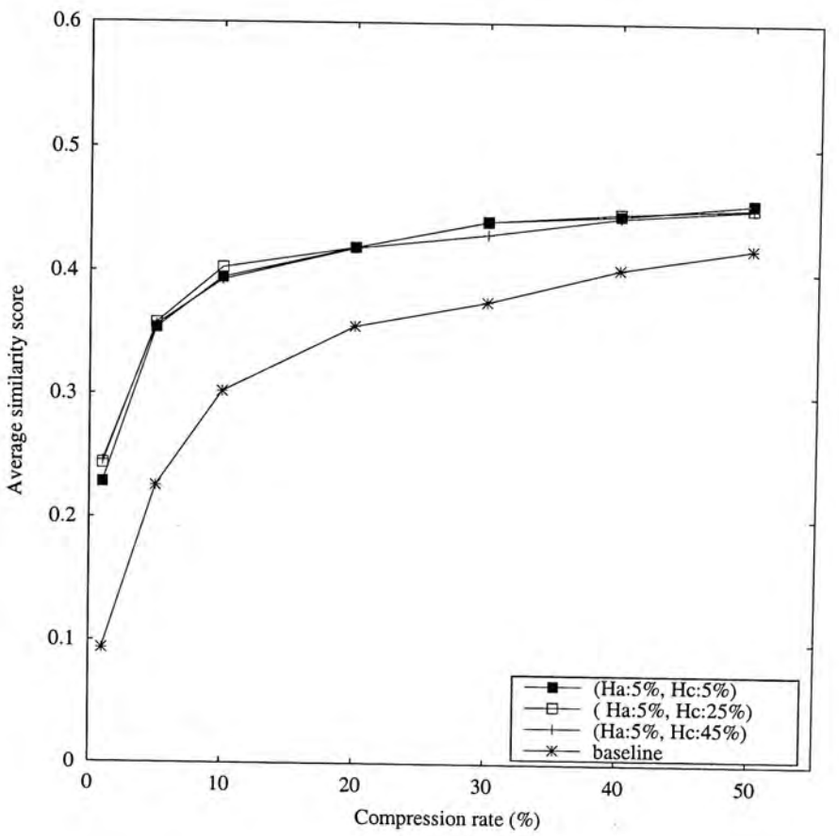


Figure 3.2: *Summaries at different compression rates with $H_a = 5\%$ (50% training to 50% testing).*

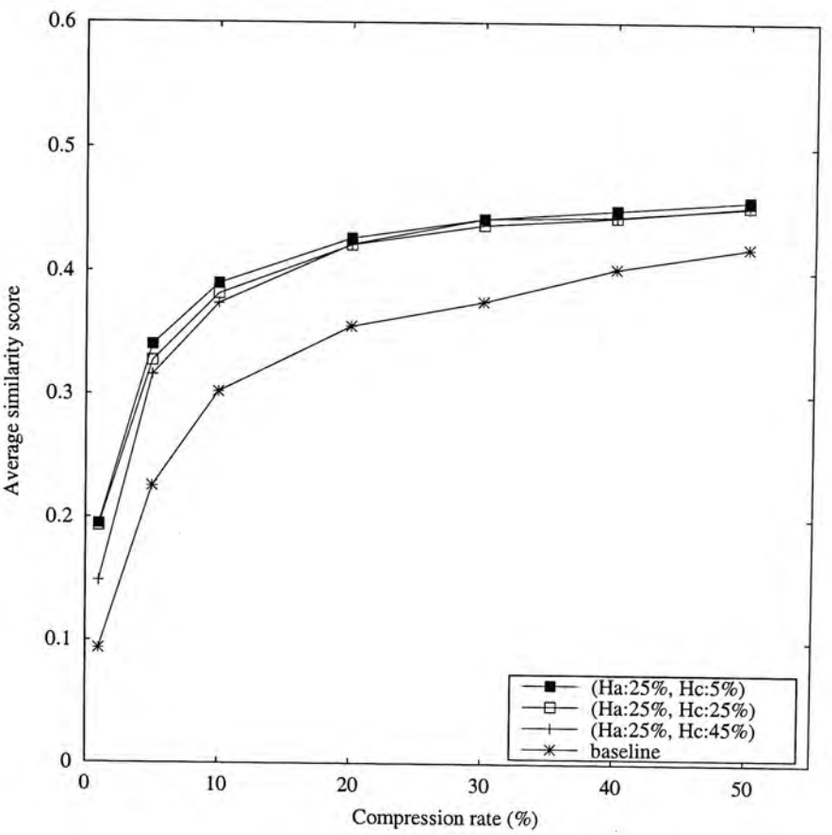


Figure 3.3: *Summaries at different compression rates with H_a = 25% (50% training to 50% testing).*

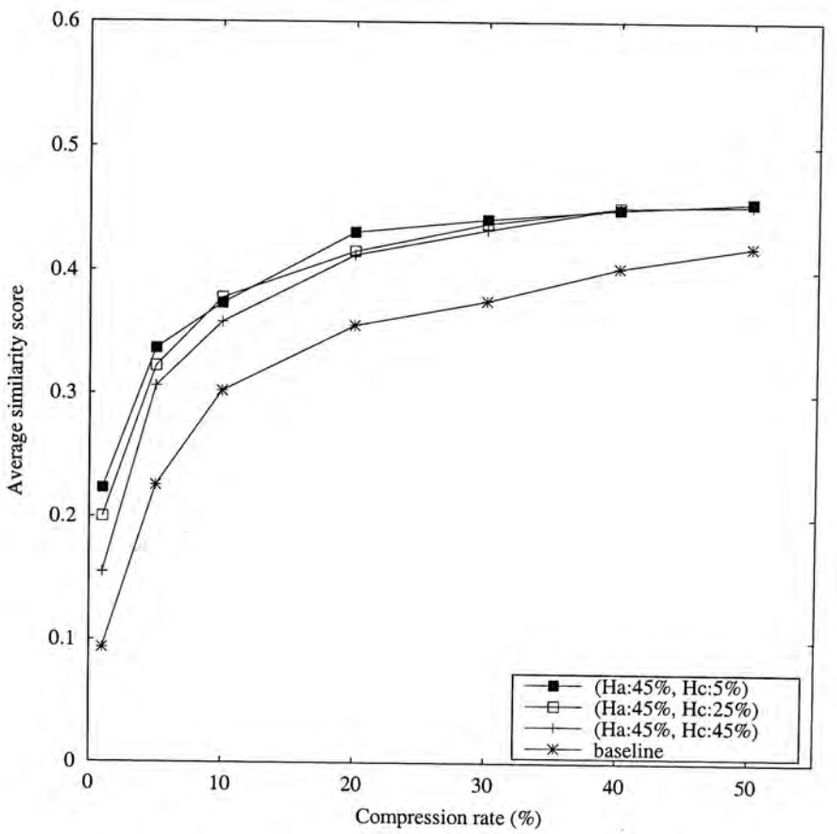


Figure 3.4: *Summaries at different compression rates with $H_a = 45\%$ (50% training to 50% testing).*

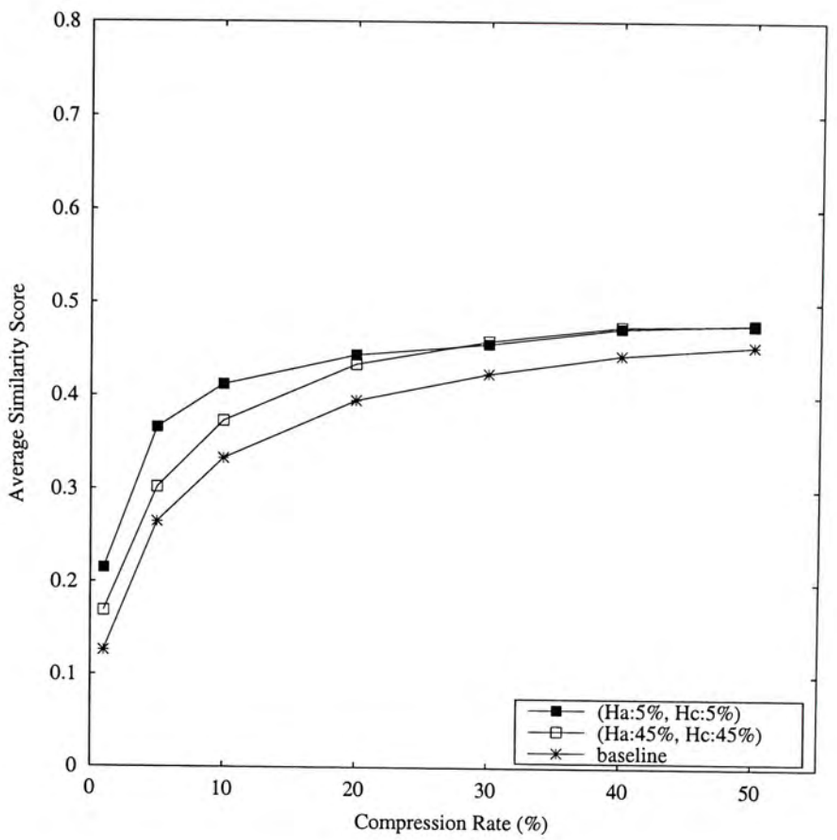


Figure 3.5: *Summaries at different compression rates in 10% training to 90% testing.*

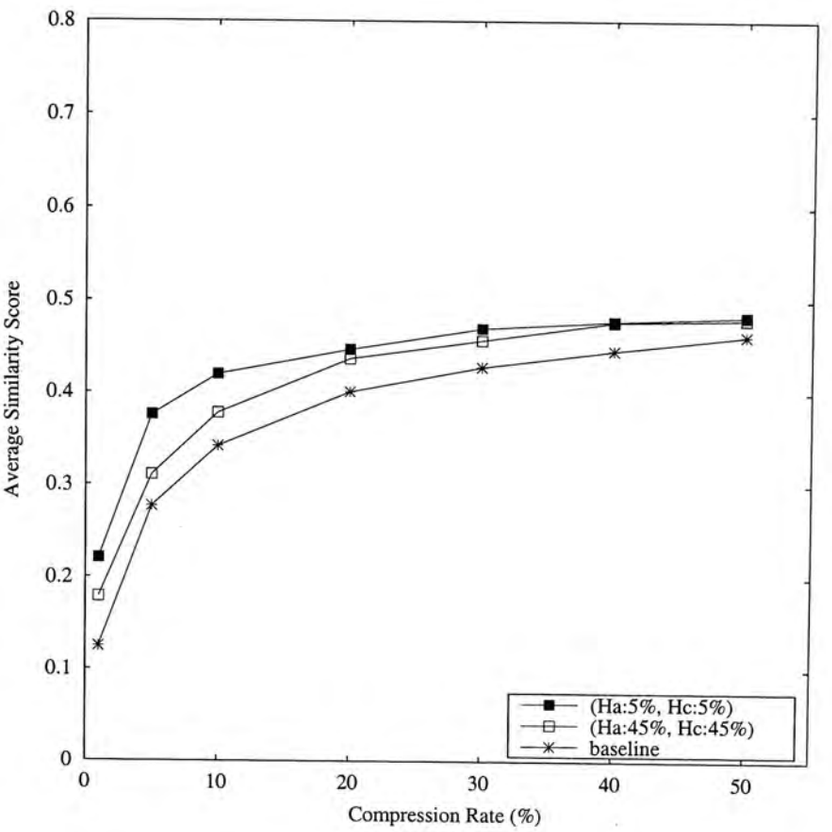


Figure 3.6: *Summaries at different compression rates in 20% training to 80% testing.*

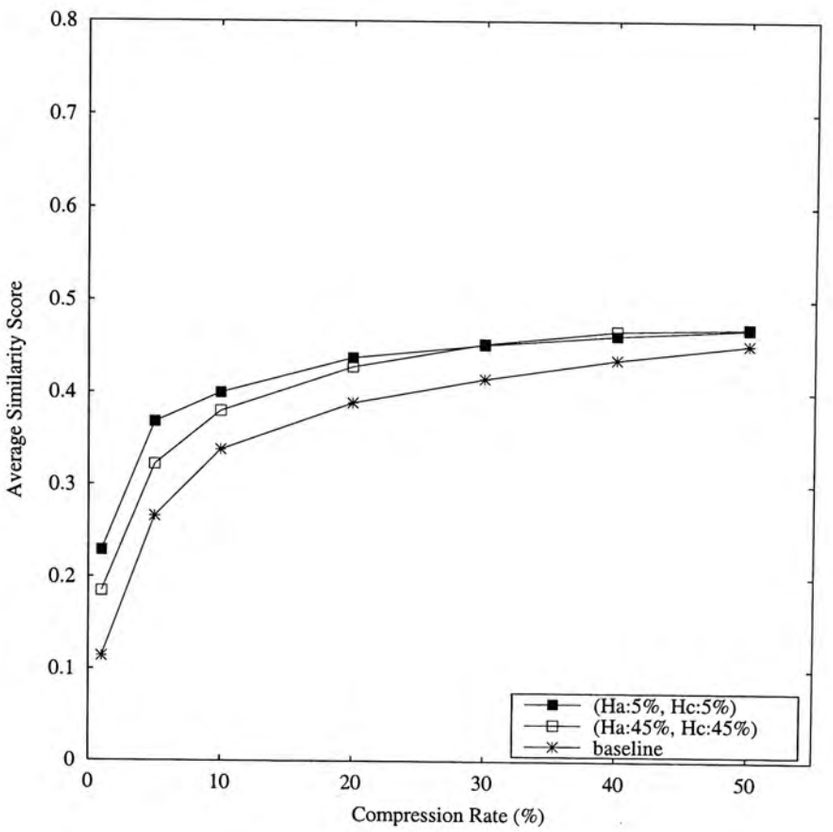


Figure 3.7: *Summaries at different compression rates in 30% training to 70% testing.*

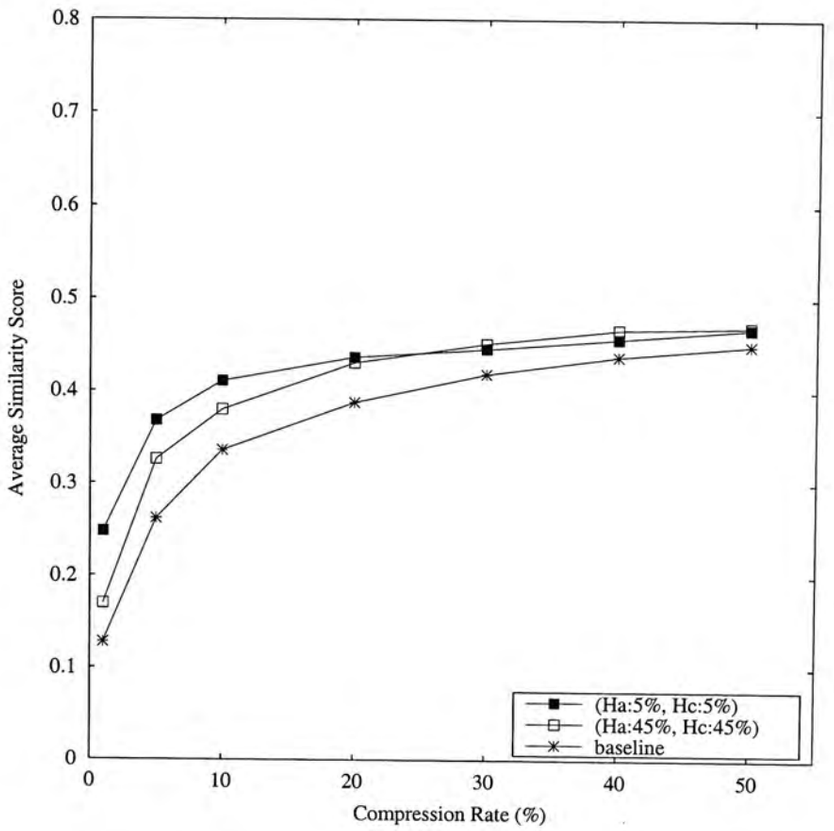


Figure 3.8: *Summaries at different compression rates in 40% training to 60% testing.*

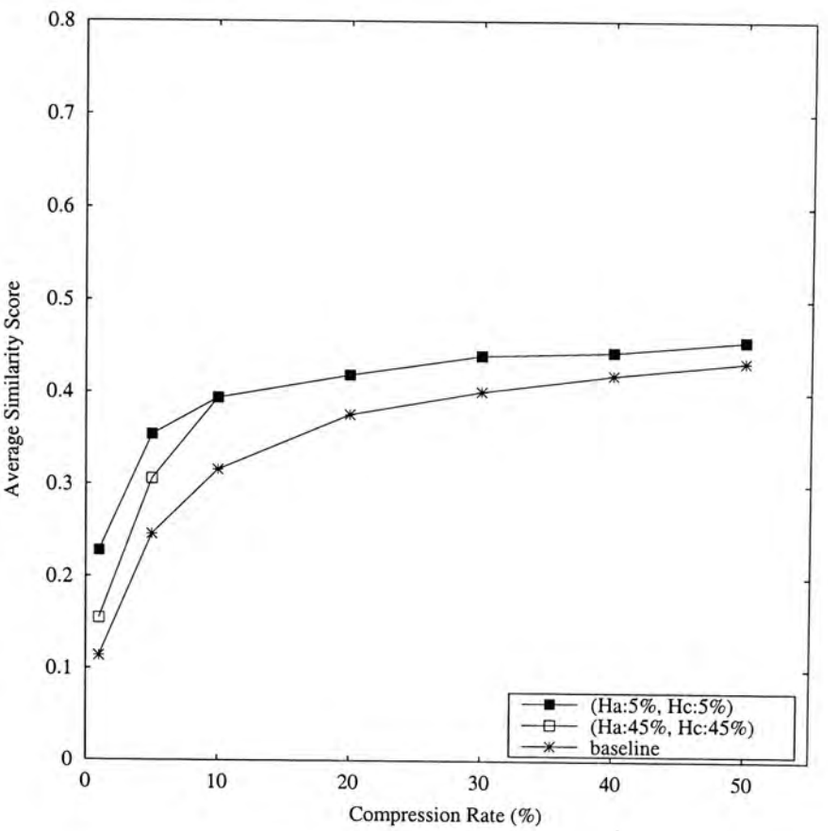


Figure 3.9: *Summaries at different compression rates in 50% training to 50% testing.*

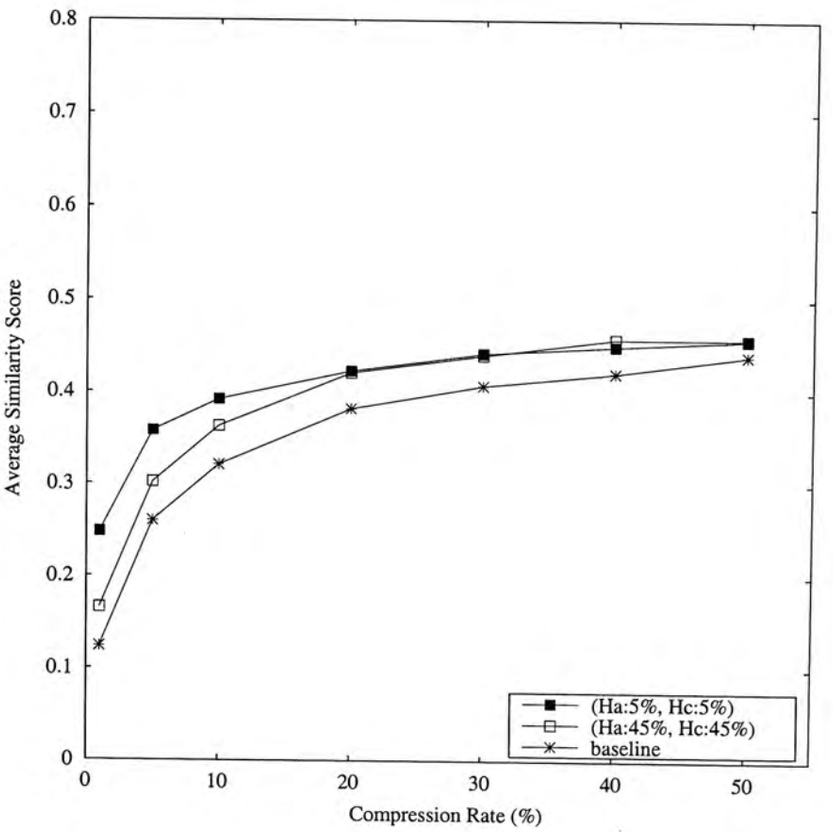


Figure 3.10: *Summaries at different compression rates in 60% training to 40% testing.*

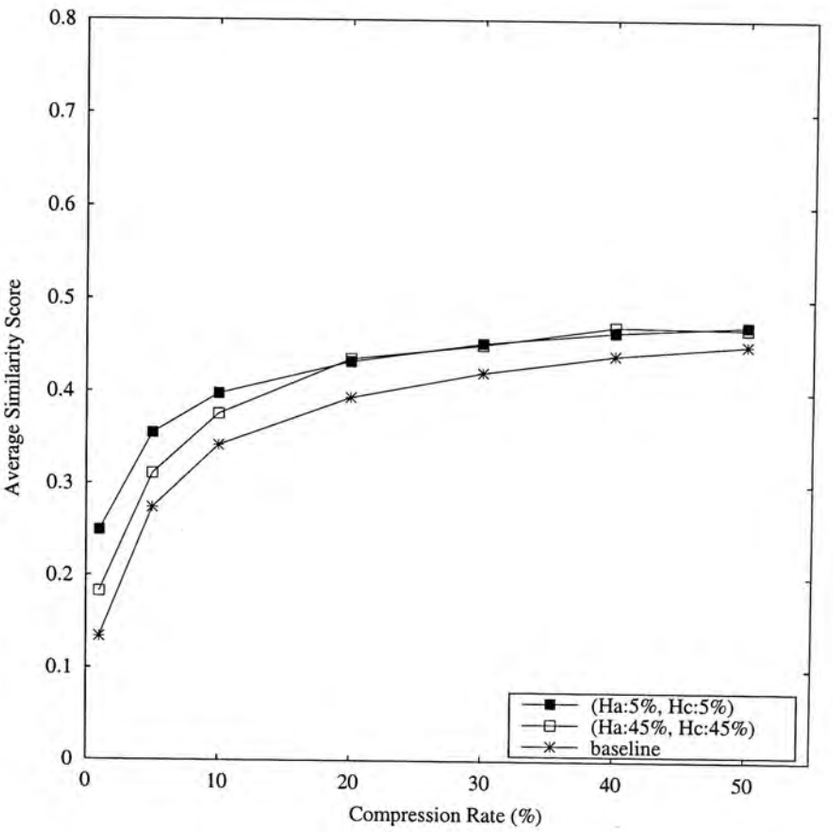


Figure 3.11: *Summaries at different compression rates in 70% training to 30% testing.*

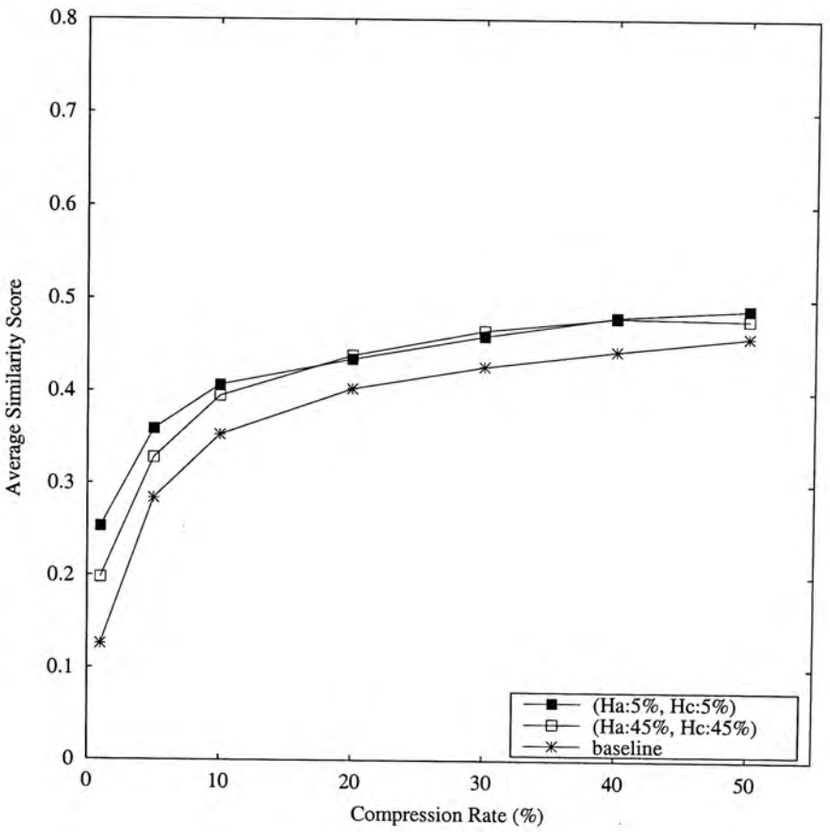


Figure 3.12: *Summaries at different compression rates in 80% training to 20% testing.*

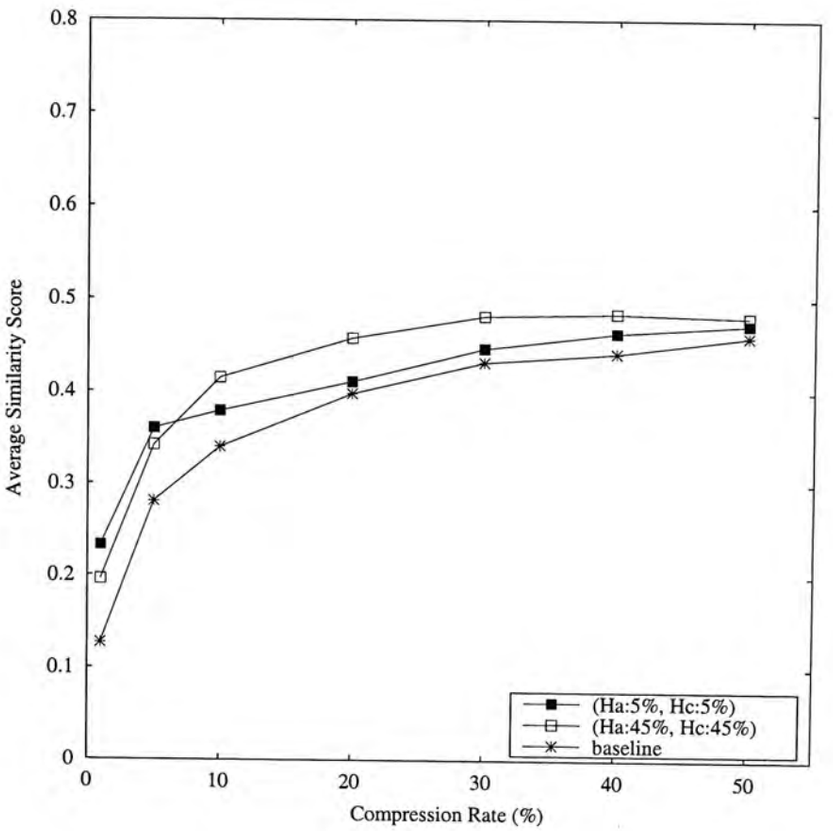


Figure 3.13: *Summaries at different compression rates in 90% training to 10% testing.*

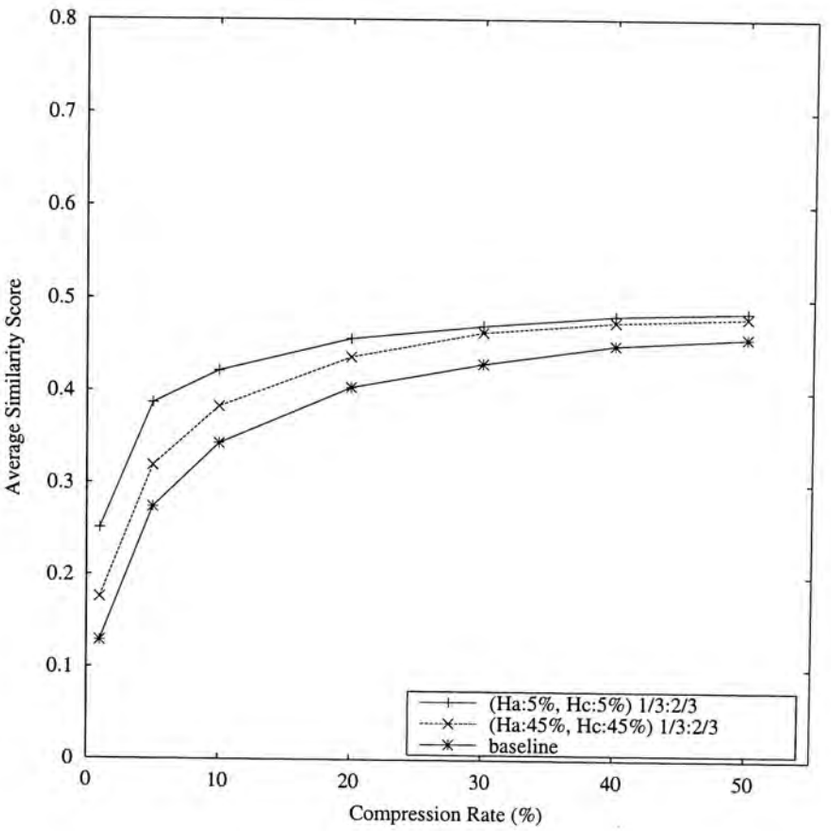


Figure 3.14: *Summaries at different compression rates with 1/3 portion in training to 2/3 portions in testing.*

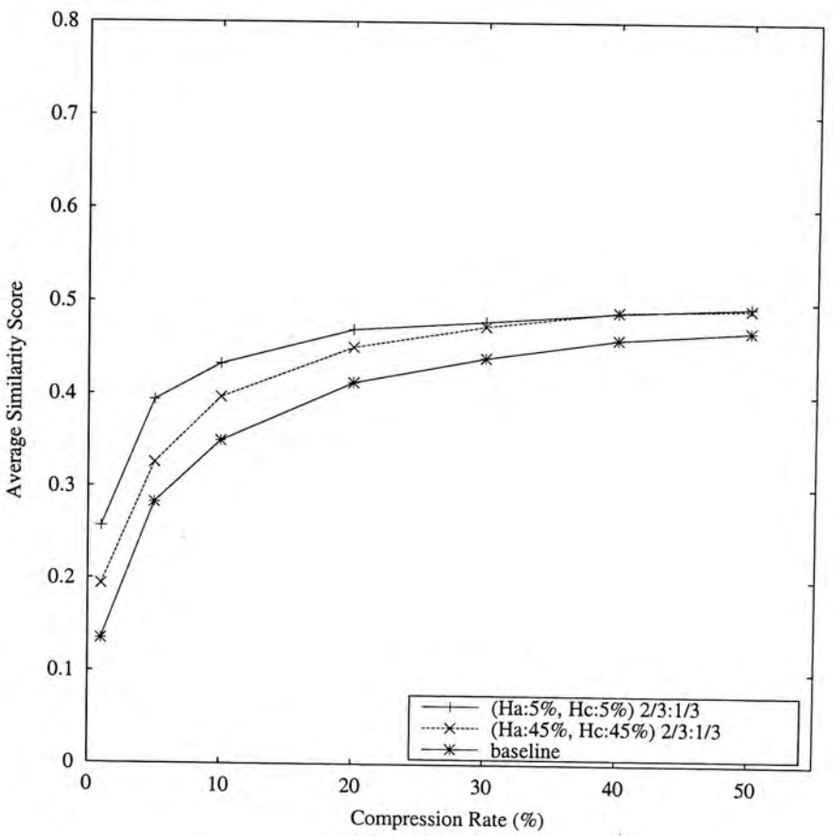


Figure 3.15: *Summaries at different compression rates with 2/3 portions in training to 1/3 portion in testing.*

In Figures 3.5 to 3.15, each figure depicts three curves. They were obtained by running H_a=5% and H_c=5%, H_a=45% and H_c=45%, and the baseline under the specified portions of training and testing sets. All figures show that curves representing H_a=5% and H_c=5%, and H_a=45% H_c=45% always obtained higher similarity scores than the baseline (randomly selecting sentences). It provides a strong evidence that our extracts generated by thematic term approach were able to give more presentable information.

Under 5% compression rate, curves with H_a=5% and H_c=5% produced higher similarity scores than that with H_a=45% and H_c=45% in all figures. Furthermore, curves with H_a=5% and H_c=5% often obtained higher scores than that with H_a=45% and H_c=45% across the 10% training to 80% training (Figures 3.5 to 3.12) and three-fold cross-validation experiments (Figures 3.14 and 3.15). This phenomenon further confirms that small amount of terms considered as thematic terms were enough in summary generation. Too many terms taken in generating sentence weights cannot help extract important sentences.

For the case of 90% training to 10% testing (Figure 3.13), curve with H_a=45% and H_c=45% abnormally produced higher similarity scores than that of H_a=5% and H_c=5% after 5% of compression rate. After looking at the β tuned in H_a=45% and H_c=45% experiment, we found that its value remained at zero for the compression rates over 5%. As the formula for

the sentence weight (see Eqn 2.3), zero β implies that important sentences determined solely by article-based thematic terms. For the case of H_a=5% and H_c=5%, β varied from 0.5 to 0.8. The phenomenon showed that too many samples in training (90% training to 10% testing) produced too many terms (H_c=45%). The sentence extraction would rely on terms coming from articles themselves. On the other hand, too few samples in testing depend less on considering corpus thematic terms in finding the important sentences.

In the Table 3.2, we show a sample summary generated by our thematic term approach. This summary is generated by specifying the compression rate of 5%; with H_a of 5% and H_c of 5%. The underlined sentences are considered representative and informative since they are also found in the handwritten abstract given in Table 3.4. Table 3.3 shows the baseline summary which is generated by randomly selecting sentences from the original document. Moreover, the original document is given in Appendix D for reference. The bracket value at the end of each sentence in Tables 3.2 and 3.3 represents the sentence identification number from its original document.

File:	**9505001.xml**
Title:	**Response Generation in Collaborative Negotiation**
Total length:	**159 sentences**
Summary length:	**7 sentences**
Handwritten abstract length:	**5 sentences**
Generated summary based on thematic term approach (similarity score: 72.4%)	
This paper presents a model for engaging in collaborative negotiation *to resolve conflicts in agents' beliefs about domain knowledge.* (7) *This paper focuses on the evaluation and modification of proposed beliefs, and details a strategy for engaging in collaborative negotiations.* (40) Since *collaborative agents* are expected to engage in effective and efficient dialogues, the system should address the unaccepted belief that it predicts will most quickly resolve the top-level conflict. (89) Thus *the algorithm* recursively applies itself to the evidence proposed as support for _bel which was not accepted by the system (102) Thus, it is important that a collaborative agent selects sufficient and effective, but not excessive, evidence to justify an intended mutual belief. (122) *This paper has presented a computational strategy for engaging in collaborative negotiation to square away conflicts in agents' beliefs . (155)* It also supports effective and efficient dialogues by identifying the focus of modification based on its predicted success in *resolving the conflict* about the top-level belief and by using heuristics motivated by research in social psychology to select a set of evidence to justify the proposed modification of beliefs. (157)	

Table 3.2: *Our system generated summary of 9505001.xml. The bracket (at the end of each sentence) indicates the sentence identification number from the original document. The content of the underlined sentences is also found in the handwritten abstract given in Table 3.4.*

Baseline summary (similarity score: 36%)
Caswey et al. , introduced the idea of utilizing a belief revision mechanism to predict whether a set of evidence is sufficient to change a user's existing belief and to generate responses for information retrieval dialogues in a library domain. (16) Conflict resolution strategies are invoked only if the top-level proposed beliefs are not accepted because if collaborative agents agree on a belief relevant to the domain plan being constructed, it is irrelevant whether they agree on the evidence for that belief. (52) Following Walker's weakest link assumption the strength of the evidence is the weaker of the strength of the belief and the strength of the evidential relationship. (55) The system must first determine whether justification for _bel is needed by predicting whether or not merely informing the user of _bel will be sufficient to convince him of _bel. (124) #1 Dr. Smith is not going on sabbatical next year. (151) If the user accepts the system's utterances, thus satisfying the precondition that the conflict be resolved, Modify-Node can be performed and changes made to the original proposed beliefs. (153) It also supports effective and efficient dialogues by identifying the focus of modification based on its predicted success in resolving the conflict about the top-level belief and by using heuristics motivated by research in social psychology to select a set of evidence to justify the proposed modification of beliefs. (157)

Table 3.3: *Baseline summary of 9505001.xml. The bracket (at the end of each sentence) indicates sentence identification number from the original document.*

Handwritten abstract
In collaborative planning activities, since the agents are autonomous and heterogeneous, it is inevitable that conflicts arise in their beliefs during the planning process. In cases where such conflicts are relevant to the task at hand, the agents should engage in collaborative negotiation as an attempt to square away the discrepancies in their beliefs. This paper presents a computational strategy for detecting conflicts regarding proposed beliefs and for engaging in collaborative negotiation to resolve the conflicts that warrant resolution. Our model is capable of selecting the most effective aspect to address in its pursuit of conflict resolution in cases where multiple conflicts arise, and of selecting appropriate evidence to justify the need for such modification. Furthermore, by capturing the negotiation process in a recursive Propose-Evaluate-Modify cycle of actions, our model can successfully handle embedded negotiation subdialogues.

Table 3.4: *Handwritten abstract of 9505001.xml.*

3.3 Average inverse rank (AIR) method

Since handwritten abstracts were not available for our Chinese corpus, the Average Inverse Rank (AIR) evaluation method was then developed. AIR evaluation metric belongs to the extrinsic type of evaluation method.

Generally, good summaries should be relatively shorter in length compared with the original document and at the same time capture useful information. The idea of AIR is based on how good a summary can act as a query to retrieve the original full-length document. Given a collection of the original documents, if a particular summary can retrieve its original document from the pool of the document collection, then this summary is capable of characterizing the original document. To implement this evaluation technique, we made use of an information retrieval engine known as XSmart [25]. XSmart is an extension of Smart, which is a well-known vector space information retrieval engine [12], to handle both Chinese and English. As shown in Figure 3.16, original documents were first indexed by the indexing engine in XSmart. Each summary was then treated as a query for retrieval. The result of the retrieval was a ranked list of relevant documents. The rank of the document corresponds to each query summary was determined. AIR was derived from the inverse of the rank. Using the same procedure, we obtained the inverse ranks of all the extracts. Finally, we took the average of the

inverse ranks as the evaluation metric.

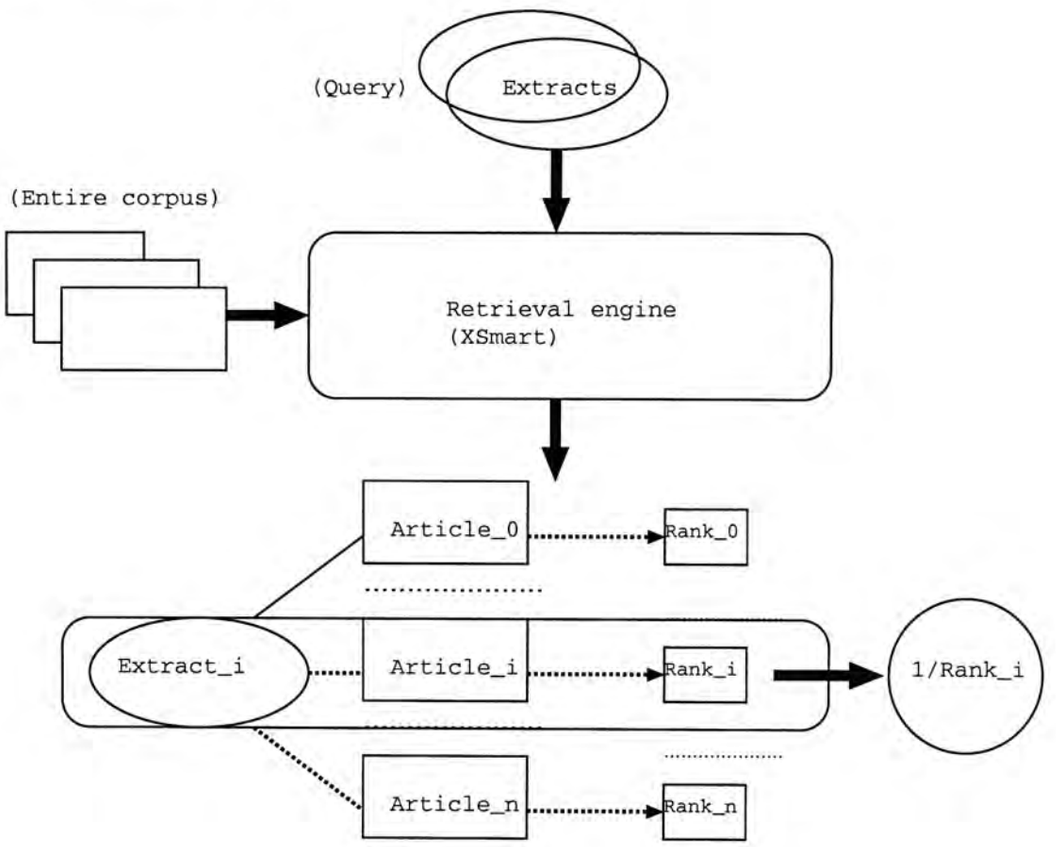


Figure 3.16: *The framework of Average Inverse Rank (AIR) evaluation method.*

3.4 Experiments using average inverse rank method

We have conducted Average Inverse Rank (AIR) evaluation to assess our thematic term summarization method on both Chinese and English documents.

Recall that AIR does not require the availability of standard summaries. In the following sub-sections, we describe the experimental settings and results.

3.4.1 Corpora and parameter training

AIR evaluations have been conducted on both English corpus and Chinese corpus separately. Each corpus was partitioned into two sets, i.e. the training set and testing set. The statistics of each corpus are shown in Table 3.5. We also conducted three-fold cross-validation in English corpus by splitting the entire corpus into three portions. Averaging the results of three experiments using two folds as training and one fold as testing. Another three runs were produced by using one fold as training and two folds as testing.

	English corpus		**Chinese corpus**	
	Training set (50%)	**Testing set (50%)**	**Training set (50%)**	**Testing set (50%)**
Total number of documents	89	89	89	89
Total number of sentences	14,876	16,945	2,805	3,071
Average number of sentences	167.1	190.4	31.5	34.5

Table 3.5: *Corpora statistics for AIR experiments in 50% training to 50% testing.*

The training set is used to tune the β value. (Recall that β is the balancing factor between the weight of terms found from the whole corpus and that

found in an article (see Section 2.5). Given a specified compression rate, we generated a summary using our thematic term approach and evaluated it using the AIR method. We varied β from 0.0 to 1.0 in 0.1 intervals. The best β was obtained by selecting the highest AIR score among all trials. The tuned β value would then be used in the open evaluation using the testing set.

We repeated the above experiments under different threshold H_a, and threshold H_c, at 5%, 25%, and 45%. This enabled us to investigate the effect of these threshold values. Finally, we also conducted experiments under different compression rates.

3.4.2 Experimental results using AIR method

Figures 3.17 to 3.21, show four plots depicting the AIR score under different threshold values as well as compression rates of the English and Chinese corpus respectively.

The three-fold cross-validation experiments are shown in Figures 3.18, 3.19 and 3.20. There are notations of 1/3:2/3 and 2/3:1/3 in these figures. The first notation (1/3:2/3) indicates that one portion is used as training and the remaining two portions as testing. Similarly, the second notation (2/3:1/3) indicates that two portions are used as training and the remaining

one as testing. Figure 3.18 shows the experiments by cross-validation using H_a=5% and H_c=5%. There are significant differences between our thematic term approach summaries and the baseline experiments. Contrarily, Figure 3.19 (with H_a=45% and H_c=45%) shows the results are almost the same as the baseline experiments. This evidence further confirms that only small portions of thematic terms should be used in generating meaningful or presentable summaries.

Experiments obtained by combining different portions (1/3:2/3 and 2/3:1/3) in training and testing under H_a=5% and H_c=5%, and H_a=45% and H_c=45% are summarized in Figure 3.20. With the same portions used in training and testing, AIR results show that the curve with H_a=5% and H_c=5% obtains better results than the curve with H_a=45% and H_c=45%. Furthermore, using more portions (2/3) in training obtained higher AIR results. This phenomenon shows that more training portions can tune a good value for β for summary generation. This implies that more presentable summaries can be generated. Also, the curve of H_a=5% and H_c=5% lies flat at 20% of compression rate. Further increase in the compression rates will not affect the performance anymore. About 20% of the full-length document was sufficient to form a good surrogate for the original document.

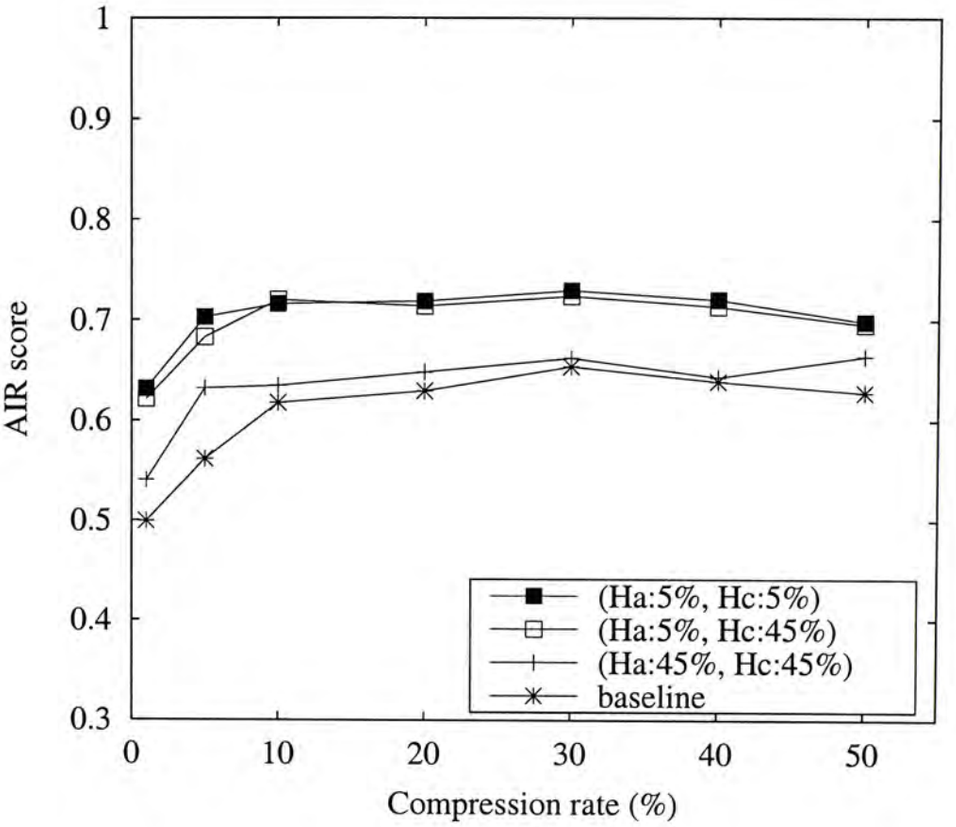


Figure 3.17: *AIR results for English corpus with 50% training to 50% testing.*

As shown in Figure 3.21, our (Chinese) summaries also achieved higher AIR scores than that of the baseline. Moreover, curve lies flat at 10% showing that about 10% of the full-length document was sufficient to be a good surrogate for the original document.

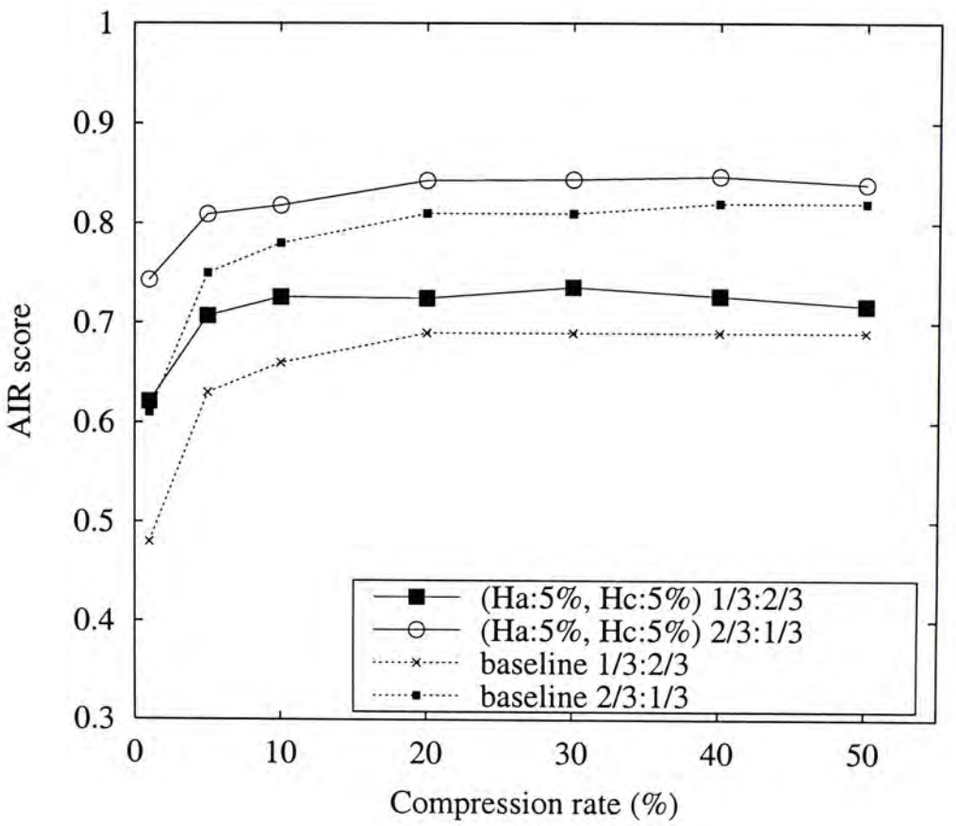


Figure 3.18: *AIR results for English corpus with H_a of 5% and H_c of 5% after conducting three-fold cross-validation.*

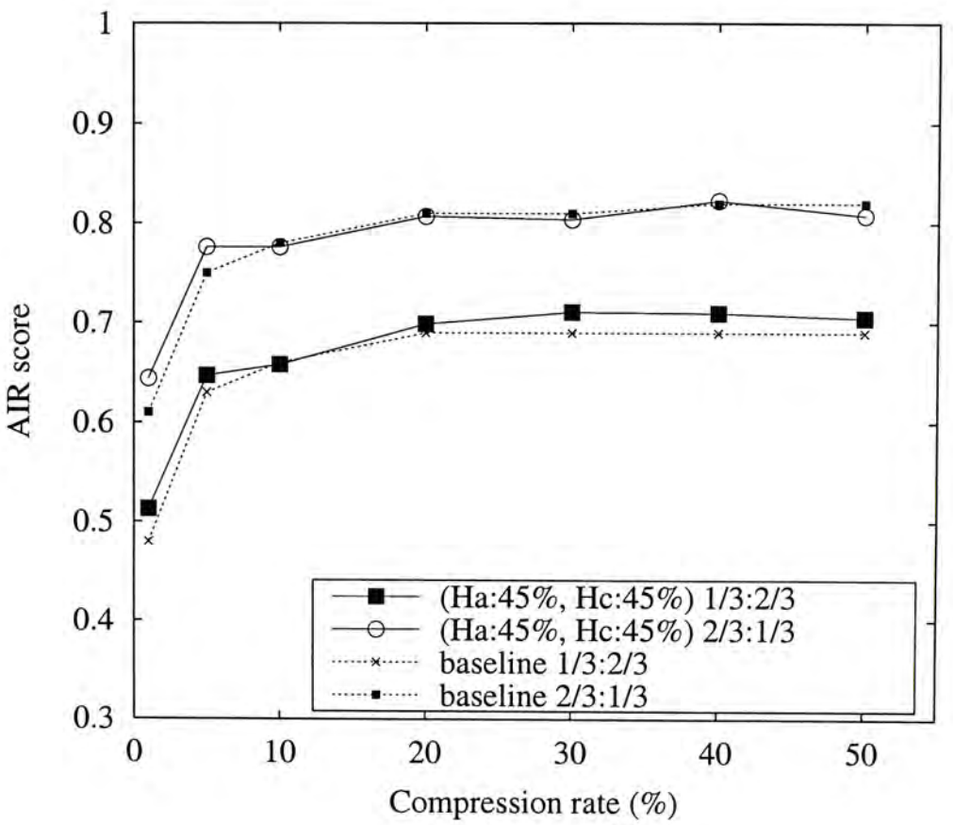


Figure 3.19: *AIR results for English corpus with H_a of 45% and H_c of 45% after conducting three-fold cross-validation.*

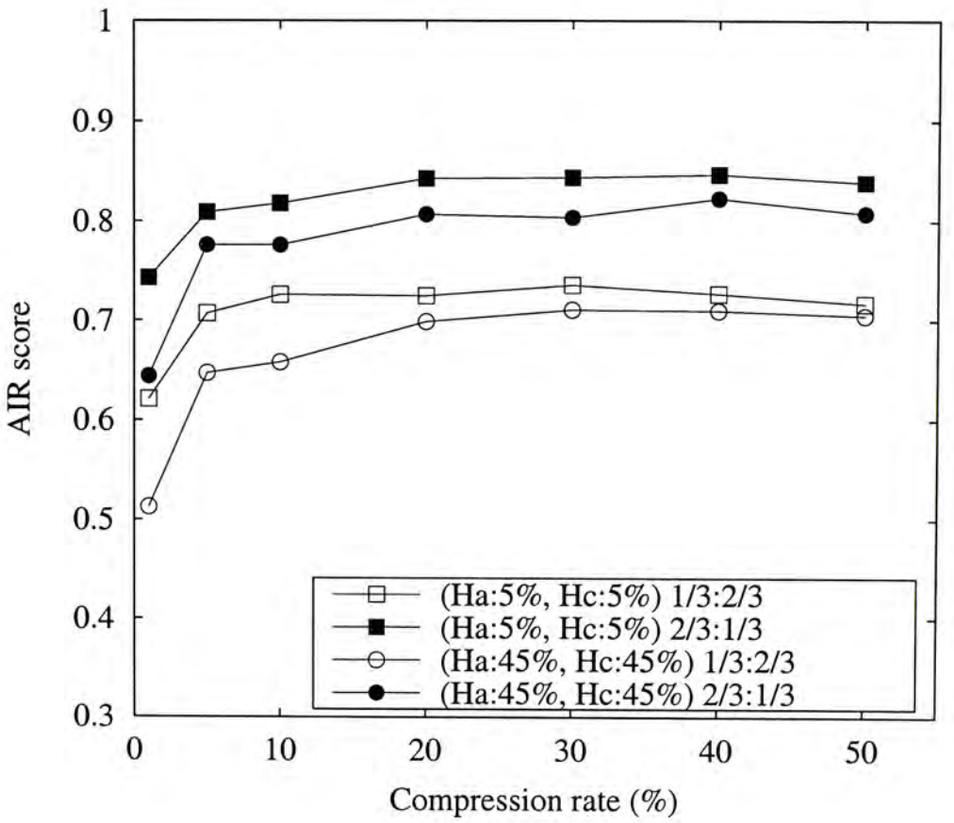


Figure 3.20: *Overall AIR results for English corpus in three-fold cross-validation experiments.*

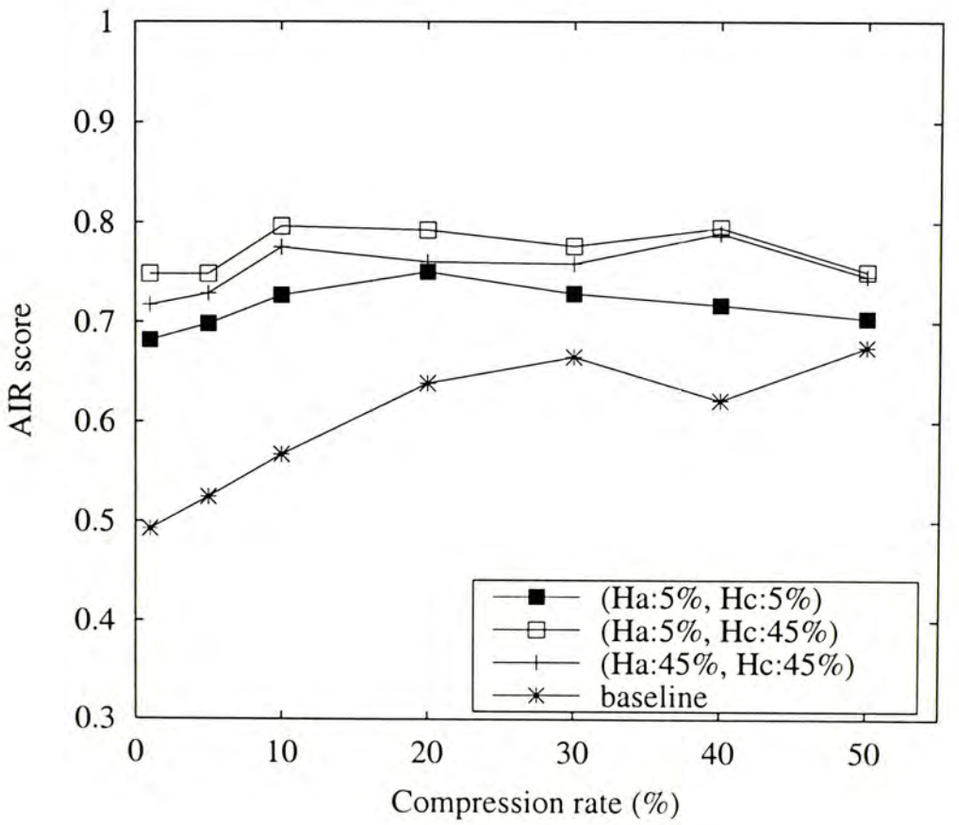


Figure 3.21: *AIR results for Chinese corpus with 50% training to 50% testing*

.

3.5 Comparison between the content-based similarity measure and the average inverse rank method

As described before, the content-based similarity measure cannot be used when there are no handwritten summaries provided. Also, if the vocabularies used in handwritten summaries are different from the passage, terms overlapping will be small resulting in a low similarity value. This phenomenon is illustrated by following examples. These examples were evaluated by the content-based similarity measure and the average inverse rank (AIR) method. Their results are shown in Table 3.6. We generated extracts by considering the compression rate of 5%, with H_a (threshold for article-based thematic terms) set at 5% and H_c (threshold for corpus-based thematic terms) set at 5%.

Article id	Similarity value	Rank
9408004	0.022	1
9504033	0.199	1

Table 3.6: *Results of articles (9408004.xml and 9504033.xml) evaluated by AIR evaluation method and content-based similarity method were shown.*

As shown in Table 3.7 and Table 3.8, writers simply used phrases to integrate the idea of the passages. Those phrases include *"some"*, *"probabilistic*

This paper is an attempt to bring together two approaches to language analysis. The possible use of *probabilistic information* in principle-based grammars and parsers is considered, including discussion on *some* theoretical and computational problems that arise. Finally a partial implementation of these ideas is presented, along with *some* preliminary results from testing on a small set of sentences.

Table 3.7: *Handwritten abstract of 9408004.xml.*

information", *"improve the accuracy"*, *"more accurate"*, etc. However, our summaries provided more meaningful content, indicated by the italicized sentences in Table 3.9 and Table 3.10, when compared to the phrases in the handwritten abstracts.

For instance, *accurate* in a handwritten abstract of 9504033.xml (see Table 3.8) referred to the writer's model which generated more accurate result than other. Our summary (see Table 3.10) stated more clearly about the figure of the accuracy, i.e. 81%. When our summaries were evaluated by the average inverse rank (AIR) evaluation method, the AIR=1 (see Table 3.6). These outcomes imply that extracts could represent their original effectively.

Corpus Statistics Meet the Noun Compound: Some Empirical Results A variety of statistical methods for noun compound analysis are implemented and compared. The results support two main conclusions. First, the use of conceptual association not only enables a broad coverage, but also *improves the accuracy.* Second, an analysis model based on dependency grammar is substantially *more accurate* than one based on deepest constituents, even though the latter is more prevalent in the literature.

Table 3.8: *Handwritten abstract of 9504033.xml.*

While the former borrows from advanced linguistic specifications of syntax, the latter has been more concerned with extracting distributional regularities from language to aid the implementation of NLP systems and the analysis of corpora. Using the schemata in this way suggests that the building of structure is category independent, i.e. it is just as likely that a verb will have a (filled) specifier position as it is for a noun. In order to force the choice of the 'best' parse on to the verb, the probabilities of theta grids for nouns, prepositions, etc. was kept constant. *The grammar employed is a partial characterization of Chomsky's Government-Binding theory , and only takes account of very local constraints (i.e. X-bar, Theta and Case); a way of encoding all constraints in the proper branch formalism will be needed before a grammar of sufficient coverage to be useful in corpora analysis can be formulated.* *It would be an elegant result if a construction such as the passive were to use probabilities for chains, Case assignment etc. to select a parse that reflected the lexical changes that had been undergone, e.g. the greater likelihood of an NP featuring in the verb's theta grid.*

Table 3.9: *Summary of 9408004.xml generated by using content-based similarity measure.*

The dependency model attempts to choose a parse which makes the resulting relationships as acceptable as possible.
The dependency model has also been proposed by Kobayasi et al (1994) for analysing Japanese noun compounds, apparently independently.
A simple calculation shows that using their own preprocessing heuristics to guess a bracketing provides a higher accuracy on their test set than their statistical model does.
A test set of syntactically ambiguous noun compounds was extracted from our 8 million word Grolier's encyclopedia corpus in the following way.
Eight different training schemes have been used to estimate the parameters and each set of estimates used to analyse the test set under both the adjacency and the dependency model.
To determine the difference made by conceptual association, the pattern training scheme has been retrained using lexical counts for both the dependency and adjacency model, but only for the words in the test set.
Left-branching is favored by a factor of two as described in the previous section, but no estimates for the category probabilities are used (these being meaningless for the lexical association method).
Three training schemes have been used and the tuned analysis procedures applied to the test set.
However, for the pattern training scheme an improvement was made to the dependency model, producing the highest overall accuracy of 81%.
The model also has the further commendation that it predicts correctly the observed proportion of left-branching compounds found in two independently extracted test sets.

Table 3.10: *Summary of 9504033.xml generated by using content-based similarity measure.*

3.6 Chapter summary

In this chapter, we propose two evaluation methods for accessing our thematic term summarization approach. The first method is the traditional content-based similarity metric. The main limitation of content-based similarity metric is that it requires the availability of standard summaries. In view of this, we propose a new method to evaluate the quality of summary by assessing the representative ability of a summary in acting as a surrogate within an entire data collection based on an information retrieval model. This method is called Average Inverse Rank (AIR) method. Experimental results on both Chinese and English summaries show that our extracts are effective under AIR evaluation metric.

Chapter 4

Bilingual Event-Driven News Summarization

In this chapter, we describe a framework for bilingual event-driven news summarization. Dictionary-based term translation is applied to handle a bilingual corpora. Unsupervised learning is used to discover new events. Specifically, incremental K-means clustering is used for processing sentences from a given topic, with the aim of capturing a coherent event in every cluster. We design two selection criteria to rank the clusters: *intra-cluster consistency* and *cluster-topic relevance*. Intra-cluster consistency intends to locate highly cohesive clusters, in which similar instances are grouped appropriately. Cluster-topic relevance aims at selecting clusters that are highly related to elements mentioned in the news topic. Our summarization approach attempts

to select the "best" clusters (or events) and use their corresponding sentences to form a coherent summary.

4.1 Corpora

We used news stories provided by Topic Detection and Tracking Project (TDT2) multi-lingual collection as our bilingual corpora. English news are collected daily from six news sources. They are Voice of America, Public Radio International — The World, ABC — World News Tonight, CNN Headlines News, Associated Press World Services, and New York Times Newswire Services. In addition, Chinese news stories, being extracted from three sources, are represented in GB code. Those three sources include Xinhua News Agency, Zaobao News Agency, and Voice of America-Mandarin Chinese news program. The time-span of the stories is between 1st April 1998 and 30th June 1998.

We selected the two topics in the multi-lingual collection, which contained a reasonable number of Chinese stories ranging from 50 to 100. In TDT2 corpus, there is a label on each story showing the degree of relevance towards the topic. The labels include YES, BRIEF, and REJECT. YES is assigned if at least 10% of the story is devoted to that topic. BRIEF is assigned if less than 10% of the story is devoted to that topic. REJECT is assigned if it is

incorrectly segmented, obviously not news, a continuation of a story, or an error in formatting. Our experiments focused on stories labeled with YES.

4.2 Topic and event definitions

We present the definition of topics and events. According to TDT [4] : "A topic is something that happens at some specific time and place, and the unavoidable consequences". Some examples of topics are "Upcoming Philippine election" and "German train derail". Generally, a news topic includes events such as facts and activities that are directly related to the topic. For example, in the topic "German train derail", some events may occur such as the time and place of the derail, the death reports, the rescue work after the derail happened, and the safety actions being carried out afterwards.

Our summarization technique attempts to extract sentences that are highly related to the events mentioned within a topic. We call these sentences "on-event" sentences under a particular topic. System generated summaries are assessed by measuring the number of relevant on-event sentences extracted. For evaluation, we invited two annotators to judge whether sentences were on-event or off-event beforehand. The entire event lists of the two corpora are shown in Appendix C.1 and Appendix C.2. Statistics of the

two bilingual corpora are shown in Table 4.1.

	20005		20091	
	Chinese	**English**	**Chinese**	**English**
Total no. of stories	47	41	14	54
Total no. of sentences	1341	760	383	796
Average no. of sentence per story	28.5	18.5	27.4	14.7
No. of off-event sentences	575	291	159	287
No. of on-event sentences	766	469	224	509
Percentage of on-event sentences	58.8%		62.2%	

Table 4.1: *Statistics of the two bilingual corpora, 20005 (Upcoming Philippine election) and 20091(German train derail).*

4.3 Architecture of bilingual event-driven news summarization system

Our summarization technique aims at generating coherent summaries by discovering events from bilingual news corpora. The corpora consist of Chinese and English news covering similar events and activities of a topic. To deal with bilingual news, our method conducts a term-by-term translation process transforming English terms into Chinese representations. As a result, unsupervised learning method can be used to discover events under a uniform representation of news regardless of the difference in language.

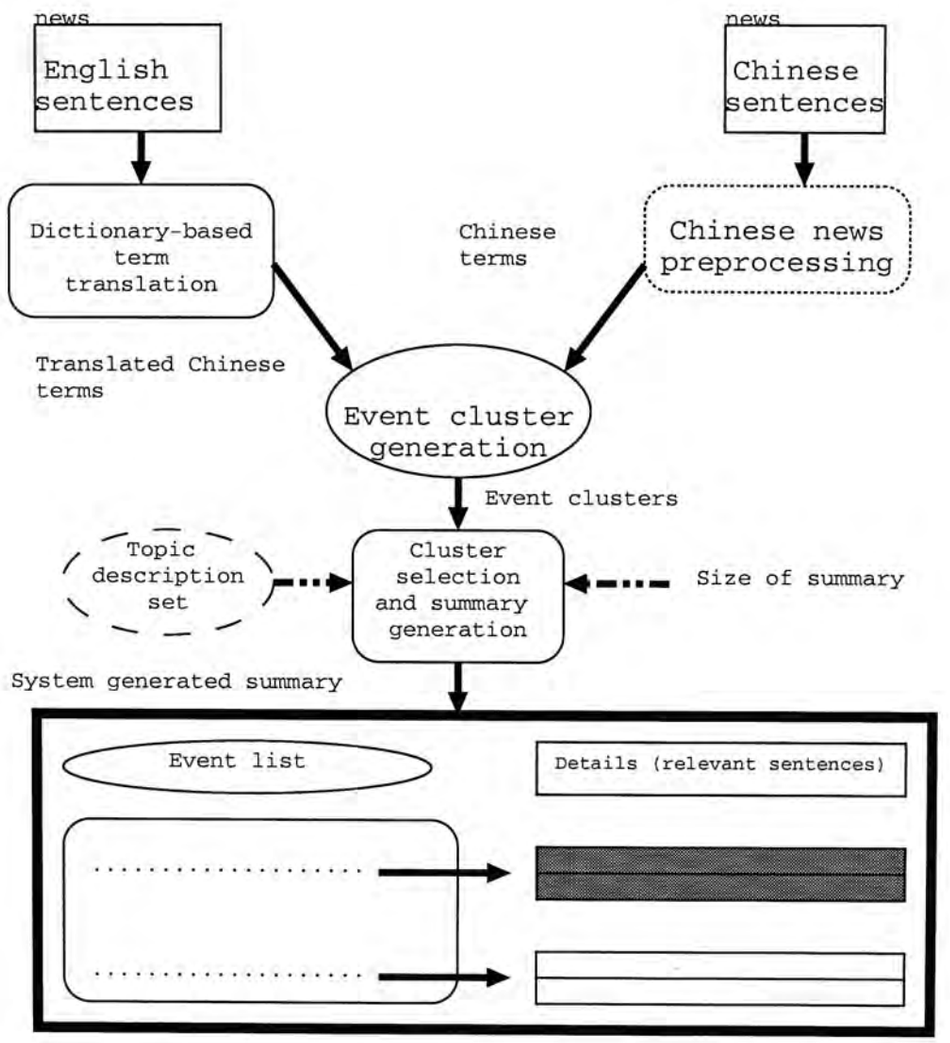


Figure 4.1: *The overall design of the summarization system. There are four core procedures contributing to the summarization task: (i) dictionary-based term translation, (ii) Chinese news preprocessing, (iii) event cluster generation, and (iv) cluster selection and summary generation. Finally, event-based summaries are generated.*

As shown in Figure 4.1, our bilingual summarization system consists of four modules. They are (i) dictionary-based term translation, (ii) Chinese news preprocessing, (iii) event cluster generation, and (iv) cluster selection and summary generation.

For English news sentences, we translate all terms into Chinese. We make use of a bilingual lexicon to look up the corresponding Chinese translations [8, 20, 24, 38]. To further disambiguate the translated terms, each Chinese term is associated a weight (see later), which indicates the degree of relevance of being an appropriate translation. For Chinese news sentences, we perform word segmentation to group Chinese characters into meaningful terms. At this stage, both English and Chinese news are transformed into a uniform representation, which is a set of Chinese terms. These terms are fed into the event cluster generation process.

The objective of the event cluster generation is to group related sentences that cover similar events into clusters. We use the *incremental K-means clustering* method to deal with the discovery of unseen events.

We design two selection criteria, namely *intra-cluster consistency* and *cluster-topic relevance*, to carry out the cluster selection process. In this process, we intend to select good clusters to generate the summary. Making use of these metrics, highly cohesive and topic relevant clusters are then selected to construct our summaries.

4.4 Bilingual event-driven approach summarization

We first describe the dictionary-based term translation for handling English news. We then present the preprocessing task for handling Chinese news. Next, we introduce the event discovery based on the *incremental K-means clustering* algorithm. Finally, we present our method for summary generation by introducing two selection criteria for choosing highly representative clusters to be included in the summary.

4.4.1 Dictionary-based term translation applying on English news articles

Dictionary-based term translation aims to translate all English sentences into Chinese terms for the subsequent step of the event discovery process. Before carrying out term translation, we apply lemmatization on each English term by making use of WordNet [31]. The lemmatization process reduces various word forms into a common term. Moreover, morphological information such as tense is handled. In WordNet, given an English term or a phrase as a query, it returns the lexicalized form under each syntactic category, namely, noun, verb, adjective, and adverb. We take the first available lexicalized term

as the output. For example, if "leaves" is given, the first lexicalized form, i.e., "leaf" in the noun category will be obtained. If "leave" is checked, the first lexicalized form will be "leave" in the noun category as well.

The next step is to make use of an English-Chinese bilingual lexicon. This bilingual lexicon is derived from a Chinese-English bilingual lexicon v2.0 provided by LDC (Linguistics Data Consortium)[3]. There are 187,439 bilingual entries and the Chinese terms are coded in *GB*. We transformed it as English-Chinese lexicon and its structure looks like a dictionary with an English term or phrase mapping to one Chinese translation. Table 4.2 shows some entries in the bilingual lexicon. Typically, given an English term, it is likely that more than one translated Chinese candidates exist in the lexicon. Not all translations found in the bilingual lexicon are appropriate for a given context. Also, English terms may be written as a phrase such as "growth rate" (增长率). It is not effective to solely consider a single English word in searching for its Chinese translations. In view of these issues, we develop our term translation approach in two steps:

(i) Phrase translation method

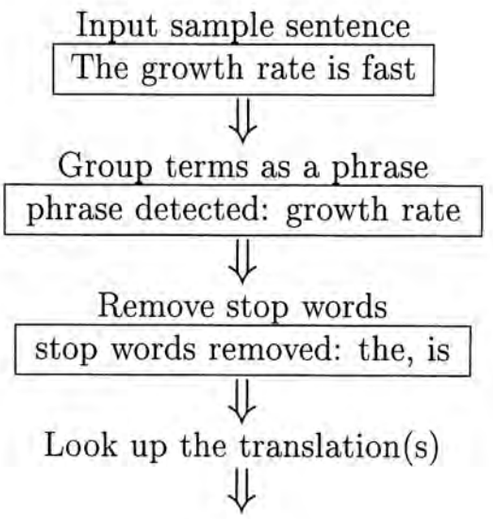
(ii) Translation term disambiguation method

English term / phrase	Chinese translation	English term / phrase	Chinese translation
right	不错	in turn	交互
right	当	in view of	基于
right	对	growth rate	增长率
right	对啊	right away	马上
should	当		
I agree	对啊		
yes	对啊		

Table 4.2: *This table shows some entries in the bilingual lexicon. It contains one-to-many and many-to-one mappings. English phrases are also provided in those mappings.*

(i) Phrase translation method

Phrase translation is applied to an English sentence in an attempt to group consecutive words into meaningful phrases. After phrases are detected, we can look up the bilingual lexicon to get the Chinese translations. Typically, our bilingual lexicon contains multiple entries corresponding to an English phrase. Given an English sentence, we first consider n consecutive English terms. In our system, we take n to be 6. When the current n fails to find its Chinese translation, we reduce this number by one until some Chinese translation(s) is (are) found. We demonstrate the phrase translation method on a sample sentence, "The growth rate is fast", in Figure 4.2. The flow chart of the phrase translation method is shown in Figure 4.3.



Result	
growth rate	⇒ 增长率
fast	⇒ 禁食 or 块 or 快 or 快速 or 牢 or 速 or 讯 or 迅速 or 遽 or 骛

Figure 4.2: *A sample sentence ("The growth rate is fast") for the phrase translation method.*

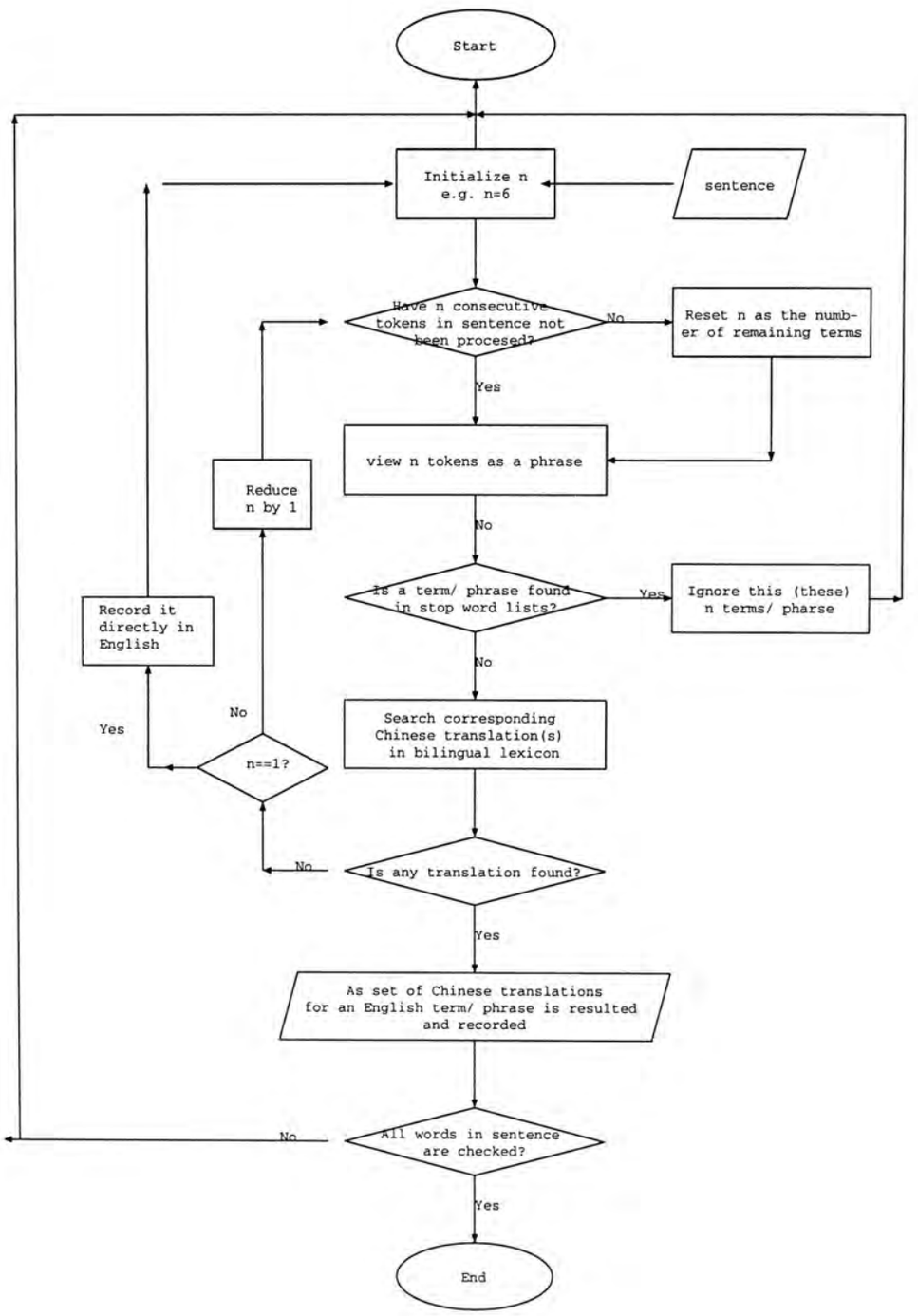


Figure 4.3: *The flow chart of the phrase translation method.*

(ii) Translation term disambiguation method

After the phrase translation, each English term / phrase is used as a key to look up the entries in the bilingual lexicon. Typically, there is a set of Chinese terms associated with an English term / phrase. Some translated Chinese terms are inappropriate for the context. Therefore, we develop a method to calculate the weight of each translated Chinese candidate found from the lexicon. This weight indicates the degree of relevance of the translation.

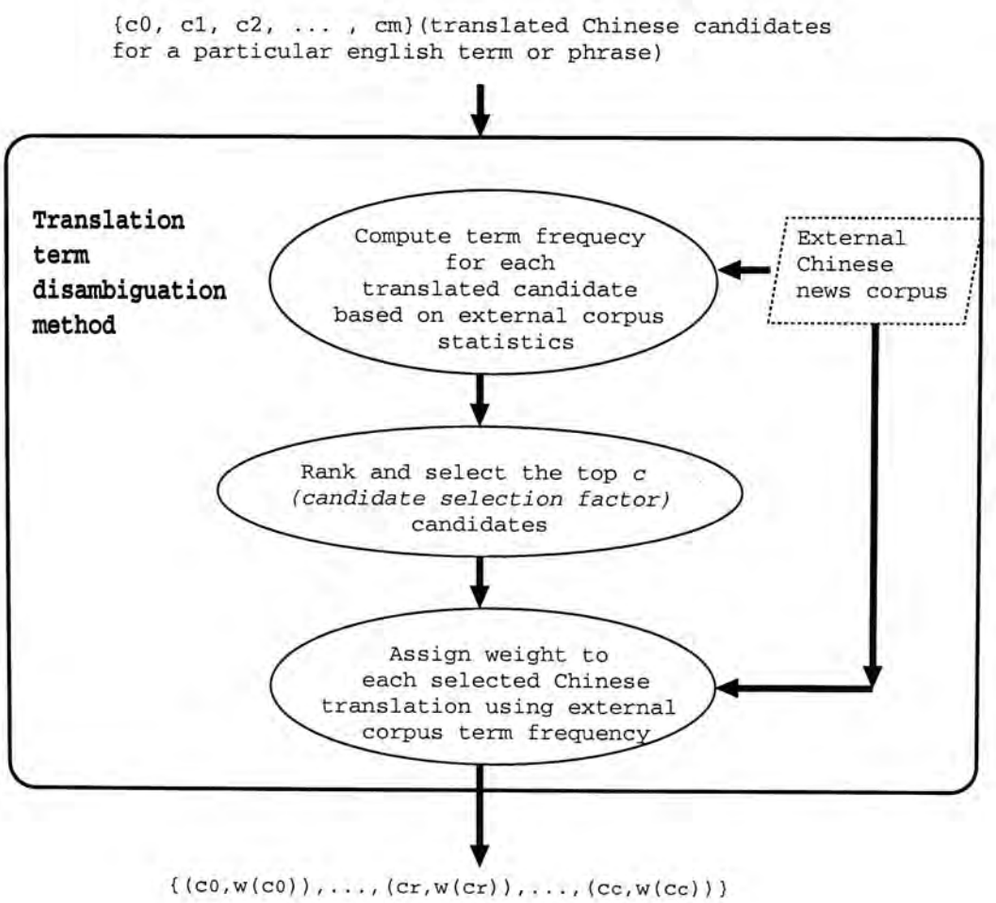


Figure 4.4: *The translation term disambiguation method.*

Our method relies on external corpus statistics to determine if translated Chinese candidates are appropriate translations. The assumption is those terms that occur more frequent in a corpus are more likely to be the correct translations. Here, we make use of an external Chinese corpus containing news articles from Wen Wei Po to derive the external corpus statistics. This external corpus consists of more than 6,100 news articles of local and world news. There are in total 3,503,631 segmented tokens and 28,171 distinct tokens.

For each English term / phrase, there is a set of translated Chinese candidates. For each candidate, we calculate the term frequency defined as the number of occurrence of it appeared in the external corpus. Afterwards, we rank all the translated Chinese candidates according to the term frequencies and pick the top c terms to be the translations for this English term / phrase. We refer c as the *candidate selection factor*. However, if a particular translated candidate is not found in the external training corpus, then there is no term frequency obtained. To deal with this situation, we assign a very small value to the term frequency for those Chinese terms that are absent in the external training corpus. This value is set to 0.0001 in our experiments. As a result, a number of c translated Chinese candidates are selected as the translation output of the corresponding English term / phrase.

Among all selected Chinese translations, we conduct a normalization for

each candidate. Suppose a weight of the selected Chinese term, c_r, is calculated by considering term frequency, denoted as $tf(c_r)$, found in the external corpus. The normalization formula is as follows:

$$w(c_r) = \frac{tf(c_r)}{\sum_{j=0}^{c} tf(c_j)} \tag{4.1}$$

where $w(c_r)$ represents the weight calculated for a Chinese translation c_r in a sentence. Let e be an English term / phrase. After the translation term disambiguation method, it is translated into the following representation:

$$\{(c_0, w(c_0)), \ldots, (c_r, w(c_r)), \ldots, (c_c, w(c_c))\}$$

For example, consider the English term "$fast$", there are ten translated Chinese candidates. Suppose c is set to 5. The term frequencies of each of the ten Chinese candidates are computed. The top five candidates are then selected. An illustrative example of the translation term disambiguation method is shown in Figure 4.5.

English term / phrase

fast

⇓

Obtain ten translated Chinese candidates

⇓

Rank in descending order based on term frequency of external corpus

⇓

Chinese term	term frequency
快	636.0000
迅速	492.0000
块	195.0000
速	192.0000
快速	189.0000
牢	16.0000
骂	1.0000
遽	1.0000
讯	0.0001
禁食	0.0001

⇓

Compute the final weight (see Eqn 4.1) of the selected top c (i.e. 5) Chinese translations

⇓

Chinese term	term frequency	weight
快	636.0000	0.3732
迅速	492.0000	0.2887
块	195.0000	0.1144
速	192.0000	0.1127
快速	189.0000	0.1109

Figure 4.5: *The output of a sample English term ("fast") using the translation term disambiguation method.*

4.4.2 Preprocessing for Chinese news articles

We discuss briefly the preprocessing work for Chinese news articles. We apply word segmentation on each Chinese sentence. As a result, Chinese characters are segmented into words which are more meaningful than discrete single characters. The word segmentation module, obtained from Linguistic Data Consortium (LDC), makes use of dynamic programming to find the path which has the highest multiple of word probability. In general, this segmentation works well for most Chinese texts except for some unseen proper names. After segmentation is completed, we remove the stop words.

For each Chinese segmented token, we also need to assign a weight to compose a weight vector for subsequent processing. Since it is an original Chinese sentence, no translation is needed. We simply assign a weight of 1.0 for each segmented Chinese term in a sentence.

4.4.3 Event clusters generation

We employ an *incremental K-means clustering* algorithm [46] to form coherent event clusters among Chinese sentences. News actually comes in daily and can be accumulated for the clustering process. Obviously, more events arise as more news arrive. Therefore, the number of events will increase as the time goes by. The traditional K-means algorithm is inadequate to

cope with new events that raise dynamically. In traditional K-means algorithm, the number of clusters represented by K is fixed which implies that a constant number of clusters (K) will be obtained. For instance, if we apply traditional K-means clustering method and K is set to 3, all related sentences mentioning similar events will be gathered into three clusters. When more new events occur, it is inappropriate to cluster the new events into the three existing clusters. To cope with this problem, we employ the *incremental K-means clustering* method.

An effective incremental K-means clustering method should be able to gather all sentences related to similar events to form coherent event clusters. As mentioned before, it is required that a new cluster should be produced when a new event is detected. Therefore, the value of K, representing the number of cluster, changes dynamically as new events are detected. Thus, the initial value of K can be simply be set to 1.

In order to perform incremental K-means clustering, three important factors should be considered:

(i) representation of sentence and cluster means

(ii) the selection of clustering criterion

(iii) the threshold value, θ, used in judging if a sentence can be merged into a cluster

We describe these factors in the following paragraphs.

(i) The representation of sentence and cluster means

Each sentence is represented by a vector with a weight for each term. This weight is computed by the term translation method. We define S_i to be the weight vector of a sentence i:

$$S_i = \{w(c_0), \ldots, w(c_r), \ldots, w(c_m)\}$$

If sentence i is a translated Chinese sentence, $w(c_r)$ refers to the weight of a term c_r calculated by the term translation method described in Section 4.4.1. If sentence i is an original Chinese sentence, $w(c_r)$ is assigned to 1.0 directly. m denotes the total number of distinct terms within a sentence. In a cluster j, the mean vector, M_j, is represented as follows:

$$M_j = \{W_j(c_0), \ldots, W_j(c_q), \ldots, W_j(c_n)\}$$

$$\text{and} \quad W_j(c_q) = \sum_{\forall S_i \in \{Cluster j\}} w_i(c_q)$$

where M_j represents the mean vector of the cluster j. $W_j(c_q)$ is the summation of all weights of the Chinese token, c_q, found in all sentences within a cluster j. n denotes the total number of unique terms found in the cluster j. Generally, n is larger than m since terms within a sentence is a subset of terms found in a cluster. Both a sentence vector (S_i) or a cluster mean (M_j) are normalized by the maximum value found in the corresponding vector.

(ii) The selection of clustering criterion

We choose the cosine similarity measure to be the clustering criterion. To compute a similarity score, we take the dot product between a particular sentence vector (S_i) and the cluster mean vector (M_j). If a sentence i is highly similar to a cluster j, the similarity score tends to be 1. The similarity score, $\Delta(S_i, M_j)$ is introduced as follows:

$$\Delta(S_i, M_j) = \frac{\sum_r (w_i(c_r) \cdot W_j(c_r))}{\sqrt{\sum_r (w_i(c_r))^2 \cdot \sum_r (W_j(c_r))^2}} \tag{4.2}$$

where S_i denotes the vector of sentence i and M_j represents the mean vector of a cluster j. $w_i(c_r)$ refers to the weight of a Chinese token c_r in a sentence i. Moreover, $W_j(c_r)$ is the summation of all weights of Chinese token, c_r, appeared in the cluster j.

(iii) The threshold value, θ, used in judging if a sentence can be merged into a cluster

In order to facilitate the dynamic increase in the number of clusters, we need to define a threshold value to determine the creation of a new cluster. Here, we define θ as the threshold value to judge if a new event cluster should be created or not. For each sentence i, we compute the similarity score against all existing clusters using the cosine similarity measure. We then rank all

similarity scores in descending magnitude. The system compares the highest similarity score with θ. If the similarity score is larger than the threshold, θ, it implies that this sentence i possesses similar characteristics to merge into a cluster j. The hypothesis is that if a sentence i consists of certain amount of events that are also covered by a cluster j, then this sentence i should be clustered into the cluster j. However, if the highest similarity value fails to exceed θ, it implies that all existing clusters are not related to this sentence i. A new cluster is then formed by including this sentence as a member. This procedure is triggered when a new sentence covering events that do not exist before. We illustrate the creation of a new cluster in Figure 4.6 and we present the flow chart in implementing the incremental K-means clustering algorithm in Figure 4.7.

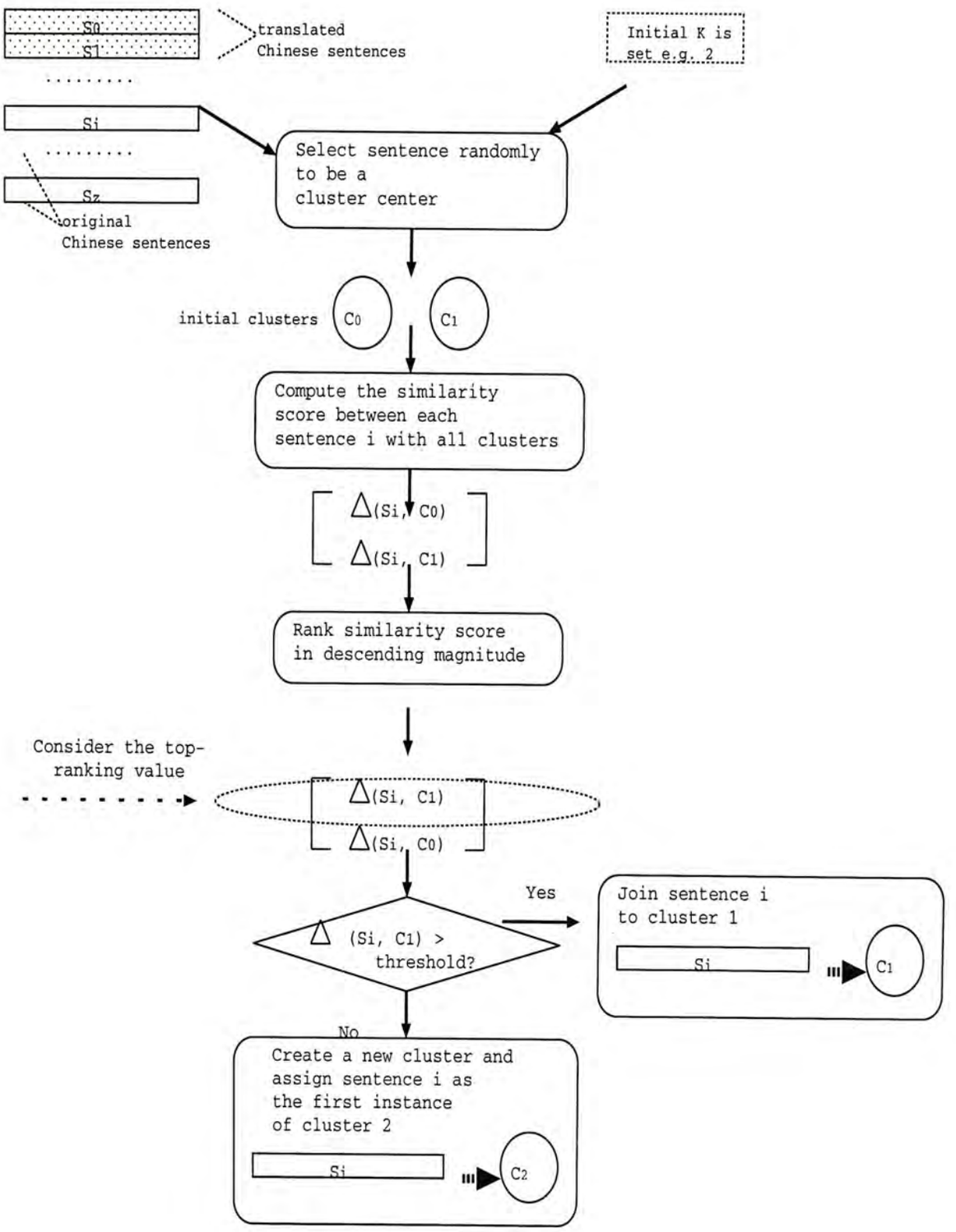


Figure 4.6: *Demonstration of an increase in the number of event clusters.*

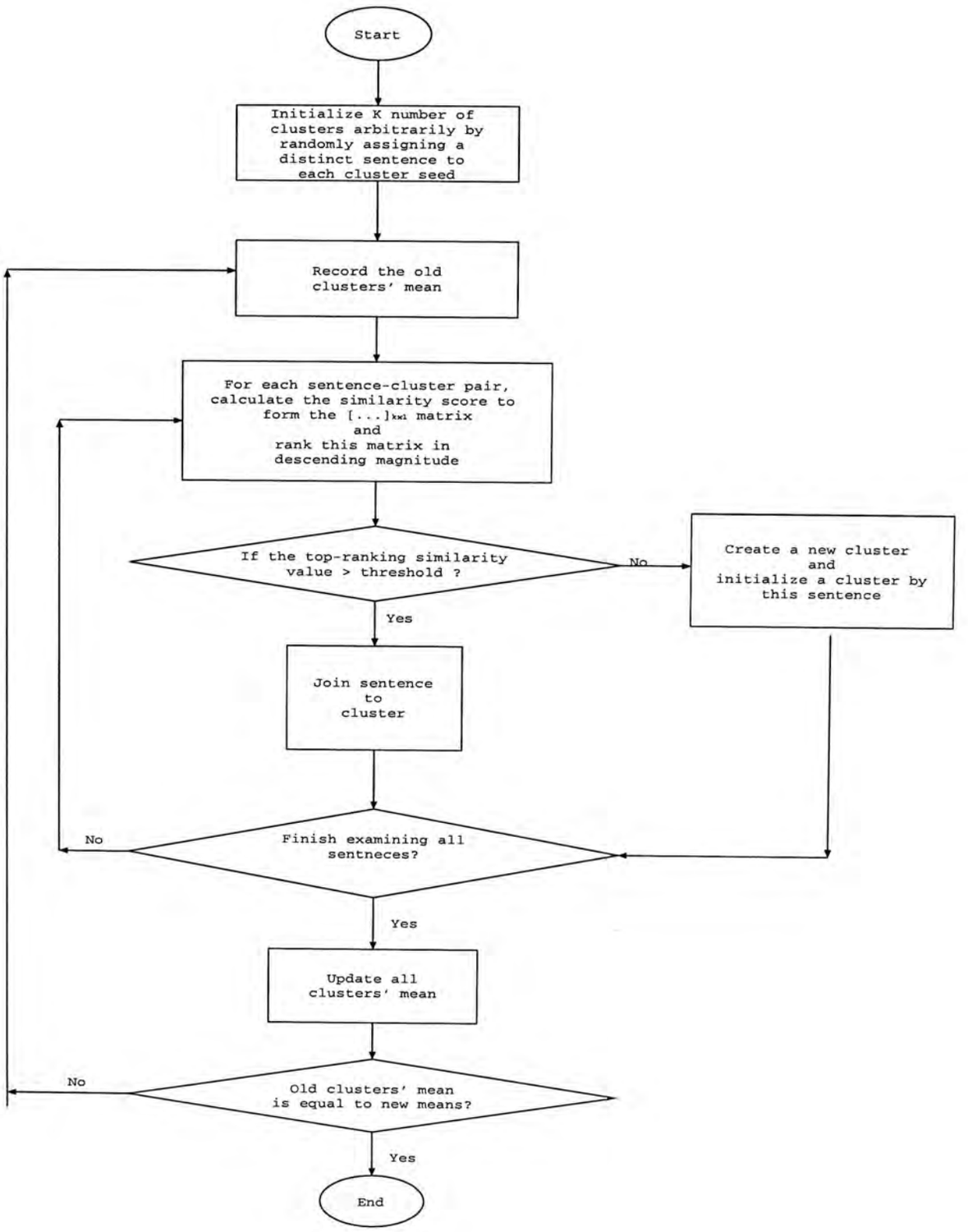


Figure 4.7: *Flow chart of the incremental K-means algorithm.*

4.4.4 Cluster selection and summary generation

When there is no change in the clusters' mean, coherent event clusters are obtained. We now consider how many representative event clusters should be selected to generate a summary. We design two selection criteria to achieve this task: (i) *intra-cluster consistency* and (ii) *cluster-topic relevance.* Intra-cluster consistency measures how sentences within a cluster are related to one another. The cluster-topic relevance reflects the degree of relevance for this cluster towards the topic. If both values are high, then this cluster is likely to be an on-event and highly cohesive cluster.

The overall score for each cluster is the combination of the effect of intra-cluster consistency and cluster-topic relevance. We define the overall score, O_j, of a cluster j, as the sum of the normalized intra-cluster consistency and the normalized cluster-topic relevance values.

(i) Intra-cluster consistency

We make use of the inverse of sum-of-square-error (SSE) to compute the intra-cluster consistency. This SSE is widely used as a criterion for clustering.

$$A_j = \frac{1}{SSE_j}$$

$$A_j = \frac{1}{\sum_i (M_j - S_i)^2} \tag{4.3}$$

where A_j denotes the intra-cluster consistency score computed for cluster j. SSE_j is the sum-of-square-error of cluster j. S_i refers to one of the sentence vector i within cluster j. M_j represents the mean vector of cluster j.

If instances of the cluster j are highly consistent with each other, the difference between instances and the cluster center is small implying that the sum-of-square-error will be small as well. Taking the inverse of this value produces the intra-cluster consistency value.

(ii) Cluster-topic relevance

To design the cluster-topic relevance score, we measure the content of a cluster covering the on-topic features. We take the English topic descriptions, provided by TDT2, to construct the on-topic features set $\{F\}$. The first step is to conduct the dictionary-based term translation on the English topic descriptions. The English sentences are converted into Chinese representations which are used to compose the on-topic features set $\{F\}$. Examples of $\{F\}$ are shown in Table 4.3 and Table 4.4.

The cluster-topic relevance score is defined as the probability of a cluster containing on-topic features set $\{F\}$ as follows:

$$B_j = \frac{|\{W\} \cap \{F\}|}{|\{W\}|} \tag{4.4}$$

where B_j refers to the cluster-topic relevance score of a cluster j. $\{W\}$ is the

term sets of a cluster j and $\{F\}$ is on-topic features set. The more on-topic features captured by a cluster, the higher is the relevance of the cluster in relation to the news topic.

The intra-cluster consistency score and the cluster-topic relevance score are first calculated for all clusters. We then normalize intra-cluster consistency score of each cluster by the maximum intra-cluster consistency score found across all clusters. Likewise, the on-topic relevance score of a cluster is normalized by the maximum on-topic relevance value obtained in all clusters. The overall cluster score is then computed by the sum of the normalized intra-cluster consistency and normalized cluster-topic relevance scores. We then rank the overall cluster score in descending order. Given a specific size of summary, the top-ranking clusters are used to compose the summary content sequentially. We can also build the event list by terms representing the mean of a cluster. By taking the sentences within the selected clusters, our summaries are generated (see Table 4.6).

Topic descriptions provided by TDT2

The replacement of 6 members of the Philippine Cabinet, most of whom resigned because they want to run in the upcoming election.

⇓

Dictionary-based term translation

phrase detected:	to run
stop words removed :	because, in, most, of, the, they, want and whom

⇓

On-topic features set {F} is obtained

replacement	⇒ 国民 / 全国性 / 国立
6	⇒ 6
member	⇒ 代替 / 替换 / 代替的人
philippine	⇒ 菲律宾的 / 菲律宾人的
cabinet	⇒ 橱 / 柜 / 小型柜橱 / 棱
resign	⇒ 辞职 / 辞
to run	⇒ 开 / 办 / 管 / 执行
upcoming	⇒ 即将来临的 / 预定将要
election	⇒ 选举 / 推选

Table 4.3: *This table shows the topic descriptions and on-topic features set* $\{F\}$ *for the topic "Upcoming Philippine election".*

Topic descriptions provided by TDT2

At least a hundred people were killed in Germany's worst post-war train crash. Prior to this investigations of possible reasons for the crash, reports from and about survivors, and actions taken regarding the safety of the other forty-three high-speed trains currently in service.

⇓

Dictionary-based term translation

phrase detected:	nil
stop words removed :	a, about, and, at, currently, for, from, least, of, possible, regarding, the, this, to, and were

⇓

On-topic features set {F} is obtained

hundred	⇒ 一百 / 佰
kill	⇒ 打死 / 戮 / 猃 / 戡
germany	⇒ 德 / 德国 / 德意志
worst	⇒ 最坏的 / 最差的
post-war	⇒ post-war
train	⇒ 训练 / 火车 / 列车
reason	⇒ 理故 / 故 / 理由 / 缘 / 端
crash	⇒ 坠毁 / 粉碎
investigation	⇒ 调查
survivor	⇒ 生还者 / 幸存者
action	⇒ 行动 / 作为 / 作用 / 行为 / 动作
safety	⇒ 安全 / 安全性 / 安危 / 安然
forty-three	⇒ forty-three
high-speed	⇒ 高速的 / 快速
service	⇒ 服务 / 功 / 役

Table 4.4: *This table shows the topic descriptions and on-topic features set {F} for the topic "German train derail".*

4.5 Evaluation for summarization based on event-driven approach

The generated summary intends for grasping relevant events and activities of a topic. Therefore, the evaluation method should focus on how many on-event sentences can be extracted.

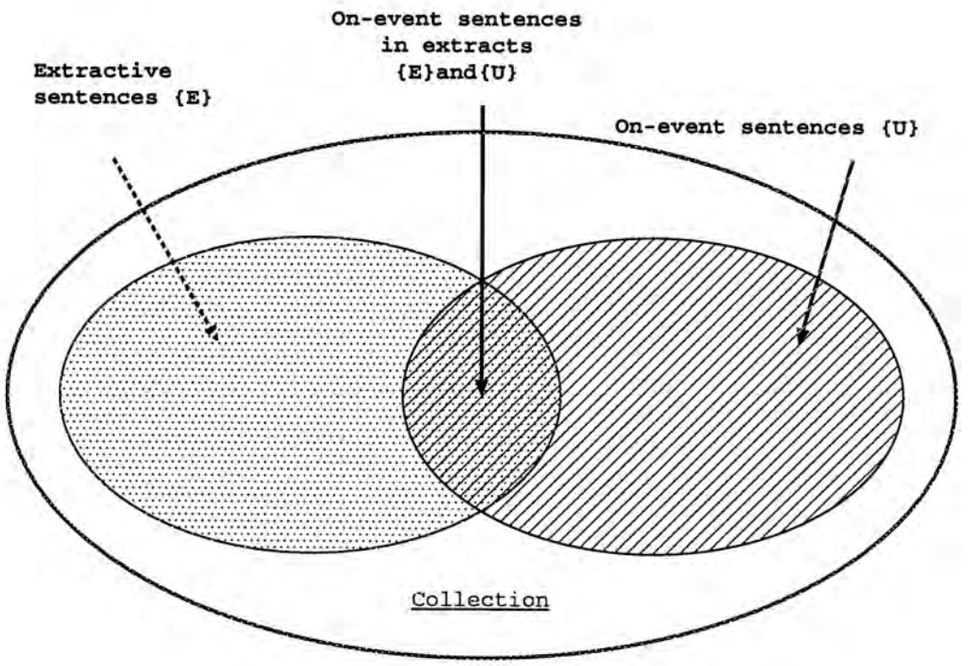


Figure 4.8: *This diagram shows the relationship between extractive sentences and on-event sentences.*

In Figure 4.8, $\{E\}$ is the set of sentences generated by selecting some good clusters based on two selection criteria in cluster selection and summary generation process. $\{U\}$ is the set of on-event sentences in the entire training / testing set. We define recall and precision measures as below:

$$R = \frac{\{E\} \cap \{U\}}{\{U\}} \tag{4.5}$$

$$P = \frac{\{E\} \cap \{U\}}{\{E\}} \tag{4.6}$$

where R is the recall score that measures how many of extracted sentences are marked as on-event over the set of on-event sentences in an entire training / testing set. P is the precision score that measures how many of extracted sentences that are correctly marked as on-event by our system.

If summarization system catches all sentences to be the extract, good recall is obtained. However, too many extracts will contain large amount of sentences that may not be on-event sentences. This case leads to a bad precision. On the other hand, summarization can attain a good precision by selecting only very few sentences. But it may lead to a low recall. We therefore consider f-measure [41] to be a single quality measure for our summaries. We use f-score which is derived from f-measure as an evaluation metric for our experiments. The formal definition of f-measure is:

$$\text{f-measure} = \frac{(\omega^2 + 1.0)P \cdot R}{(\omega^2 \cdot P) + R} \tag{4.7}$$

where ω is the relative weight of recall and precision.

In our experiments, we set ω to 1.0 producing f-score which is the harmonic mean of recall and precision value. If both recall and precision are

high, f-score becomes closer to 1.

$$\text{f-score} = \frac{2P \cdot R}{P + R} \tag{4.8}$$

4.6 Experimental results on event-driven summarization

The evaluation we focuses on how many on-event sentences are successfully extracted in the output summary. We employ the f-score as the evaluation metric.

4.6.1 Experimental settings

First, we divided the entire corpora into two distinct portions under each topic. These portions were used for training and testing purposes. Statistical details are shown in Table 4.5.

On the other hand, we conducted three-fold cross-validation experiments to evaluate our approach. Since our experiments were run on daily basis, we divided the data collection into three portions that were approximately equal in the number of days. Training was done on one portion and testing was performed on the other two portions. For each combination, we conducted

three trials for summary generation. As a result, we produced nine sets of experimental results. We plot the graph by taking the average results of these nine runs.

Corpora	**Training set**		**Testing set**	
Upcoming Philippine election	on-event	off-event	on-event	off-event
Total no. of sentences	631	455	604	411
Percentage among the entire collection	51.7%		42.9%	
German train derail	on-event	off-event	on-event	off-event
Total no. of sentences	374	164	362	249
Percentage among the entire collection	45.7%		54.6%	

Table 4.5: *Statistical details of the training set and the testing set for each topic are shown.*

The training set aimed at tuning a suitable threshold, θ, for the incremental K-means clustering method. The suitable θ was used in testing. In the experiments, θ was tuned from 0.001 to 0.300 in 0.001 intervals. If θ is too small, under-clustering resulted, i.e. the number of cluster produced is smaller than expected. As a result, many sentences are crowded in few clusters. On the contrary, if θ is large, many clusters (more than expected) are formed. The worst case is that one sentence belongs to one cluster only.

As mentioned before, experiments were conducted under different compression rates. Compression rate is defined as the ratio of the total sentence number in the summary to that in the entire collection. We examine vari-

ous compression rates at 10%, 20%, 30%, 40%, and 50% in the experiments. Larger compression rates were not attempted as we targeted to keep length as short as possible.

We also investigate a baseline method for generating summaries by randomly selecting sentences under various compression rates. Ten runs were conducted and the results were averaged to produce fairly baselines for comparison. Comparative study between our event-based method and a baseline method was conducted.

4.6.2 Results and analysis

Figures 4.9 to 4.12 show the results based on f-scores conducted with various compression rates. On the event-driven summarization plots in Figure 4.9 and Figure 4.11, there are (x, z) values. x denotes the precision score. z denotes the total number of clusters formed after the clustering process has completed. When the three-fold cross-validation procedure was used (see Figure 4.10 and Figure 4.12), the notation 1/3:2/3 denotes that news articles from one portion are used for training and the remaining two portions are used for testing. Similarly, 2/3:1/3 denotes that news from two portions are used for training and one portion is taken as testing.

As shown in Figure 4.9 to 4.12, summaries produced by the event-driven

approach always outperform the baseline method. It implies that the two selection criteria are effective in selecting coherent clusters to form the summaries. Highly topic relevant and consistent clusters are selected to compose the summaries.

Figure 4.9 and Figure 4.11 also show a precision value of 60% at 10% of compression rate. Encouragingly, Figure 4.9 depicts good results under each compression rate with high precision value. The last element in the bracket, z, indicates the final number of clusters generated. Figure 4.9 shows that the number of final clusters generated lies between 77 and 89 across different compression rates. This variation is produced by different initialized sentences and clustering threshold (θ) used in the clustering process.

We present a sample summary as shown in Table 4.6. Event items represent the clusters' mean are displayed under the topic shown on the top of the Table 4.6. An event list item can be viewed as a brief content capturing the major happenings of the topic, followed by the details describing the corresponding events. For instance, there are event list items showing the death report of the train derail accident. Our summaries show that the changing number of deaths was reported subsequently. The first two Chinese sentences stating the death number was around 100 persons. A few days later, an English sentence was selected, showing that the death number was confirmed to be 96 persons. Information related to the death report was updated on a

daily basis. Therefore, one language in the bilingual summaries can serve as complementary content for another one.

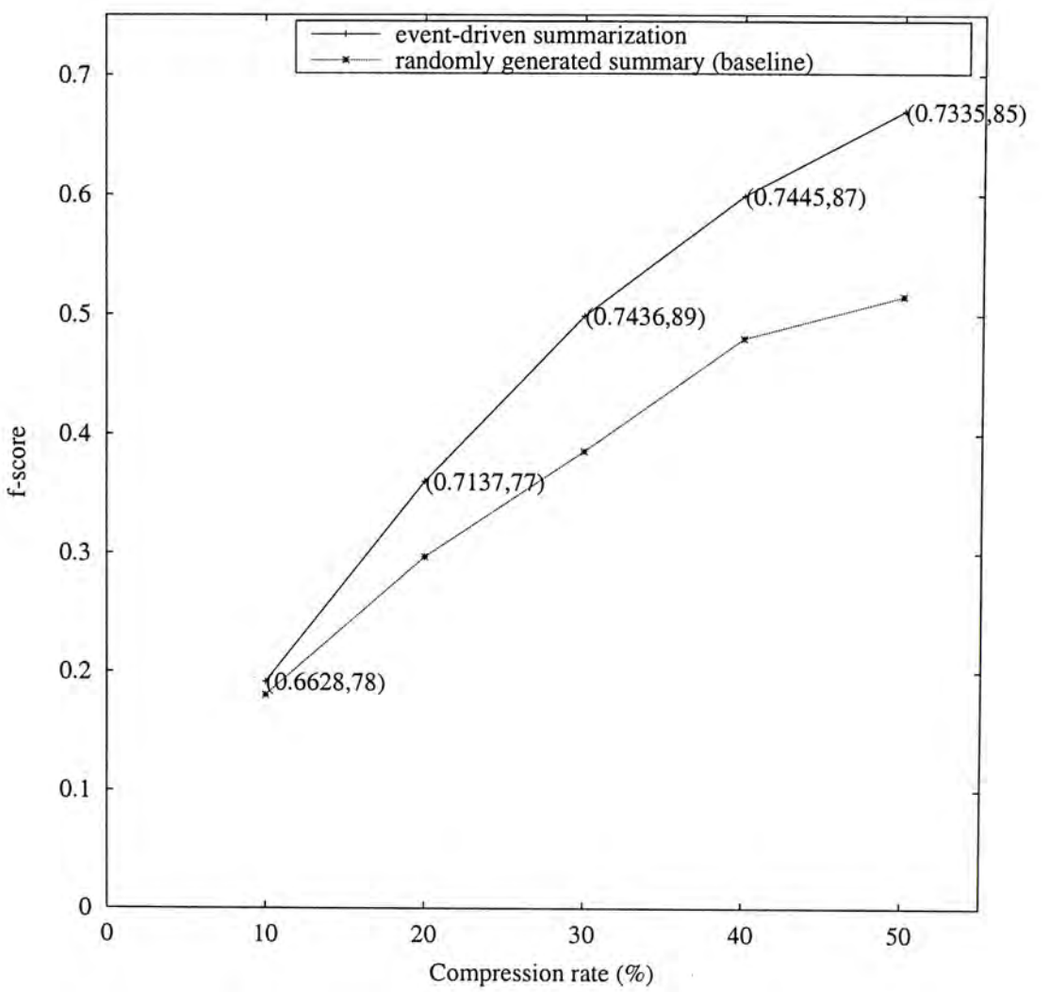


Figure 4.9: *This figure shows the change of f-score (average the results after five runs) with different compression rates. The precision scores are constantly above 66% (corpus: Upcoming Philippine election). The event-driven summarization method always outperforms the baseline. On the event-driven summarization, there are (x, z) values. x denotes the precision score and z denotes the total number of clusters found by the incremental K-means clustering method.*

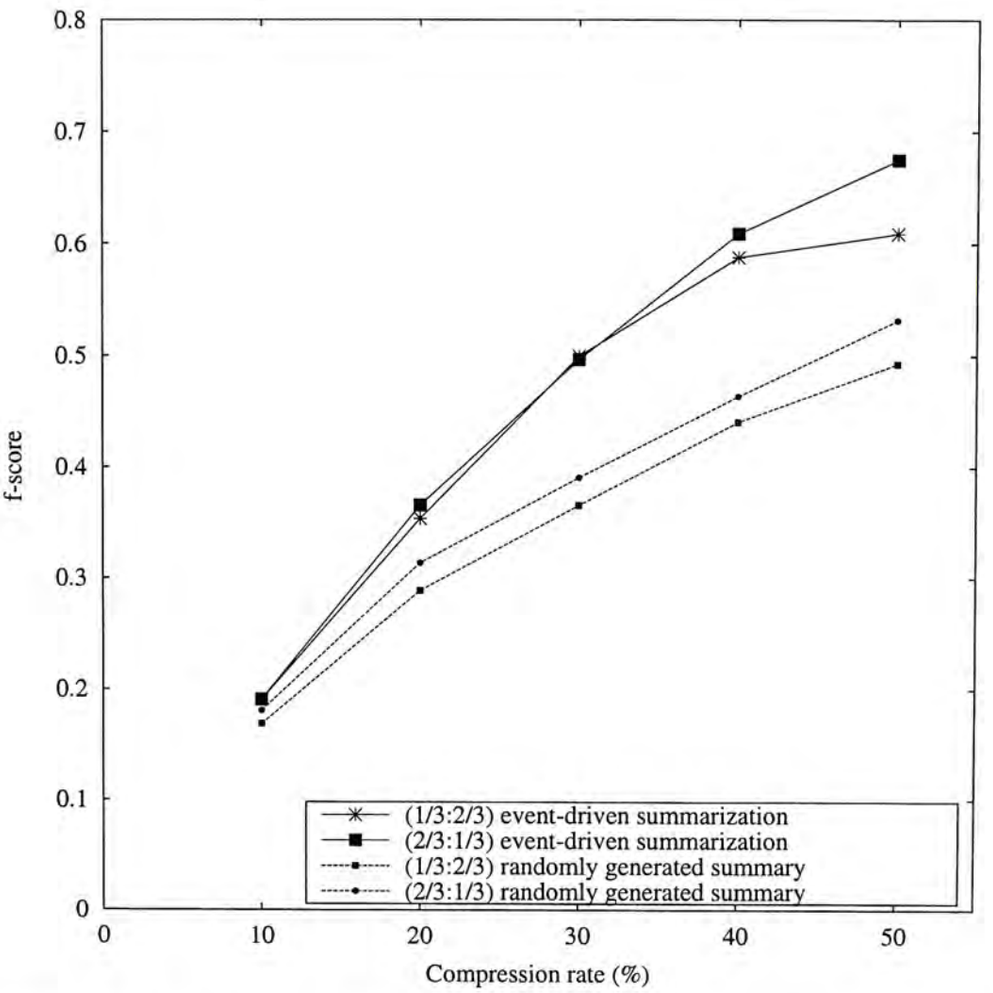


Figure 4.10: *This figure shows the change of f-score (average the results after nine runs) with various compression rates (corpus: Upcoming Philippine election) in three-fold cross evaluation.*

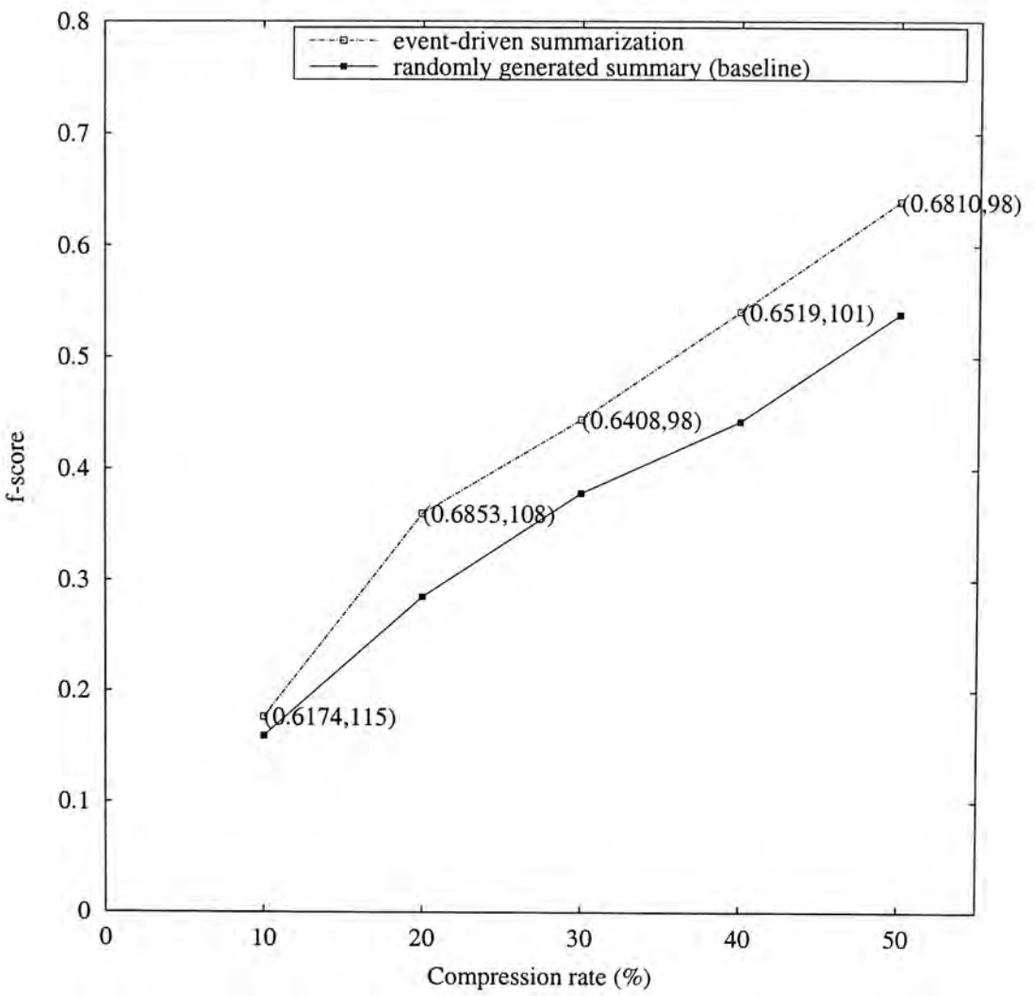


Figure 4.11: *This figure shows the change of f-score (average the results after five runs) with various compression rates (corpus: German train derail). About 61% of precision can still be obtained at 10% of compression rate. The event-driven summarization method always outperforms the baseline. On the event-driven summarization, there are (x, z) values. x denotes the precision score and z refers to the total number of clusters found by the incremental K-means clustering method.*

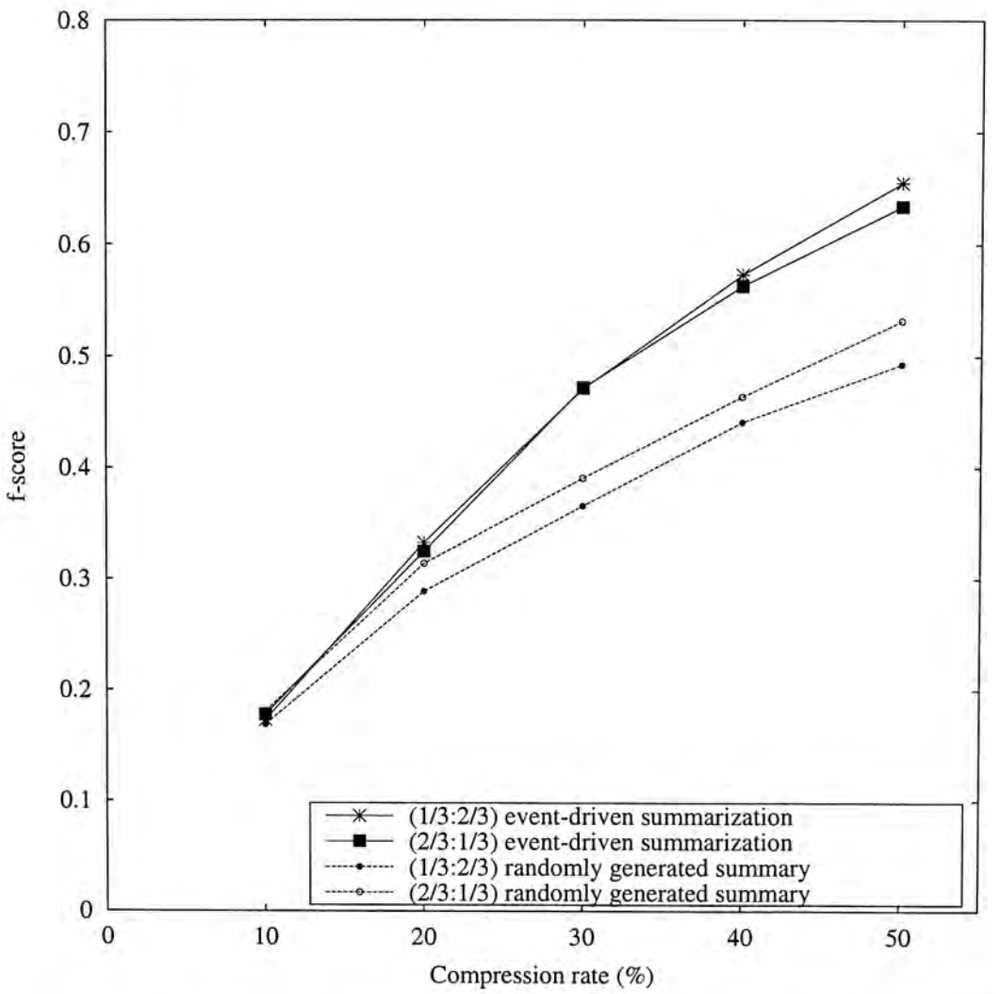


Figure 4.12: *This figure shows the change of f-score (average the results after nine runs) with various compression rates (corpus: German train derail) in three-fold cross evaluation.*

Topic: German train derail

event: time and place	event: death report	event: progress of the rescue work
慕尼黑 (Munich) 汉堡 (Hamburg) 星期三 (Wednesday) ↓ summary 由 慕尼黑 开 出 到 汉堡 汉堡 的 特快 列车 号 出 事故。 The high-speed train she boarded early Wednesday derailed and smashed in to the side of a bridge near the town of Celle , 140 kilometres (85 miles) south of Hamburg . In a sign that the German railway is taking no chances, an ICE train traveling from Munich to Hamburg Wednesday stopped in Celle about midday after personnel aboard reported loud noises.	丧 (death) 上升 (increase) 死亡人数 (death number) ↓ summary 使 死亡人数 上升 到 至少 一百 人 这 起事 故 造成 一百 零 二 人 死 。 With certainty there were 96 dead, and possibly the number will rise to 99%.	彻夜 (whole night) 尸体 (dead body) 残骸 (wreckage) ↓ summary 营救 人员 从 火车 残骸 里 拖 出 具 尸体。 It will include interviews of surviving passengers and rail personnel as well as maintenance workers. 救援 工作 人员 又 从 火车 的 残骸 里 找 到 了 几 具 尸体。 因此 拯救 人员 彻夜 使用 架 起重机 近 名 救护 、 工程 人员 来自 全国各地。
Explanations	*Explanations*	*Explanations*
These sentences stated that news topic was related to the accident of high-speed train. They pointed out the exact place (Celle) and They also gave information about when the accident happened. (Wednesday) gave more detailed information about the train coming from Munich to Hambury.	All three sentences came from different days of reported news. The changing number of deaths was reported repeatedly.	Ambulance man found dead bodies from the train wreckage. There were some passengers who survived. Ambulance man found dead bodies again, proving that rescue work was continuing. Ambulance man used machines in rescue work. Rescue work was done by many people who came from all around the world.

Table 4.6: *Example of a generated summary (event list and details) for the topic of "German train derail".*

4.7 Chapter summary

In this chapter, we present a novel summarization technique using event-driven approach to generate bilingual summary. This summarization method can deal with both Chinese and English news. We develop the dictionary-based term translation method via two steps. The first step is the phrase translation method. The second step is the translation term disambiguation method. We then apply incremental K-means clustering algorithm to generate coherent event clusters dynamically. Event clusters may be on-event or off-event. Therefore, we develop the cluster-topic relevance and intra-cluster consistency to select good clusters. Only highly topic relevant and consistent clusters are selected in summary generation. Our summary can always achieve above 60% precision at only 10% of the data. Also, our summary can obtain better f-score than the baseline method.

Chapter 5

Applying Event-Driven Summarization to a Parallel Corpus

This chapter presents the feasibility of our bilingual event-driven summarization technique for application to a parallel corpus. A set of experiments were conducted to demonstrate the effectiveness of our summarization algorithm. First, the characteristics of the parallel corpus used in the experiments are outlined. Based on the characteristics of the parallel corpus, two evaluation methods used in order to assess our summarization methodology are presented. Finally, the results and analysis are given.

5.1 Parallel corpus

The parallel corpus we used is a collection of English and Chinese documents. It consists of English news articles accompanied with corresponding Chinese articles. It is extracted from the web site of the press releases by the Hong Kong SAR government [2]. It is also available from our Area of Excellence (AoE) Web Repository [5]. Each press release is provided as a pair of Chinese and English texts. All texts have sentence-by-sentence correspondence in the parallel corpus. The following samples are of English on-event sentences and the corresponding on-event Chinese sentences.

English on-event sentence:
The Government has made use of portal technology to showcase e-Business development with the launch of the Central Cyber Government Office (CCGO) in August 2000 and the Electronic Services Delivery Scheme (ESD) in December 2000.

Corresponding Chinese on-event sentence:
政府 利用 入门 网 站 的 技术 ， 在 二 〇 〇 〇 年 八月 推出 「 数码 政府 合 署 」 ， 并 在 二 〇 〇 〇 年 十二 月 推出 「 公共 服务 电子 化 」 计划 ， 向 社会 各界 展示 电子 贸易 的 发展 。

The content is related to various subject matter such as fire accidents, traffic accidents, reports of legislative council, etc.

5.2 Parallel documents preparation

There are in total 8,518 news articles of Chinese and English news for a time-span between 1st January 2001 and 30th June 2001. We chose a topic among the articles to conduct experiments on event-driven summarization. After browsing through the collection, we chose the topic related to the "Electronic Service Delivery (ESD) scheme" which was a new scheme launched by the Hong Kong SAR government at the beginning of January in 2001. Under this scheme, on-line government services were widely provided to the citizens of Hong Kong. This topic covered a number of events that happened across the six months period. Examples of some events were about the promotion of the scheme, the provision of various kinds of services, and additional plans built for the scheme, etc.

We made use of an information retrieval engine to identify those documents related to the topic. We first constructed the query "ESD, ESDlife, electronic service delivery and digital 21". Here, "ESD" is the short form for electronic service delivery, which is a key initiative under the Digital 21 information technology strategy of the Hong Kong SAR government. ESDlife provides an on-line electronic platform for the ESD services. People in Hong Kong commonly refer to these key phrases in discussions about the ESD scheme.

The query was fed to SMART, an information retrieval engine. SMART ranked all the English news articles (i.e. 4,259 articles) in our corpus. We examined the top nineteen documents (see Table 5.1 for statistics) that are relevant to the topic. We then combined the corresponding nineteen Chinese documents to form a set of thirty-eight documents. They were used in our subsequent experiments. For each pair of English-Chinese documents, we manually performed sentence alignment.

Documents related to the electronic service delivery (ESD) scheme	**Chinese**	**English**
Total no. of stories	19	19
Total no. of sentences	366	366
Average no. of sentence per story	19.3	19.3
No. of off-event sentences	188	188
No. of on-event sentences	178	178
Percentage of on-event sentences	48.6%	48.6%

Table 5.1: *Statistical details of the parallel documents (government news reports related to the electronic service delivery (ESD) scheme between 1st January 2001 and 31st June 2001) are shown. Since it is a sentence-by-sentence parallel in our corpus, statistics of Chinese and English documents are the same.*

The next step was to prepare the events in this topic for evaluation. We invited two annotators to extract all the events. The annotators judged whether each sentence should be labeled as on-event or off-event. The entire event list is shown in Appendix C.3. We show four sentences marked as on-event or off-event in the Table 5.2. Two on-event sentences were related to

the Post e-Cert joined in the ESD scheme (This event is labeled as 11 in the event list shown in Appendix C.3).

On-event sentences	**Event item no.**	**Off-event sentences**
New Initiatives on e-Cert Service in Support of the ESD Scheme	11	Each person can only submit one application.
The Postmaster General, Mr Luk Ping-chuen, today (January 16) announced Hong kong Post's new initiatives on e-Cert service in support of the formal launch of the Electronic Service Delivery (ESD) Scheme on January 19.	11	Organisation of Information Security Awareness Seminars for Secondary Schools Students

Table 5.2: *This table shows sentences (from an article on 16th January 2001) labeled as on-event or off-event manually. Event item number 11 referred to the event that the Post e-Cert joined the ESD scheme. Details of the event list is shown in Appendix C.3.*

5.3 Evaluation methods for the event-driven summaries generated from the parallel corpus

We computed the precision (P) and recall (R) by using the number of extracted on-event in the summary and the number of on-event sentences. The details of the recall and precision metrics is described in Section 4.5. Specif-

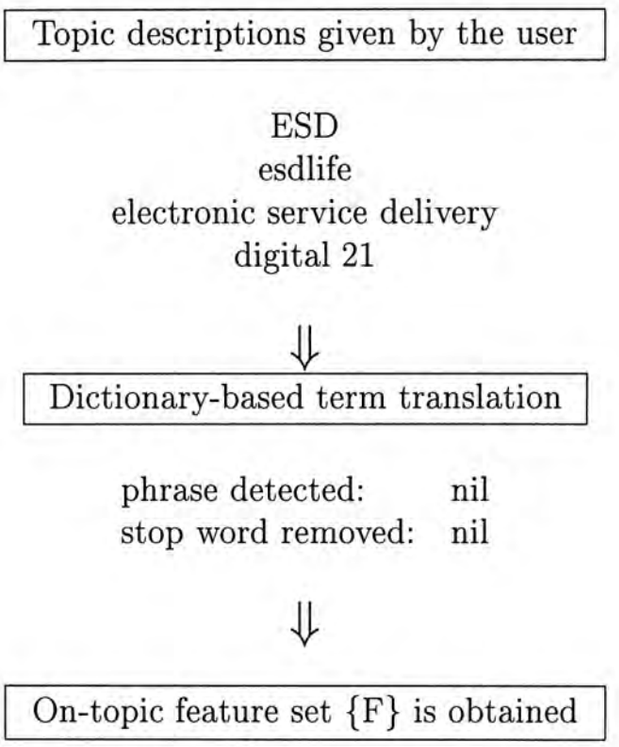
ically, the f-score is used as one of the evaluation metrics.

Furthermore, making use of the sentence-by-sentence parallel property of the corpus, we design an additional evaluation metric called *parallel event precision*. It measures how many on-event Chinese / English sentence pairs are extracted. The rationale is that parallel sentences belonging to the same event should be captured by a bilingual summarization system. We define the parallel event precision, P_m, as follows:

$$P_m = \frac{\{U_m\}}{\{E\} \cap \{U\}}$$

where $\{E\}$ is the set of sentences (in both Chinese and English) that are generated by our summarizer. $\{U\}$ denotes the set of on-event sentences in the entire training / testing set. $\{U_m\}$ denotes the number of extracted on-event Chinese / English sentence pairs. If a large number of on-event sentences extracted coexists with their corresponding English / Chinese sentences, a high P_m will be obtained.

Recall that a topic description is needed to compute one of the selection criteria: cluster-topic relevance, in cluster selection and summary generation processes. We used the query described in Section 5.2 as the topic description. The resulted on-topic feature set $\{F\}$ based on this topic description is shown in Table 5.3.



ESD	⇒ esd
esdlife	⇒ esdlife
electronic	⇒ 电子
service	⇒ 服务 / 功 / 役
delivery	⇒ 分娩
digital	⇒ 数字 / 数据 / 数码 / 数位
21	⇒ 二十一

Table 5.3: *This table shows the topic description and the on-topic feature set* $\{F\}$ *for the topic "Electronic service delivery (ESD) scheme".*

5.4 Experimental results and analysis

5.4.1 Experimental settings

We split the set of parallel documents into two partitions. The first partition consisted of the earlier half of the days, i.e. between 1st January 2001 and 31st March 2001. The second partition consisted of the remaining documents, i.e. between 1st April 2001 and 30th June 2001. The first partition was used for training and the second partition for testing. The basic information on the training and testing portions is given in Table 5.4. After taking the average of five runs, final results were obtained.

We also performed three-fold cross evaluation. Based on the total number of days, we divided entire data set into three parts. When one part was used in training, the two remaining parts were used for testing. We conducted three trails for each combination. As a result, the result of one-third in training and two-third in testing was obtained by averaging nine experimental results. We also did the two-third in training and one-third in testing results using the same procedure.

The training portion is used to obtain an effective threshold, θ, for incremental K-means clustering. Under each predefined compression rate, θ was varied from 0.001 to 0.300 in 0.001 intervals. The chosen θ in the training phase was used in the testing phase. If θ was set below 0.001, under-clustering

Documents related to electronic service delivery (ESD) scheme	**Training**	**Testing**
Total no. of sentences	298	434
No. of on-event sentences	196	160
Percentage of on-event sentences	65.8%	36.9%
Time-span	1st January 2001 to 31st March 2001	1st April 2001 to 30th June 2001

Table 5.4: *Statistical details of the parallel documents (government news reports related to electronic service delivery (ESD) scheme) for the training and testing sets.*

occurred, i.e. the number of cluster produced is smaller than the expected. In this situation, many sentences are crowded in few clusters only. If a large θ was used, many clusters were formed. The worst case is that each cluster consists of one sentence only.

Recall that the compression rate is defined as the ratio of the total number of sentence in a summary to that in the entire processing collection. In our experiments, various compression rates at 10%, 20%, 30%, 40%, and 50% will be studied. We also investigate a baseline method by randomly selecting sentences. For each combination of training and testing sets, we got the average baseline results by performing ten runs.

5.4.2 Results and analysis

We present our results in Figures 5.1 and 5.2. There are (x, z) values on the event-driven summarization plot in Figure 5.1. x denotes the precision score (P). z denotes the total number of clusters formed after the clustering process has completed. In Figure 5.2, notation 1/3:2/3 denotes that news articles from one portion are used for training and the remaining two portions are used for testing. The notation of 2/3:1/3 denotes that news of two portions are used for training and one portion is taken as testing.

As shown in Figure 5.1 and Figure 5.2, there is a substantial difference in the performance of our system generated summaries compared with the baseline method (randomly sentence selection). The precision score (the first element in the bracket) in the event-driven summarization plot always stays above 60% under all the investigated compression rates. This suggests that our summaries can achieve a fairly stable precision across different compression rates. Specifically for the curve with 1/3:2/3 (one-third in training and two-third in testing) in Figure 5.2, about 84% precision is obtained at the compression rate of 10%. As expected, when more data are used for testing, the outcome should be better.

Since the recall increases as the size of summary increases, the f-score raises gradually. The event-driven summarization curve tends to converge

around 50% because the number of on-event sentences extracted approaches to the total number of on-event sentences in the entire collection.

The final number of clusters produced was the same across different compression rates when 50% of training and 50% of testing was set. By investigating the data, we found that more or less the same set of sentences contained in 15 clusters under different θ. Table 5.5 shows those terms with weight over 0.5 found in the cluster mean. It shows that stable clusters included in the summary are stable under different compression rates.

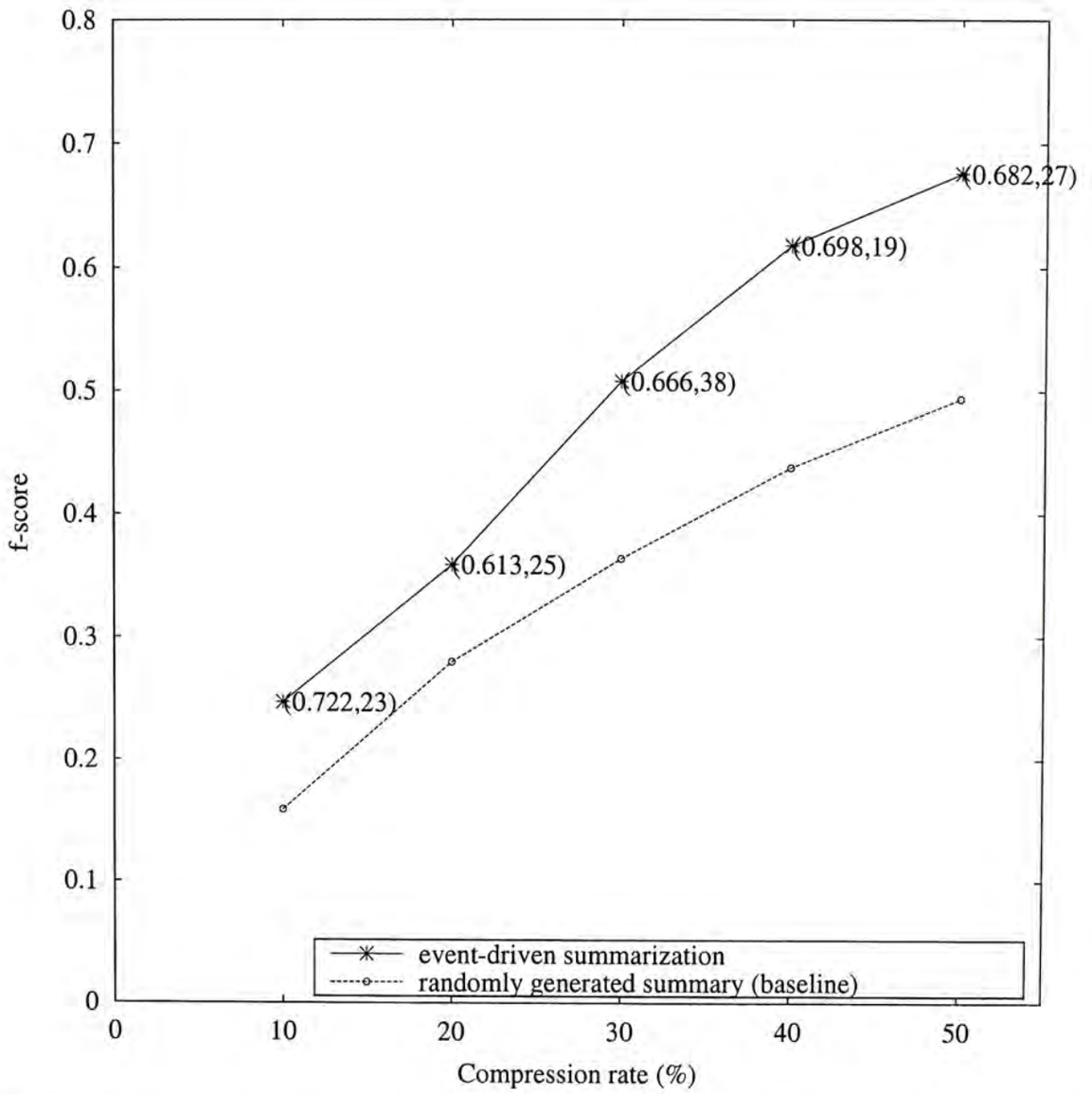


Figure 5.1: *This figure compares the performance of generated summaries (50% of days in training and 50% of days in testing) with the baseline method (i.e. by randomly selecting sentences from the collection). For event-driven summarization, there are (x, z) values. x denotes the precision score and z denotes the total number of clusters found by the incremental K-means clustering method.*

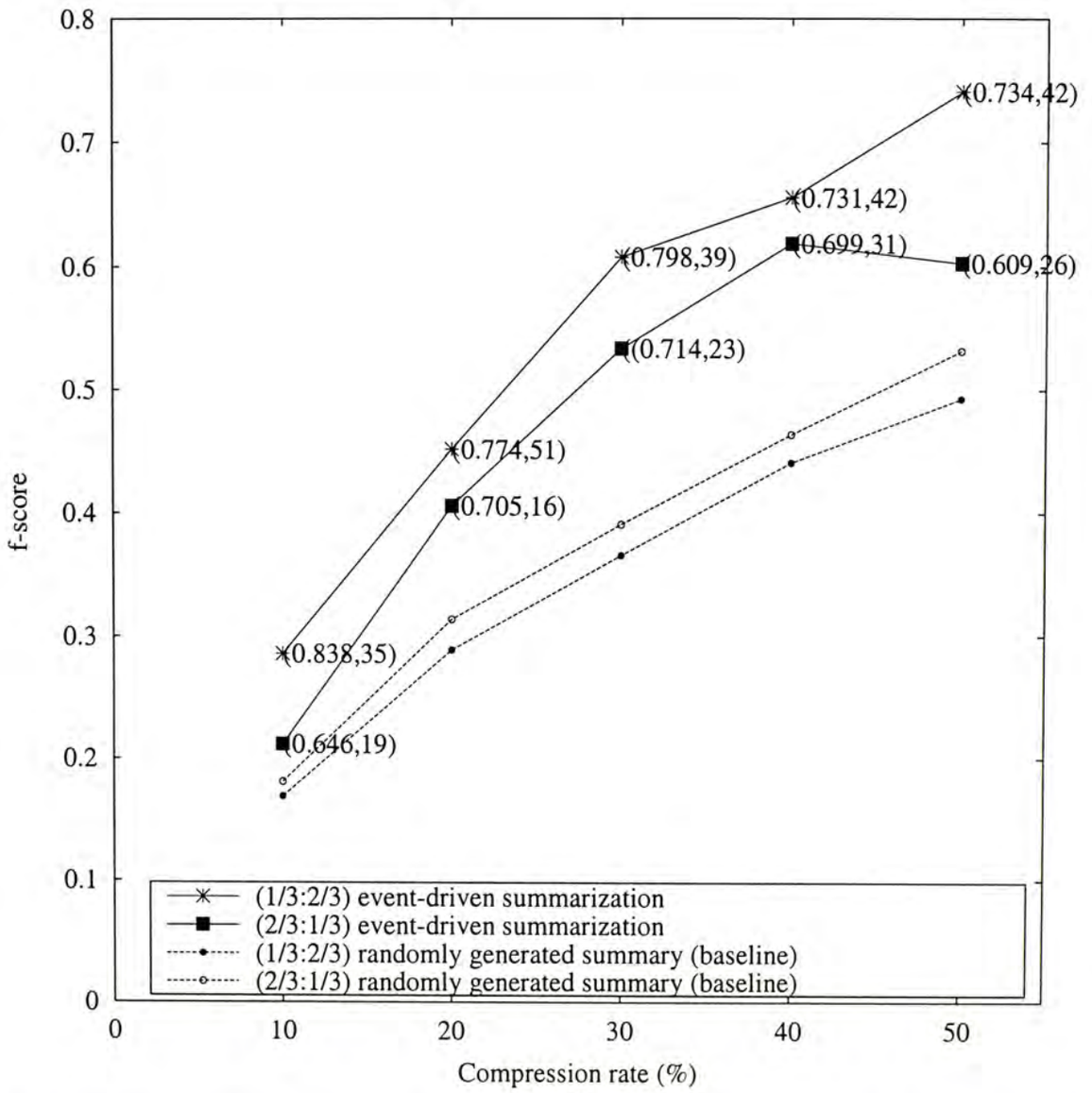


Figure 5.2: *This figure compares the performance of generated summaries with the baseline method (i.e. by randomly selecting sentences from the collection). Summaries are resulted by using different combinations (1/3:2/3 and 2/3:1/3) of training and testing sets.*

Figure 5.3 shows the evaluation of parallel event precision, P_m, across different compression rates. Generally, all P_m increases with the increase in compression rates. At low compression rates (10% and 20%), a high P_m is achieved for our event-driven summarization when training and testing portions are in 50% to 50%. As shown in Figure 5.1, precision (i.e. the fraction of the total number of on-event sentences appeared in the extracted sentences) obtains about 72% which is the second highest among different training and testing portions. Even the precision is the greatest (about 84%) of the curve (1/3:2/3) in Figure 5.2, the P_m (about 10%) is the lowest in Figure 5.3. This situation implies that even many testing samples are considered and many on-event sentences are extracted, they do not contain many parallel pairs (i.e. on-event sentences extracted coexists with their corresponding English/ Chinese sentences). This observation suggests that our summarizer is able to consider common elements in bilingual settings. If the training and testing portions are set in 50% to 50%, the P_m (parallel event precision) and P (precision) results are better at low compression rates (10-20%). That is many extracted sentences are on-event sentences and they coexist with their Chinese/ English sentences.

In a particular run with 50% training and 50% testing, there are 22 sentences out of 28 on-event sentences in the summary of 10% compression rate having the parallel characteristics. We show five pairs of bilingual summary

Compression rates				
10%	20%	30%	40%	50%
cluster 1 服务 (service) 电子 (electronic) 计划 (scheme)	cluster 1 服务 (service) 电子 (electronic) 计划 (scheme)	cluster 1 服务 (service) 电子 (electronic) 计划 (scheme)	cluster 1 服务 (service) 电子 (electronic) 计划 (scheme)	cluster 1 服务 (service) 电子 (electronic) 计划 (scheme)
		cluster 2 行 (hong) 缸 (kong)	cluster 2 行 (hong) 缸 (kong)	cluster 2 行 (hong) 缸 (kong)
			cluster 3 银行 (bank) 四月 (april) 二〇〇一年 (2001)	cluster 3 银行 (bank) 四月 (april) 二〇〇一年 (2001)
			cluster 4 香港 (hongkong) 邮政 (post)	cluster 4 香港 (hongkong) 邮政 (post)
				cluster 5 资讯 (information) 科技 (technology)
				cluster 6 8 11 pm

Table 5.5: *This table shows those terms with weight over 0.5 found in the clusters which are included in the summary. It is found that event clusters are stable under different compression rates. English word in the bracket is the translation of the above Chinese term by human for convenient read.*

sentences in Table 5.6. The missing six sentences lied in different clusters.

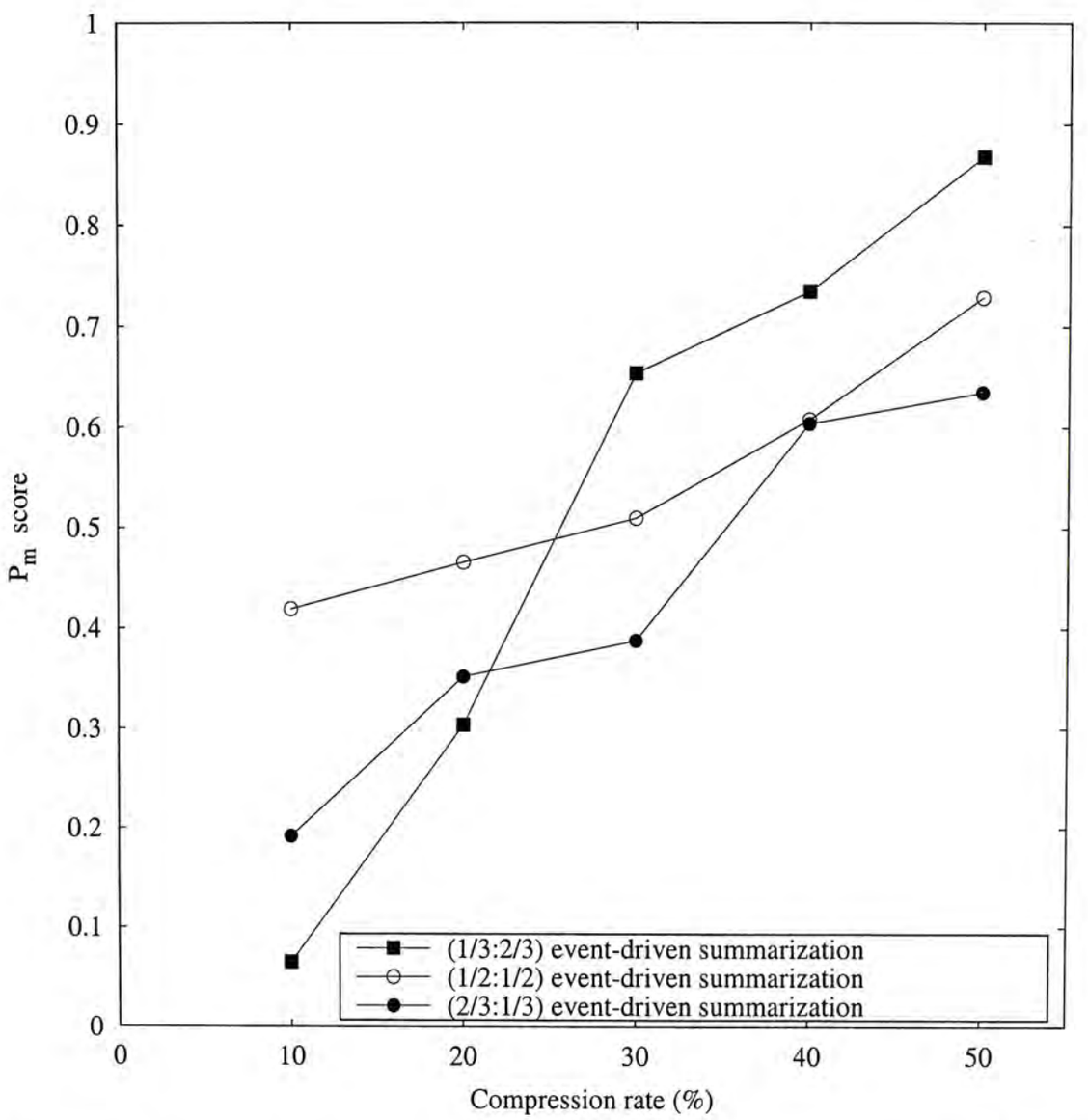


Figure 5.3: *This figure shows the performance of summarization using the parallel event precision, P_m. P_m scores always increases as the compression rate increases. Particularly, 50% to 50% of training and testing set obtains better P_m score at low compression rates (10-20%). This evidence shows that our summarizer is effective in considering common elements in bilingual corpus.*

稍後，数码证书持有人便可透过香港交易所第三代自动对盘系统，在网上买卖股票，亦可到公共服务电子化计划的网址享用政府服务，进行交易	↔	In the near future, it can be used for Internet stock trading (AMS/3 of Hong Kong Exchange) and to conduct transactions at the web site of Electronic Service Delivery (ESD) for government services.
政府出版的刊物将於二〇〇一年内透过公共服务电子化计划在网上售卖	↔	Online sale of Government publications will be implemented under the Electronic Service Delivery (ESD) Scheme within 2001.
第六轮公共服务电子化计划流动展览	↔	Sixth round of Electronic Service Delivery scheme roving show
由明日（五月一日）起的一个月内，市民可在多个大型购物商场，亲身了解「公共服务电子化」计划及熟习该计划的资讯服务站的操作方法。	↔	Members of the public will be able to familiarize themselves with the Electronic Service Delivery (ESD) scheme and the use of the ESD public kiosks at various shopping centres next month starting tomorrow (May 1).
随著「公共服务电子化」计划在今年一月正式推出後，第六轮「公共服务电子化」计划流动展览将於五月一日至三十日期间，轮流在六个大型购物商场举行。	↔	Following the formal launch of the ESD scheme in January this year, the sixth round of the ESD roving show will be staged at six popular shopping centres between May 1 and 31.

Table 5.6: *This table shows the bilingual on-event sentences in the summary of compression rate=10% .*

5.5 Chapter summary

This chapter aims to test the feasibility of applying our event-driven summarization approach for a parallel corpus. We have described the preparation of parallel documents. The experimental results demonstrate that our event-driven summarization approach is effective in extracting the on-event sentences from the bilingual corpus.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Our research goal is to investigate summarization approaches for capturing the main theme and events covered by a set of documents automatically. Our work can be divided into two parts, namely, the thematic term approach, and the event-driven approach for bilingual news summarization. The first part is a domain independent, single document summarization system. The thematic terms are extracted automatically from the entire corpus as well as from the corresponding documents using information retrieval techniques. A score for each sentence is then computed by considering both kinds of thematic terms. As a result, the extractive summary is generated by selecting the high-scoring sentences. We further investigated the feasibility of this

approach by applying it to a Chinese corpus.

In order to assess the performance of our summary, we adopt an intrinsic evaluation method by comparing the system generated summary with a standard human written summary. For English summaries, the performance evaluation was conducted by the content-based evaluation method. However, content-based evaluation could not be used for Chinese summaries since no handwritten summaries were unavailable. Consequently, we propose a new alternative method to evaluate the quality of summary by assessing the representative ability of a summary in acting as a surrogate within an entire data collection based on an information retrieval model. This new metric is called the Average Inverse Rank (AIR), is practically for the extrinsic evaluation. The results suggested that 20% and 10% of the full-length document is sufficient to be a good surrogate for the original document in both Chinese and English corpus respectively.

In the second part of our research, we developed a bilingual news summarization system based on event-driven approach. Firstly, we developed the dictionary-based term translation for handling English news via two steps, namely, the phrase translation method and the translation term disambiguation method. We then developed an incremental K-means clustering algorithm to discover coherent event clusters dynamically. Heuristic criteria, namely, the cluster-topic relevance and intra-cluster consistency were de-

veloped to select good clusters. Only highly topic relevant and consistent clusters were selected to compose the summary. Finally, we conducted our performance evaluation by recall, precision, and f-score measurements. The results indicated that our summary achieved around 60% precision at only 10% of the full-length documents.

We further investigated the effectiveness of our event-driven approach summarization technique by applying it to a parallel corpus. Apart from using the recall and precision method as the assessment, we proposed an additional metric, namely, the parallel event precision, to assess the quality of the event-based summaries. The experimental results showed that the precision achieved around 70% in all runs.

6.2 Future work

We suggest the following for improving the summarization performs of our method,

1. Name entities may be useful in discovering events. We plan to investigate how they could be incorporate in our event-driven approach to produce better summaries.

2. Explore the possibility of reducing redundant sentences in the summary.

3. In principle, our summarization approach could apply to the textual content of other kinds of data, eg. audio, video, etc. This area is also worth exploring in the future.

Bibliography

[1] http://lr-www.pi.titech.ac.jp/tsc/index-en.html.

[2] http://www.info.gov.hk/isd/news/index.htm and http://www.info.gov.hk/isd/news/cindex.htm.

[3] http://www.ldc.upenn.edu/Projects/Chinese/LDC_ch.htm#e2cdict.

[4] http://www.nist.gov/speech/tests/tdt/tdt2002/evalplan.htm.

[5] http://www.se.cuhk.edu.hk/ aoe.

[6] J. Allan, G. Rahul, and K. Vikas. Temporal summaries of news topics. In *Proceedings of the SIGIR*, pages 10–18, New Orleans, Louisiana, USA, 2001.

[7] R.K. Ando, B.K. Boguraev, R.J. Byrd, and M.S. Neff. Multi-document summarization by visualizing topical content. In *Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 79–88, Seattle, WA, 2000.

[8] L. Ballesteros and W.B. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, Melbourne, Australia, August 1998.

[9] M. Banko, V.O. Mittal, and M.J. Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Conference of the Association for Computational Linguistics, Hongkong, China*, pages 318–325, October 2000.

[10] R. Barzilay, K. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, Maryland, USA, June 1999.

[11] P.B. Baxendale. Man-made index for technical literature - an experiment. In *IBM Journal of Research and Development*, volume 2, pages 354–361. 1958.

[12] C. Buckley. Implementation of the smart information retrieval system. In *Cornell University Technical Report*, pages 85–686. 1985.

[13] W.T. Chuang and J. Yang. Extracting sentence segments for text summarization: a machine learning approach. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 152–159, Athens, Greece, 2000.

[14] M. Daniel. Discourse trees are good indicators of importance in text. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization. MIT Press*, pages 123–136, 1999.

[15] R.L. Donaway, K.D. Kevin, and A.M. Laura. A comparision of rankings produced by summarization evaluation measures. In *Proceedings of the ANLP/NAACL-2000 Workshop on Automatic Summarization*, pages 69–78, Seattle, Washington, USA, May 2000.

[16] H.P. Edmundson. New methods in automatic extracting. In *Journal of the Association for Computing Machinery*, volume 16, pages 264–285. April 1968.

[17] H. Eduard and C.Y. Lin. Automated text summarization and the summarist system. In *Proceedings of the TIPSTER Text Program, Phase III*, pages 197–214, 2000.

[18] J. Goldstein, M. Kantrowitz, V.O. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, pages 121–128, Berkeley, CA, August 1999.

[19] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–25, New Orleans, Louisiana, USA, 2001.

[20] M.L. Helen, C. Berlin, K. Sanjeev, L. Gina-Anne, K.L. Wai, and O. Douglas. Mandarin-English Information (MEI): Investigating translingual speech retrieval. In *Proceedings of the Human Language Technology Conference (HLT)*, pages 187–194, San Diego, California, 2001.

[21] M.Y. Kan and K.R. McKeown. Information extraction and summarization: Domain independence through focus types. In *Columbia University Computer Science Technical Report*, 1999.

[22] K. Knight and D. Marcu. Statistics-based summarization step one: Sen-

tence compression. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 703–710, Austin, TX, USA, 2000.

[23] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington, USA, July 1995.

[24] K.L. Kwok. Exploiting the ldc chinese-english bilingual wordlist for cross language information retrieval. In *International Journal of Computer Processing of Oriental Languages*, volume 14, pages 173–191. June 2001.

[25] W. Lam, C.Y. Wong, and K.F. Wong. Performance evaluation of character-, word-, and n-gram-based indexing for chinese text retrieval. In *2nd International Workshop on Information Retrieval with Asian Languages (IRAL)*, pages 68–80, 1997.

[26] S.H. Lo, M.L. Helen, and W. Lam. Multi-document summarization by visualizing topical content. In *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI)*, pages 113–118, Orlando, Florida, USA, July 2002.

[27] H.P. Luhn. The automatic creation of literature abstracts. In *IBM Journal of Research and Development*, volume 2, pages 159–165. April 1958.

[28] I. Mani. *Automatic Summarization.* John Benjamins Pub Co., June 2001.

[29] I. Mani, T. Firmin, D. House, G. Klein, B. Sundheim, and L. Hirschman. The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of EACL 1999*, pages 77–85, Bergen, Norway, June 8-12, 1999.

[30] D. Marcu. The automatic construction of large-scale corpora for summarization research. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 137–144, 1999.

[31] A.G. Miller. Wordnet: A lexical database for english. In *Proceedings of the Communications of the ACM Vol. 38, No.11*, pages 39–41, November 1995.

[32] V. Mittal and A. Berger. Ocelot: a system for summarizing web pages. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 144–151, Athens, Greece, 2000.

[33] V. Mittal and A. Berger. Query-relevant summarization using faqs. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 294–301, Hong Kong, 2000.

[34] Y. Nakao. An algorithm for one-page summarization for a long based on thematic hierarchy detection. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 302–309, Hong Kong, 2000.

[35] J.L. Neto, A.D. Santos, C.A.A. Kaestner, A.A. Freitas, and J.C. Nievol. A trainable algorithm for summarizing news stories. In *Proceedings of the PKDD Workshop on Machine Learning and Textual Information Access*, Lyon, France, September 2000.

[36] T. Nomoto and Y. Matsumoto. A new approach to unsupervised text summarization. In *Proceedings of the 24th Annual International ACM*

SIGIR Conference on Research and Development in Information Retrieval, pages 26–34, New Orleans, Louisiana, USA, 2001.

[37] C.D. Paice. The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases. In *Norman, O., Robertson, S., van Rijsbergen, C., and Williams, P., editors, Information Retrieval Research, London : Butterworth*, 1981.

[38] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–63, Melbourne, Austrailia, 1998.

[39] M.F. Porter. An algorithm for suffix stripping. In *Program, 14 no. 3*, pages 130–137, July 1980.

[40] D.R. Radev, J. Hongyan, and B. Malgorzata. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP/NAACL Workshop on Summarization*, pages 21–29, Seattle, Washington, USA, May 2000.

[41] C.J. Rijsbergen. *Information Retrieval.* Butterworths, London, 1979.

[42] G.C. Stein, G.B. Wise, T. Strzalkowski, and A. Bagga. Evaluating summaries for multiple documents in an interactive environment. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC00)*, pages 1651–1657, Athens, Greece, May 2000.

[43] S. Teufel and M. Moens. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In Inderjeet Mani and

Mark Maybury, editors, *Advances in automatic Text Summarization. MIT Press*, pages 58–65, 1999.

[44] S. Teufel and M. Moens. Sentence extraction as a classification task. In *Proceedings of ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pages 58–65, Madrid, Spain, 1999.

[45] P.D. Turney. Learning algorithms for keyphrase extraction. In *Information Retrieval*, volume 2, pages 303–336. 1999.

[46] F. Walls, H. Jin, S. Sista, and R. Schwartz. Topic detection in broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*, pages 193–198, Herndon, Virginia, 1999.

Appendix A

English Stop Word List

!	”	***	***
***	***	’	(
)	*	+	,
-	–	.	.I
.W	/	:	;
¡	=	¿	?
@	EMPTYLINE	***	
***	***	***	‘
a	a’s	able	about
above	according	accordingly	across
actually	add	after	afterwards
again	against	ain’t	all
allow	allows	almost	alone
along	already	also	although
always	am	among	amongst
an	and	another	any
anybody	anyhow	anyone	anything
anyway	anyways	anywhere	apart
appear	appreciate	appropriate	are
area	aren’t	around	as
aside	ask	asking	associated
at	available	away	awfully
b	back	be	became
because	become	becomes	becoming
been	before	beforehand	behind
being	believe	below	beside
besides	best	better	between

beyond	both	brief	but
by	c	c'mon	c's
came	can	can't	cannot
cant	cause	causes	cent
certain	certainly	changes	chief
clearly	co	com	come
comes	concerning	consequently	consider
considering	contain	containing	contains
corresponding	could	couldn't	course
currently	d	definitely	described
despite	did	didn't	different
do	does	doesn't	doing
don't	done	double	down
downwards	during	e	each
eas	edu	eg	eight
either	else	elsewhere	enough
entirely	especially	et	etc
even	ever	every	everybody
everyday	everyone	everything	everywhere
ex	exactly	example	except
f	far	few	fifth
first	five	followed	following
follows	for	former	formerly
forth	four	from	further
furthermore	g	get	gets
getting	gif	give	given
gives	go	goes	going
gone	got	gotten	greetings
h	had	hadn't	happens
hardly	has	hasn't	have
haven't	having	he	he's
hello	help	hence	her
here	here's	hereafter	hereby
herein	hereupon	hers	herself
hi	him	himself	his
hither	hopefully	how	howbeit
however	i	i'd	i'll
i'm	i've	ie	if
ignored	immediate	in	inasmuch
inc	includ	indeed	indicate
indicated	indicates	inner	insofar

instead	into	introduc	inward
is	isn't	it	it'd
it'll	it's	its	itself
j	just	k	keep
keeps	kept	know	known
knows	l	last	lately
later	latter	latterly	least
less	lest	let	let's
like	liked	likely	little
look	looking	looks	ltd
m	mainly	many	may
maybe	me	mean	meanwhile
merely	might	more	moreover
most	mostly	mr	much
must	my	myself	n
name	namely	nd	near
nearly	necessary	need	needs
neither	never	nevertheless	new
next	nine	no	nobody
non	none	noone	nor
normally	not	nothing	novel
novemb	now	nowhere	o
obviously	of	off	often
oh	ok	okay	old
on	once	one	ones
only	onto	or	other
others	otherwise	ought	our
ours	ourselves	out	outside
over	overall	own	p
particular	particularly	per	perhaps
placed	please	plus	possible
pp	prepar	presumably	probably
provides	q	que	quite
qv	r	rather	rd
re	really	reasonably	regarding
regardless	regards	relatively	respectively
right	s	said	same
saw	say	saying	says
second	secondly	see	seeing
seem	seemed	seeming	seems
seen	self	selves	sensible

sent	serious	seriously	seven
several	shall	she	short
should	shouldn't	since	six
so	some	somebody	somehow
someone	something	sometime	sometimes
somewhat	somewhere	soon	sorry
specified	specify	specifying	still
stor	story	sub	such
sup	sure	t	t's
take	taken	tell	tends
th	than	thank	thanks
thanx	that	that's	thats
the	their	theirs	them
themselves	then	thence	there
there's	thereafter	thereby	therefore
therein	theres	thereupon	these
they	they'd	they'll	they're
they've	think	third	this
thorough	thoroughly	those	though
three	through	throughout	thru
thus	to	together	too
took	toward	towards	tried
tries	truly	try	trying
twice	two	u	un
under	unfortunately	unless	unlikely
until	unto	up	upon
us	use	used	useful
uses	using	usually	uucp
v	value	various	very
via	viz	vs	w
want	wants	was	wasn't
way	we	we'd	we'll
we're	we've	welcome	well
went	were	weren't	what
what's	whatever	when	whence
whenever	where	where's	whereafter
whereas	whereby	wherein	whereupon
wherever	whether	which	while
whither	who	who's	whoever
whole	whom	whose	why
will	willing	wish	with

within	without	won't	wonder
would	wouldn't	www	x
y	year	yes	yesterday
yet	you	you'd	you'll
you're	you've	your	yours
yourself	yourselves	z	zero
zoom	***	—	***

Table A.1: *English stop word list*

Appendix B

Chinese Stop Word List

、	。	•	—	——	~
…	……	‘	’	“	”
〔	〕	〈	〉	《	》
！	（	）	，	——	／
：	；	？		按照	八
把	包含	包括	保梗福赌	被	本
比	比较	比如	并	并非	并且
不	不过	不久前	不能	部	才
长此以往	成	成为	初	出	此
次	从	从而年	大	但	但是
当	当今	当夜	当前	当然	当日
当时	当晚	等	到	道	得
的	的话	都	底	地	第一
第一个	电	而已常	对	对於	多
而	而且	非更国	二	发	乏
法	反而	何	附近	该	高
个	各	或	庚果很	工	公元去
公元前	关於	几天来	机	过	过要时
好	和	将	既	还	还
还有	会	今近均来	将於	及	即
几百颗	几千颗	量	今年过	既是	家
简称	将	了	经始	交	皆是
届	届时		开里	今日	今天
今晚	今後			九	就
就是	具			看	可
可能	可以			例	两个
两	两翼			凌晨	另外

六	吗	没	没有	每	们
明年	明天	名	末	目前	目前为止
那	那时	那些	那麽	内	内外
能	能够	你	你们	年	偏
七	其实	其他	其它	前	面
前年	清晨	请	去	去冬今春	前面年
全	却	然	然而	让	去年
人们	任何	认为	仍	仍然	人
如	如果	如今	若	三	仍有
上年	上午	甚至	生	十	上
时	是	是次	适当	说	十分
所	所有	他	他们	它	四
她	台	太	特别	提	它们
天主教堂	同	同期	同时	外	天
为	为例	未	文	问	晚秋
我们	无	五	细	下	我
先前	现	现阶段	现今	现时	下午
现在	相反	想	项	像	现有
新	心	信	星期六	需	小
需要	须要	许多	学	讯	需求
也	也是	也许	夜间	一	要
已	已经	以	以及	以上	依然
易	亦	意	翌年	因	以至
因而	因为	印刷厂	应	应当	因此
由	由此因春	由於	有	又	用
元月	原	约	月底	再	与
早	早	早秋	则	站	在
者	这	这个	这时	这是	章
真	知	之	之一	只	这些
只是	至	中	中旬	种	只能
资	子	子时	自	自己	著
最大	最近	最後	昨天	卅	最
於	於是	麽			後

Table B.1: *Chinese stop word list*

Appendix C

Event List Items on the Corpora

C.1 Event list items for the topic "Upcoming Philippine election"

1. 时间 (有关 选举 及 公布结果) (the time of the election and the result announced) 2. 总统 的 背景 如 任期 (the background of the president) 3. 竞选 活动 如 听政会 、 演讲 等 (the campaign of the election) 4. 政党/候选人 的 描述 、 背景 及 活动 (the description/background of the candidates and the activities related to them) 5. 公众 对 选举 的 期望 及 态度 (the expectation and the attitude of the public towards the election) 6. 教会 对 候选人/选举活动 作出 的 回应 (the response of the church towards the candidates and the election) 7. 投票 活动 及 状况 (the description of the voting) 8. 有关 选举 活动 引起 的 罪恶 , 如 抢击 、 暴力 、 凶杀 、 游击队 活动 和 舞弊 等 (the description of the crime came out due to the election) 9. 警方 对 选举事件 作出 的 保安工作 (security action by the police for the election) 10. 描述 因 选举 造成 的 比索 升降 (the effect on currency (peso) due to the election) 11. 对 选举 所作 的 调查 及 结果 (the investigation conducted for the election) 12. 点票 活动 及 结果 (the counting and the result of the vote) 13. 公布 选举 结果 (the announcement of the result for the election)

Table C.1: *Event lists items found in the bilingual corpus (Upcoming Philippine election).*

C.2 Event list items for the topic "German train derail"

1. 客车出轨 发生 的 时间 、 地方 及 原因 (the time, the place and the reason of the derail)
2. 描述 事件 中 出现 的 乘客/司机 , 他(们)/她(们) 在 事件 发生 前/时/後 的 状态 (the description of the passengers/drivers who got involved in the derail)
3. 描述 事件 中 出现 的 火车/车轮/铁桥 , 它 在 事件 发生 前/时/後 的 状态 (the description of the train/wheel/bridge that were related to the derail)
4. 死伤报告 (the death or injury report)
5. 救援工作 (the rescue work)
6. 现场 清理 工作 (the clean up work)
7. 跟进 工作 如 火车 要 进行 安全检查 、 停止 服务 或 减低 车速 等 (the follow up work: the safety check for the trains, the termination of the train service, reduction in speed, etc)

Table C.2: *Event lists items found in the bilingual corpus (German train derail).*

C.3 Event list items for the topic "Electronic service delivery (ESD) scheme"

1. 公共 服务 电子化 投入 服务 时间 (the time of the launch of the ESD scheme)
2. ESD 开幕 事件 (the opening ceremony of the ESD scheme)
3. 介绍 ESD 的背景 (the background of the ESD scheme)
4. 公共 服务 电子化 讲助 (the seminars for the ESD scheme)
5. 公共 电子化 服务 有关 要求 (the requirement of the ESD scheme)
6. ESD 展览会 (the show of the ESD scheme)
7. 取得 ESD 服务 的 途径 (the way to get the ESD service)
8. ESD 提供 的 服务 (the services provided in the ESD scheme)
9. 生活站 所 提供的 ESD 服务 (the services provided in the esdlife web site)
10. 社区 数码 站 (the information related to the ESD kiosks)
11. 邮政 电子 核证 服务 参与 公共 电子 化 计划 (Post e-Cert joined in the ESD scheme)
12. 邮政 电子 核证 服务 有关 资诉 (the information related to the ESD scheme)
13. 邮政局 对 ESD 的 支持 (the support of the ESD by the Post office)
14. 为 ESD 发行 的 纪念封 (an official souvenir cover to commemorate the launch of the ESD scheme)
15. 智能咭 资诉 (the information related to the smart card)
16. 检讨 或 研究工作 (the reviewing work on the ESD services)

Table C.3: *Event lists items found in parallel documents (government news reports related to ESD scheme between 1st January 2001 and 31st June 2001).*

Appendix D

The sample of an English article (9505001.xml).

Introduction
In collaborative consultation dialogues, the consultant and the executing agent collaborate on developing a plan to achieve the executing agent's domain goal. Since agents are autonomous and heterogeneous, it is inevitable that conflicts in their beliefs arise during the planning process. In such cases, collaborative agents should attempt to square away the conflicts by engaging in collaborative negotiation to determine what should constitute their shared plan of actions and shared beliefs. Collaborative negotiation differs from non-collaborative negotiation and argumentation mainly in the attitude of the participants, since collaborative agents are not self-centered, but act in a way as to benefit the agents as a group. Thus, when facing a conflict, a collaborative agent should not automatically reject a belief with which she does not agree; instead, she should evaluate the belief and the evidence provided to her and adopt the belief if the evidence is convincing. On the other hand, if the evaluation indicates that the agent should maintain her original belief, she should attempt to provide sufficient justification to convince the other agent to adopt this belief if the belief is relevant to the task at hand. This paper presents a model for engaging in collaborative negotiation to resolve conflicts in agents' beliefs about domain knowledge. Our model 1) detects conflicts in beliefs and initiates a negotiation subdialogue only when the conflict is relevant to the current task, 2) selects the most effective aspect to address in its pursuit of conflict resolution when multiple conflicts exist, 3) selects appropriate evidence to justify the system's proposed modification of the user's beliefs, and 4) captures the negotiation process in a recursive Propose-Evaluate-Modify cycle of actions, thus enabling the system to handle embedded negotiation subdialogues.

Related Work

Researchers have studied the analysis and generation of arguments ; however, agents engaging in argumentative dialogues are solely interested in winning an argument and thus exhibit different behavior from collaborative agents. Sidner sid_aaaiws92,sid_aaai94 formulated an artificial language for modeling collaborative discourse using proposal/acceptance and proposal/rejection sequences; however, her work is descriptive and does not specify response generation strategies for agents involved in collaborative interactions. Webber and Joshi web_jos_coling82 have noted the importance of a cooperative system providing support for its responses. They identified strategies that a system can adopt in justifying its beliefs; however, they did not specify the criteria under which each of these strategies should be selected. Walker wal_coling94 described a method of determining when to include optional warrants to justify a claim based on factors such as communication cost, inference cost, and cost of memory retrieval. However, her model focuses on determining when to include informationally redundant utterances, whereas our model determines whether or not justification is needed for a claim to be convincing and, if so, selects appropriate evidence from the system's private beliefs to support the claim. Caswey et al. , introduced the idea of utilizing a belief revision mechanism to predict whether a set of evidence is sufficient to change a user's existing belief and to generate responses for information retrieval dialogues in a library domain. They argued that in the library dialogues they analyzed, "in no cases does negotiation extend beyond the initial belief conflict and its immediate resolution." . However, our analysis of naturally-occurring consultation dialogues , shows that in other domains conflict resolution does extend beyond a single exchange of conflicting beliefs; therefore we employ a recursive model for collaboration that captures extended negotiation and represents the structure of the discourse. Furthermore, their system deals with a single conflict, while our model selects a focus in its pursuit of conflict resolution when multiple conflicts arise. In addition, we provide a process for selecting among multiple possible pieces of evidence.

Features of Collaborative Negotiation

Collaborative negotiation occurs when conflicts arise among agents developing a shared plan during collaborative planning. A collaborative agent is driven by the goal of developing a plan that best satisfies the interests of all the agents as a group, instead of one that maximizes his own interest. This results in several distinctive features of collaborative negotiation: 1) A collaborative agent does not insist on winning an argument, and may change his beliefs if another agent presents convincing justification for an opposing belief. This differentiates collaborative negotiation from argumentation . 2) Agents involved in collaborative negotiation are open and honest with

one another; they will not deliberately present false information to other agents, present information in such a way as to mislead the other agents, or strategically hold back information from other agents for later use. This distinguishes collaborative negotiation from non-collaborative negotiation such as labor negotiation . 3) Collaborative agents are interested in others' beliefs in order to decide whether to revise their own beliefs so as to come to agreement . Although agents involved in argumentation and non-collaborative negotiation take other agents' beliefs into consideration, they do so mainly to find weak points in their opponents' beliefs and attack them to win the argument. In our earlier work, we built on Sidner's proposal/acceptance and proposal/rejection sequences and developed a model that captures collaborative planning processes in a Propose-Evaluate-Modify cycle of actions . This model views collaborative planning as agent A proposing a set of actions and beliefs to be incorporated into the shared plan being developed, agent B evaluating the proposal to determine whether or not he accepts the proposal and, if not, agent B proposing a set of modifications to A's original proposal. The proposed modifications will again be evaluated by A, and if conflicts arise, she may propose modifications to B's previously proposed modifications, resulting in a recursive process. However, our research did not specify, in cases where multiple conflicts arise, how an agent should identify which part of an unaccepted proposal to address or how to select evidence to support the proposed modification. This paper extends that work by incorporating into the modification process a strategy to determine the aspect of the proposal that the agent will address in her pursuit of conflict resolution, as well as a means of selecting appropriate evidence to justify the need for such modification.

Response Generation in Collaborative Negotiation

In order to capture the agents' intentions conveyed by their utterances, our model of collaborative negotiation utilizes an enhanced version of the dialogue model described in to represent the current status of the interaction. The enhanced dialogue model has four levels: the domain level which consists of the domain plan being constructed for the user's later execution, the problem-solving level which contains the actions being performed to construct the domain plan, the belief level which consists of the mutual beliefs pursued during the planning process in order to further the problem-solving intentions, and the discourse level which contains the communicative actions initiated to achieve the mutual beliefs . This paper focuses on the evaluation and modification of proposed beliefs, and details a strategy for engaging in collaborative negotiations.

Evaluating Proposed Beliefs

Our system maintains a set of beliefs about the domain and about the user's

beliefs. Associated with each belief is a strength that represents the agent's confidence in holding that belief. We model the strength of a belief using endorsements, which are explicit records of factors that affect one's certainty in a hypothesis , following ,. Our endorsements are based on the semantics of the utterance used to convey a belief, the level of expertise of the agent conveying the belief, stereotypical knowledge, etc. The belief level of the dialogue model consists of mutual beliefs proposed by the agents' discourse actions. When an agent proposes a new belief and gives (optional) supporting evidence for it, this set of proposed beliefs is represented as a belief tree, where the belief represented by a child node is intended to support that represented by its parent. The root nodes of these belief trees (top-level beliefs) contribute to problem-solving actions and thus affect the domain plan being developed. Given a set of newly proposed beliefs, the system must decide whether to accept the proposal or to initiate a negotiation dialogue to resolve conflicts. The evaluation of proposed beliefs starts at the leaf nodes of the proposed belief trees since acceptance of a piece of proposed evidence may affect acceptance of the parent belief it is intended to support. The process continues until the top-level proposed beliefs are evaluated. Conflict resolution strategies are invoked only if the top-level proposed beliefs are not accepted because if collaborative agents agree on a belief relevant to the domain plan being constructed, it is irrelevant whether they agree on the evidence for that belief . In determining whether to accept a proposed belief or evidential relationship, the evaluator first constructs an evidence set containing the system's evidence that supports or attacks _bel and the evidence accepted by the system that was proposed by the user as support for _bel. Each piece of evidence contains a belief _beli, and an evidential relationship supports(_beli,_bel). Following Walker's weakest link assumption the strength of the evidence is the weaker of the strength of the belief and the strength of the evidential relationship. The evaluator then employs a simplified version of Galliers' belief revision mechanism, to compare the strengths of the evidence that supports and attacks _bel. If the strength of one set of evidence strongly outweighs that of the other, the decision to accept or reject _bel is easily made. However, if the difference in their strengths does not exceed a pre-determined threshold, the evaluator has insufficient information to determine whether to adopt _bel and therefore will initiate an information-sharing subdialogue to share information with the user so that each of them can knowledgably re-evaluate the user's original proposal. If, during information-sharing, the user provides convincing support for a belief whose negation is held by the system, the system may adopt the belief after the re-evaluation process, thus resolving the conflict without negotiation.
Example

To illustrate the evaluation of proposed beliefs, consider the following utterances: I think Dr. Smith is teaching AI next semester. Dr. Smith is not teaching AI. He is going on sabbatical next year. Figure shows the belief and discourse levels of the dialogue model that captures utterances () and (). The belief evaluation process will start with the belief at the leaf node of the proposed belief tree, On-Sabbatical(Smith,next year)). The system will first gather its evidence pertaining to the belief, which includes 1) a warranted belief that Dr. Smith has postponed his sabbatical until 1997 (Postponed-Sabbatical(Smith,1997)), 2) a warranted belief that Dr. Smith postponing his sabbatical until 1997 supports the belief that he is not going on sabbatical next year (supports(Postponed-Sabbatical(Smith,1997), On-Sabbatical(Smith,next year)), 3) a strong belief that Dr. Smith will not be a visitor at IBM next year (visitor(Smith, IBM, next year)), and 4) a warranted belief that Dr. Smith not being a visitor at IBM next year supports the belief that he is not going on sabbatical next year (supports(visitor(Smith, IBM, next year), On-Sabbatical(Smith, next year)), perhaps because Dr. Smith has expressed his desire to spend his sabbatical only at IBM). The belief revision mechanism will then be invoked to determine the system's belief about On-Sabbatical(Smith, next year) based on the system's own evidence and the user's statement. Since beliefs (1) and (2) above constitute a warranted piece of evidence against the proposed belief and beliefs (3) and (4) constitute a strong piece of evidence against it, the system will not accept On-Sabbatical(Smith, next year). The system believes that being on sabbatical implies a faculty member is not teaching any courses; thus the proposed evidential relationship will be accepted. However, the system will not accept the top-level proposed belief, Teaches(Smith, AI), since the system has a prior belief to the contrary (as expressed in utterance (1)) and the only evidence provided by the user was an implication whose antecedent was not accepted.

Modifying Unaccepted Proposals

The collaborative planning principle in , suggests that "conversants must provide evidence of a detected discrepancy in belief as soon as possible." Thus, once an agent detects a relevant conflict, she must notify the other agent of the conflict and initiate a negotiation subdialogue to resolve it – to do otherwise is to fail in her responsibility as a collaborative agent. We capture the attempt to resolve a conflict with the problem-solving action Modify-Proposal, whose goal is to modify the proposal to a form that will potentially be accepted by both agents. When applied to belief modification, Modify-Proposal has two specializations: Correct-Node, for when a proposed belief is not accepted, and Correct-Relation, for when a proposed evidential relationship is not accepted. Figure shows the problem-solving recipes for

Correct-Node and its subaction, Modify-Node, that is responsible for the actual modification of the proposal. The applicability conditions of Correct-Node specify that the action can only be invoked when _s1 believes that _node is not acceptable while _s2 believes that it is (when _s1 and _s2 disagree about the proposed belief represented by _node). However, since this is a collaborative interaction, the actual modification can only be performed when both _s1 and _s2 believe that _node is not acceptable – that is, the conflict between _s1 and _s2 must have been resolved. This is captured by the applicability condition and precondition of Modify-Node. The attempt to satisfy the precondition causes the system to post as a mutual belief to be achieved the belief that _node is not acceptable, leading the system to adopt discourse actions to change _s2's beliefs, thus initiating a collaborative negotiation subdialogue.

Selecting the Focus of Modification

When multiple conflicts arise between the system and the user regarding the user's proposal, the system must identify the aspect of the proposal on which it should focus in its pursuit of conflict resolution. For example, in the case where Correct-Node is selected as the specialization of Modify-Proposal, the system must determine how the parameter _node in Correct-Node should be instantiated. The goal of the modification process is to resolve the agents' conflicts regarding the unaccepted top-level proposed beliefs. For each such belief, the system could provide evidence against the belief itself, address the unaccepted evidence proposed by the user to eliminate the user's justification for the belief, or both. Since collaborative agents are expected to engage in effective and efficient dialogues, the system should address the unaccepted belief that it predicts will most quickly resolve the top-level conflict. Therefore, for each unaccepted top-level belief, our process for selecting the focus of modification involves two steps: identifying a candidate foci tree from the proposed belief tree, and selecting a focus from the candidate foci tree using the heuristic "attack the belief(s) that will most likely resolve the conflict about the top-level belief." A candidate foci tree contains the pieces of evidence in a proposed belief tree which, if disbelieved by the user, might change the user's view of the unaccepted top-level proposed belief (the root node of that belief tree). It is identified by performing a depth-first search on the proposed belief tree. When a node is visited, both the belief and the evidential relationship between it and its parent are examined. If both the belief and relationship were accepted by the evaluator, the search on the current branch will terminate, since once the system accepts a belief, it is irrelevant whether it accepts the user's support for that belief . Otherwise, this piece of evidence will be included in the candidate foci tree and the system will continue to search through the evidence in the belief tree proposed

as support for the unaccepted belief and/or evidential relationship. Once a candidate foci tree is identified, the system should select the focus of modification based on the likelihood of each choice changing the user's belief about the top-level belief. Figure shows our algorithm for this selection process. Given an unaccepted belief (_bel) and the beliefs proposed to support it, Select-Focus-Modification will annotate _bel with 1) its focus of modification (_bel. focus), which contains a set of beliefs (_bel and/or its descendents) which, if disbelieved by the user, are predicted to cause him to disbelieve _bel, and 2) the system's evidence against _bel itself (_bel. s-attack). Select-Focus-Modification determines whether to attack _bel's supporting evidence separately, thereby eliminating the user's reasons for holding _bel, to attack _bel itself, or both. However, in evaluating the effectiveness of attacking the proposed evidence for _bel, the system must determine whether or not it is possible to successfully refute a piece of evidence (i.e., whether or not the system believes that sufficient evidence is available to convince the user that a piece of proposed evidence is invalid), and if so, whether it is more effective to attack the evidence itself or its support. Thus the algorithm recursively applies itself to the evidence proposed as support for _bel which was not accepted by the system (step). In this recursive process, the algorithm annotates each unaccepted belief or evidential relationship proposed to support _bel with its focus of modification (_beli.focus) and the system's evidence against it (_beli.s-attack). _beli.focus contains the beliefs selected to be addressed in order to change the user's belief about _beli, and its value will be nil if the system predicts that insufficient evidence is available to change the user's belief about _beli. Based on the information obtained in step , Select-Focus-Modification decides whether to attack the evidence proposed to support _bel, or _bel itself (step). Its preference is to address the unaccepted evidence, because McKeown's focusing rules suggest that continuing a newly introduced topic (about which there is more to be said) is preferable to returning to a previous topic . Thus the algorithm first considers whether or not attacking the user's support for _bel is sufficient to convince him of _bel (step). It does so by gathering (in cand-set) evidence proposed by the user as direct support for _bel but which was not accepted by the system and which the system predicts it can successfully refute (i.e., _beli.focus is not nil). The algorithm then hypothesizes that the user has changed his mind about each belief in cand-set and predicts how this will affect the user's belief about _bel (step). If the user is predicted to accept _bel under this hypothesis, the algorithm invokes Select-Min-Set to select a minimum subset of cand-set as the unaccepted beliefs that it would actually pursue, and the focus of modification (_bel.focus) will be the union of the focus for each of the beliefs in this minimum subset. If attacking the evidence

for _bel does not appear to be sufficient to convince the user of _bel, the algorithm checks whether directly attacking _bel will accomplish this goal. If providing evidence directly against _bel is predicted to be successful, then the focus of modification is _bel itself (step). If directly attacking _bel is also predicted to fail, the algorithm considers the effect of attacking both _bel and its unaccepted proposed evidence by combining the previous two prediction processes (step). If the combined evidence is still predicted to fail, the system does not have sufficient evidence to change the user's view of _bel; thus, the focus of modification for _bel is nil (step). and of the algorithm invoke a function, Predict, that makes use of the belief revision mechanism discussed in Section to predict the user's acceptance or unacceptance of _bel based on the system's knowledge of the user's beliefs and the evidence that could be presented to him . The result of Select-Focus-Modification is a set of user beliefs (in _bel.focus) that need to be modified in order to change the user's belief about the unaccepted top-level belief. Thus, the negations of these beliefs will be posted by the system as mutual beliefs to be achieved in order to perform the Modify actions.

Selecting Justification for a Claim

Studies in communication and social psychology have shown that evidence improves the persuasiveness of a message. Research on the quantity of evidence indicates that there is no optimal amount of evidence, but that the use of high-quality evidence is consistent with persuasive effects . On the other hand, Grice's maxim of quantity specifies that one should not contribute more information than is required. Thus, it is important that a collaborative agent selects sufficient and effective, but not excessive, evidence to justify an intended mutual belief. To convince the user of a belief, _bel, our system selects appropriate justification by identifying beliefs that could be used to support _bel and applying filtering heuristics to them. The system must first determine whether justification for _bel is needed by predicting whether or not merely informing the user of _bel will be sufficient to convince him of _bel. If so, no justification will be presented. If justification is predicted to be necessary, the system will first construct the justification chains that could be used to support _bel. For each piece of evidence that could be used to directly support _bel, the system first predicts whether the user will accept the evidence without justification. If the user is predicted not to accept a piece of evidence (evidi), the system will augment the evidence to be presented to the user by posting evidi as a mutual belief to be achieved, and selecting propositions that could serve as justification for it. This results in a recursive process that returns a chain of belief justifications that could be used to support _bel. Once a set of beliefs forming justification chains is identified, the system must then select from this set those belief chains which, when pre-

sented to the user, are predicted to convince the user of ¬bel. Our system will first construct a singleton set for each such justification chain and select the sets containing justification which, when presented, is predicted to convince the user of ¬bel. If no single justification chain is predicted to be sufficient to change the user's beliefs, new sets will be constructed by combining the single justification chains, and the selection process is repeated. This will produce a set of possible candidate justification chains, and three heuristics will then be applied to select from among them. The first heuristic prefers evidence in which the system is most confident since high-quality evidence produces more attitude change than any other evidence form . Furthermore, the system can better justify a belief in which it has high confidence should the user not accept it. The second heuristic prefers evidence that is novel to the user, since studies have shown that evidence is most persuasive if it is previously unknown to the hearer ,. The third heuristic is based on Grice's maxim of quantity and prefers justification chains that contain the fewest beliefs.

Example

After the evaluation of the dialogue model in Figure , Modify-Proposal is invoked because the top-level proposed belief is not accepted. In selecting the focus of modification, the system will first identify the candidate foci tree and then invoke the Select-Focus-Modification algorithm on the belief at the root node of the candidate foci tree. The candidate foci tree will be identical to the proposed belief tree in Figure since both the top-level proposed belief and its proposed evidence were rejected during the evaluation process. This indicates that the focus of modification could be either Teaches(Smith,AI) or On-Sabbatical(Smith, next year) (since the evidential relationship between them was accepted). When Select-Focus-Modification is applied to Teaches(Smith,AI), the algorithm will first be recursively invoked on On-Sabbatical(Smith, next year) to determine the focus for modifying the child belief (step 3.1 in Figure). Since the system has two pieces of evidence against On-Sabbatical(Smith, next year), 1) a warranted piece of evidence containing Postponed-Sabbatical(Smith,1997) and supports(Postponed-Sabbatical(Smith,1997),On-Sabbatical(Smith, next year)), and 2) a strong piece of evidence containing visitor(Smith,IBM,next year) and supports(visitor(Smith,IBM,next year),On-Sabbatical(Smith,next year)), the evidence is predicted to be sufficient to change the user's belief in On-Sabbatical(Smith,next year), and hence Teaches(Smith,AI); thus, the focus of modification will be On-Sabbatical(Smith,next year). The Correct-Node specialization of Modify-Proposal will be invoked since the focus of modification is a belief, and in order to satisfy the precondition of Modify-Node (Figure), MB(S,U, On-Sabbatical(Smith,next year)) will be posted

as a mutual belief to be achieved. Since the user has a warranted belief in On-Sabbatical(Smith,next year) (indicated by the semantic form of utterance ()), the system will predict that merely informing the user of the intended mutual belief is not sufficient to change his belief; therefore it will select justification from the two available pieces of evidence supporting On-Sabbatical(Smith,next year) presented earlier. The system will predict that either piece of evidence combined with the proposed mutual belief is sufficient to change the user's belief; thus, the filtering heuristics are applied. The first heuristic will cause the system to select Postponed-Sabbatical(Smith, 1997) and supports(Postponed-Sabbatical(Smith, 1997),On-Sabbatical(Smith, next year)) as support, since it is the evidence in which the system is more confident. The system will try to establish the mutual beliefs as an attempt to satisfy the precondition of Modify-Node. This will cause the system to invoke Inform discourse actions to generate the following utterances: #1 Dr. Smith is not going on sabbatical next year. He postponed his sabbatical until 1997. If the user accepts the system's utterances, thus satisfying the precondition that the conflict be resolved, Modify-Node can be performed and changes made to the original proposed beliefs. Otherwise, the user may propose modifications to the system's proposed modifications, resulting in an embedded negotiation subdialogue.

Conclusion

This paper has presented a computational strategy for engaging in collaborative negotiation to square away conflicts in agents' beliefs. The model captures features specific to collaborative negotiation. It also supports effective and efficient dialogues by identifying the focus of modification based on its predicted success in resolving the conflict about the top-level belief and by using heuristics motivated by research in social psychology to select a set of evidence to justify the proposed modification of beliefs. Furthermore, by capturing collaborative negotiation in a cycle of Propose-Evaluate-Modify actions, the evaluation and modification processes can be applied recursively to capture embedded negotiation subdialogues.