Text Summarization Using Semantic Technique

by

Nguyen Thi Linh Chi

Dissertation submitted in partial fulfillment of

the requirements for the

Bachelor of Technology (Hons)

(Business Information Systems)

JANUARY 2011

Universiti Teknologi PETRONAS
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Ridzuan
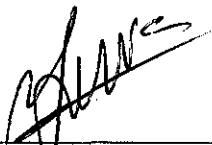
CERTIFICATION OF APPROVAL

Text Summarization Using Semantic Technique

by

Nguyen Thi Linh Chi

A project dissertation submitted to the
Business Information System Programme
Universiti Teknologi PETRONAS
in partial fulfillment of the requirement for the
BACHELOR OF TEHNOLOGY (Hons)
(BUSINESS INFORMATION SYSTEM)

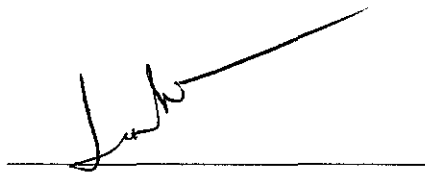Approved by,

_____
(Yew Kwang Hooi)

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

January 2011

2

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

NGUYEN THI LINH CHI

ABSTRACT

This work proposes semantic technique as a new approach for text summarization of online news/ journal articles. This text summarization project contains two part: parsing and semantic analysis. At first, for the parsing, we set up criteria to evaluate importance of sentences within text such as position, length. According to this, only sentences with high score of importance will be selected. The number of sentence selected depends on how much compact that users expect summary output would be. System combines those sentences into summary draft. Second part is semantic analysis; the technique we use here is lexical semantic approach. During this part, system will use Word Net lexical database to analyze words within summary draft. This database had words linked through semantic relations such as synonym, antonym, hyponymy, and more. Our project would end up by evaluation process. We evaluate accuracy of summary output, also compare and find differences between human summary work and system result.

ACKNOWLEDGEMENT

CONTENTS

LIST OF FIGURES

ABBREVIATIONS AND NOMENCLATURES

IR – Information Retrieval

KWIC – Keyword in Context

CHAPTER 1

INTRODUCTION

1.1. Background

Rapid growth of Internet increased the speed of information dissemination. Internet created alternatives to save and distribute documents. Documents of all time can be stored safely and firmly in huge databases forever. Then it becomes more convenient and time-saving for researcher to look for information from Internet than from books in library or book stores. However, exiting of so many available information sources on the Internet may get users confused of how to the right information that they want. Automatic text summarization is a useful tool to provide users an effective and efficient way to find the required information.

Inception of text summarization is 1950s by Luhn. He proposed an idea of summarization by finding the *most frequent words* in the document. On the other hand, Edmunson (1969) thought that *cue phrases* must be the key to extract information from original documents. Recently, Regina Barzilay and Michael Elhadad introduced another method to summarize document using lexical chains (1997). Mohamed Abdel Fattah and Fuji Ren carry out experiments with text features (2008)... All in all, the target of all automatic text summarization research conducted throughout last 60 years is to achieve the most-accurate-result method.

1.2. Problem Statement

So far, there is a lack of efficient algorithm to identify key points in online articles that may help better summarization.

Before the inception of this project, we reviewed a number of projects of predecessors and have some opinions: Though, findings of location and cue phrases of Edmunson gave better results than word frequency method of Luhn, his method is dependent on genre of text. There are some ideas of combining both Luhn and Edmunson heuristics to summarize text, however the most severe weakness of those theories are that they omitted relationship among words in text. To overcome this

matter, Regina et al. proposed an idea of Lexical cohesion using WordNet thesaurus. But, WordNet is not updated timely and sometimes causes error in identify verb as noun. Lexical chain approach also has other problems: long sentence will be set higher priority than others, unable to control length and level of details of outcome. Hence, we come up an idea of new more effective and efficient approach: Utilize semantic techniques to extract sentences from original texts and then use Lexical cohesion to find relationship among words and replace words with its more simple synonyms. With number of extracted sentences ranked from highest to lowest level of significance, system can select an appropriate number of outcome sentences to meet a given compression rate. Our project would include carrying out our new approach and evaluating accuracy of that method by comparing with human summary result.

1.3. Objectives

Our project's objective is using new text summarization technique – lexical semantic analysis to generate an accurate summary of online news/ journal articles. According to this, we would develop an excellent text summarization application which has specific domain output, meets user-desired compression rate, has grammar check and makes better sense of its meaning.

1.4. Scope of Study

Our project is to summarize online text which is from a single document (online articles/ journals), is not to summarize texts from multi-documents. Our project focuses on lexical semantic technique: We utilize WordNet thesaurus and user-built dictionary to create relationships among word within input text. Moreover, our project's scope includes text-feature analysis such as sentence position, keyword frequency, etc. Not only proposed a new technique to summarize text, our project also conduct evaluation to show how accurate our system works.

CHAPTER 2

LITERATURE REVIEW

This project's main function is to propose the semantic technique for text summarization of electronic information. Using this application, users will save time for reading online news/ journal articles but they still can find necessary information. A summary is implemented in either indicative way – to highlight some parts of passage or informative way – to contain all salient information (Joel et al, 2002).

Based on the form, Rejhan's team (2009) classified summarization into extracts and abstracts. For extracts summaries, all sentences and word sequences are got from original text. But in abstract ones, there are some word sequences not from the original text. Investigate summarization processes, they identify two approach level: surface level (this method uses shallow features such as: cue phrases, position of sentence in the text) and deeper level (generate sentence using semantic analysis technique). According to Reijhan's team, there are three main group of end-users: summary for all (generic summaries, summary for a limited domain (query-based summaries) and topic focused summaries. Summaries can also be divided into single-based and multi-document based. My text summarization project will use deeper level approach processing using semantic analysis technique and focus on single document base. The output of project is generic summaries which target a huge group of readers.

Agency Suspends Smallpox Vaccines for People With Heart Disease

Summary from the U.S.

A second health care worker has died of a heart attack (3) after receiving a smallpox vaccination (9) and officials are investigating whether vaccinations are to blame (3) for cardiac problems. (6) The vaccine never has been associated with heart trouble but as a precaution (3) the U.s. centers for Disease Control and Prevention (11) is advising people with a history of heart disease to be vaccinated (3) until further notice. (14) Strom suggested that the Bush administration reassess whether it necessary and safe to continue with its aggressive plan to inoculate millions of health care workers and emergency responders. (1)

Story keywords

vaccine, Heart, Smallpox, vaccinated, Disease

Source articles

1. Vaccination program in peril after second death (seattletimes.nwsource.com, 03/28/2003, 319 words)
2. Wired News: Smallpox Shots: Proceed With Care (Wired, 03/27/2003, 550 words)
3. 2nd worker dies after smallpox vaccination (msnbc.com, 03/28/2003, 358 words)
4. 2nd worker dies after smallpox vaccine (dallasnews.com, 03/28/2003, 499 words)
5. Smallpox vaccine is reviewed after second fatal heart attack (boston.com, 03/28/2003, 732 words)
6. Second Smallpox Vaccine Death Eyed (CBS News, 03/28/2003, 865 words)

Figure 1 - Text Summarization Example

2.1. Lexical Semantic Analysis

Semantic analysis is simply the study of meanings. There are some semantic techniques which are often used in developing text summarization as well as evaluating automatic summarization, for example: latent semantic analysis, and lexical semantic technique.

Latent semantic analysis example is text summarization project of Yihong Gong and Xin Liu in 2002. They used "singular value decomposition" to break given text into a number of linearly-independent base concepts. (Josef Steinberger et al 2004)
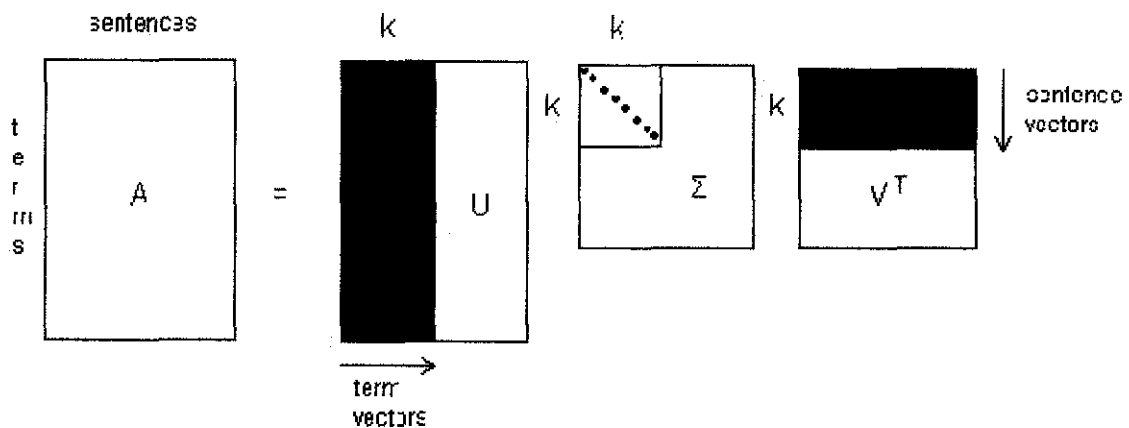
Figure 2 - Singular Value Decomposition

Earlier, lexical semantic techniques for corpus analysis are researched by a team of James Pustejovsky (Brandei University) Peter Anick (Digital Equipment Corporation), and Sabine Bergler (Concordia University) in 1993. They investigated "linguistic phenomena" for "semantic tagging of lexical chains" exploitation, and utilized semantic theory to construct deeper semantic word relations. They developed "qualia structure" which list all different aspect of a word: *constitutive, formal, telic, agentive*. The following is an example of qualia structure that James Pustejovsky team had given:

13

$$\begin{bmatrix} \text{book(x,y)} \\ \text{CONST} = \text{information(y)} \\ \text{FORMAL} = \text{physobj(x)} \\ \text{TELIC} = \text{read(T,w,y)} \\ \text{AGENTIVE} = \text{write(T,z,y)} \end{bmatrix}$$

where:

constitutive: the relation between an object and its constituent parts;

formal: that which distinguishes it within a larger domain;

telic: its purpose and function;

agentive: factors involved in its origin or "bringing it about"

(by James Pustejovsky et al)

Later, instead of creating matrix of terms and sentences and computing relationship among words, Princeton University has developed Word Net database. This database contains Senses in this database are represented sets of words sharing a common sense. There are many findings which has used semantic relations of words from Word Net, for example:

- Lexical chain as representation of context for the dectection and correction of malapropisms. (Hirst and St-Onge, 1998)

- A computational analysis of lexical cohesion with applications in information retrieval. (Stairmand, 1996).

- Using Lexical Chains for Text Summarization (Regina and Michael, 1997)

Our project is decided to use Word Net database as a tool to deploy lexical semantic technique for text summarization. In the following part, we would discuss more about lexical chains.

14

2.2. Sentence Extraction

There were two pioneers in using computer for information retrieval (IR): Luhn and Edmundson. Luhn was the manager of information retrieval research at IBM; he published his first book about information entitled "A new method of recording and searching information" in 1953. Luhn developed many computer-based IR applications. One of those is Keyword in Context (KWIC) system which includes three main elements: keywords, title, and context. According to Luhn (1958), frequency of keywords shows the salience of the passage. (Heting, 2007). Lately, finding keywords has become a crucial requirement for most of text summarization applications.

Besides, Edmundson was known for his paper (1969) regarding new methods in extracting sentences using three elements: cue words, title and headings, and structural indicators. (Rejhan Basagic et al 2009). In which, cue words use meta-linguistic markers to select important phrases: title, headings, first sentence and last sentence in the paragraph are important information to put in summary. (Regina Barzilay and Michael Elhadad, 1997).

Tradition techniques of Luhn and Edmundson nowadays get common in the field. In most text summarization project, the researchers used high-frequency words, phrases, position of sentences as basic criteria to score sentences within text. Recently, there are some groups who combine those techniques in their text summarization projects such as Kupiec et al 1995, Teufel & Moens 1997, Rejhan et al 2009.

However, Regina and Michael has indicated that the weakness of location and cue phrases abstraction is their dependence on the text genre. They gave one example that many rhetorical marks within a "Scientific American" articles change those articles into political articles. In their point of view, word frequency is a good indicator since words deliver important concepts. But, it's too "simplistic" as a source representation. Regina and Michael thought that word frequency method groups those high-frequency-word into the summary without considering links

among words. They believe that "adding information about word relations" would improve summary quality.

Hence, they proposed lexical chains heuristics for text summarization. Morris and Hirst 1991 were the first ones who used lexical chains in their NLP work. Lexical cohesion relations are used in form of categories, index entries and pointers in Roget's Thesaurus. According to Morris and Hirst, their work covered "90% of the intuitive lexical relations". But Regina and Michael had found out one limitation of their project that "they did not require the same word to appear with the same sense in its different occurrences for it to belong to a chain". In her opinion, this can cause confusions for semantically equivocal words. The team investigated and realized that most recently projects relating lexical chains are used WordNet lexical database in their algorithms. For example: projects of Hirst and St-Onge (1998), and
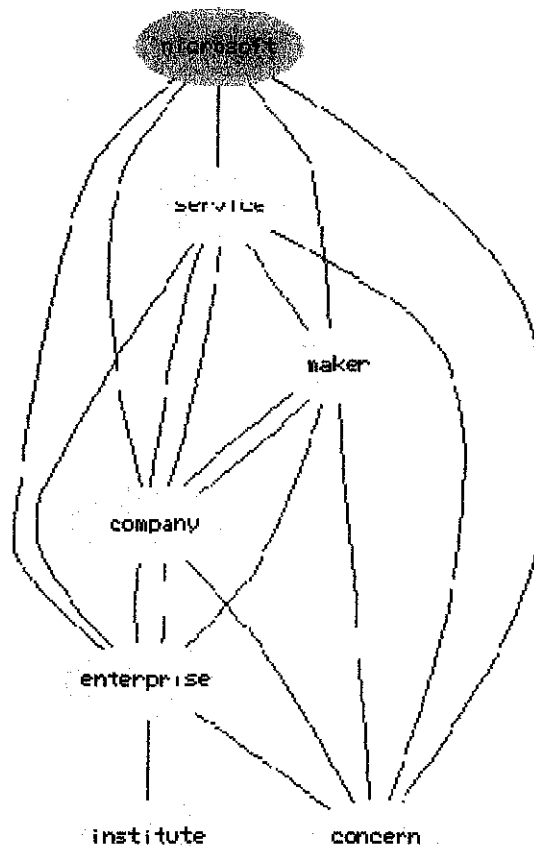
Figure 3 - Lexical Chains example

Stairmand (1996), etc. WordNet is an original core relations such as: synonymy,

polysemy, metonymy, hyponymy, meronymy, antonymy. Significantly, WordNet has some features that strongly support for NLP: glosses, links between semantically related words, group of similar terms, and more.

Here is an example when you search the word "water" in WordNet:

1. water, H2O – (binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid; freezes into ice below 0 degrees centigrade and boils above 100 degrees centigrade; widely used as a solvent)
2. body of water, water – (the part of the earth's surface covered with water (such as a river or lake or ocean); "they invaded our territorial waters"; "they were sitting by the water's edge")
3. water system, water supply, water – (facility that provides a source of water; "the town debated the purification of the water supply"; "first you have to cut off the water")
4. water – (once thought to be one of four elements composing the universe (Empedocles))
5. urine, piss, pee, piddle, weewee, water – (liquid excretory product; "there was blood in his urine"; "the child had to make water")
6. water – (a fluid necessary for the life of most animals and plants; "he asked for a drink of water")

Figure 4 - Synset Example

Regina introduced Hirst & St-Onge's procedure of building lexical chains using WordNet: Firstly, her team selects a set of candidate words. Secondly, find a chain for each candidate word so that each word in the chain related to the others. Thirdly, if the system found any word has relation to the words in the chain, system would insert that word into the chain then update the chain. However, they also indicates the weakness of this method that Hirst & St-Onge and Stairmand's algorithm are both rely on noun while WordNet sometimes makes mistake in identifying nouns. WordNet has another limitation is not updating new noun-compounds. Regina and Michael solved those problems by using "shallow parser" (created by Ido Dagan's team at Bar – Ilan University) and by "text segmentation"

17

(Hearst's algorithm 1994). While shallow parser is useful in identifying important concepts in the domain and removing unnecessary words, text segmentation helps reduce errors in finding noun keywords.

After constructing chains, Regina and Michael use empirical methodology to score chain. They presented score function as follows:

Score (chain) = Length * HomogeneityIndex

Where:

Length: The number of occurrences of members of the chain

HomogeneityIndex: 1 – the number of distinct occurrences divided by the length

Selected chains are those that meet "Strength Criterion":

Score (Chain) > Average (Scores) + 2 * StandardDeviation(Scores)

Finally, Regina & Michael extract full sentences from the input text based on select chains in three different ways: based on "first appearance of chain member in passage", based on "notion of representative words", and based on text units with "high density of chain members". Regina and Michael concluded that there is only one sentence is extracted for each chain.

Although Regina & Michael project have eliminated some weakness of predecessors' projects, their work still has some limitations: "sentence granularity", anaphora-sentence extraction, lack control of length and level of details of output.

In 2008, Bawakid and Oussalah has presented their "Semantic summarization system" using similar approach at Text Analysis Conference. Bawakid and Oussalah proposed scoring system static features (sentence position, "name entities" in each sentences) and dynamics features (semantic similarity between sentences and user query) as criteria to score sentences in the text. Bawakid & Oussalah text summarization procedures includes 3 steps:

18

1. Preprocessing: eliminate unnecessary elements.

2. Processing: use scoring system to score sentences in the text based on the above criteria.

3. Generation: choose the highest score sentences and arrange them in an appropriate sequence.

(Rejhan's team investigation).
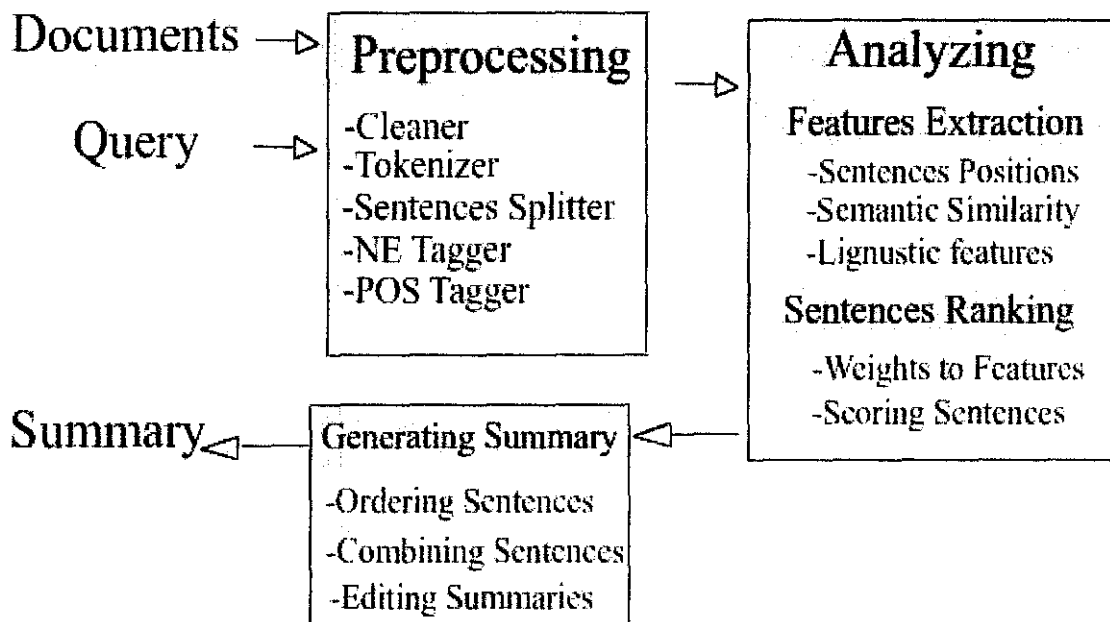
Figure 5 - Bawakid and Oussalah's Summarizer Architecture

It's said that this approach work more effectively by eliminating irrelevant contents in the text. Though the project still lacks of redundancy checking and redundancy remover features, we came to an inference that multi-feature approaches can improve the performance of text summarization systems.

Another example of using multi-feature approaches in automatic text summarization is the work of Mohamed Abdel Fattah and Fuji Ren (2008). Their system weighted 10 text features:

1. Sentence Position

2. Positive keyword in the sentence

3. Negative keyword in the sentence

4. Similarity of sentence with others

5. Similarity of sentence with title

6. Proper nouns in sentence

7. Numerical data in sentence

8. Relative length of sentence to overall article length

9. Bushy part of sentence

10. Aggregate similarity of sentence.

TABLE II
THE SUMMARIZATION PRECISION ASSOCIATED WITH EACH FEATURE FOR
DIFFERENT COMPRESSION RATES

Compression rate (CR)	10%	20%	30%
P(f1)	0.3365	0.3487	0.3424
P(f2)	0.3332	0.3318	0.3456
P(f3)	0.2897	0.2975	0.2998
P(f4)	0.3854	0.3993	0.4097
P(f5)	0.3587	0.3548	0.3790
P(f6)	0.3193	0.3285	0.3264
P(f7)	0.2612	0.2698	0.2665
P(f8)	0.2698	0.2652	0.2747
P(f9)	0.4153	0.4176	0.4283
P(f10)	0.3749	0.3682	0.3794

Figure 6 - The Summarization Precision Associated With Each Feature For Different Compression Rates

2.3. Compression Ratio

About compression ratio for text summarization system, Jade Goldstein et al conducted an investigation and inferred that "summary length was independent of document length, and that compression ratios became smaller with the longer documents."

20
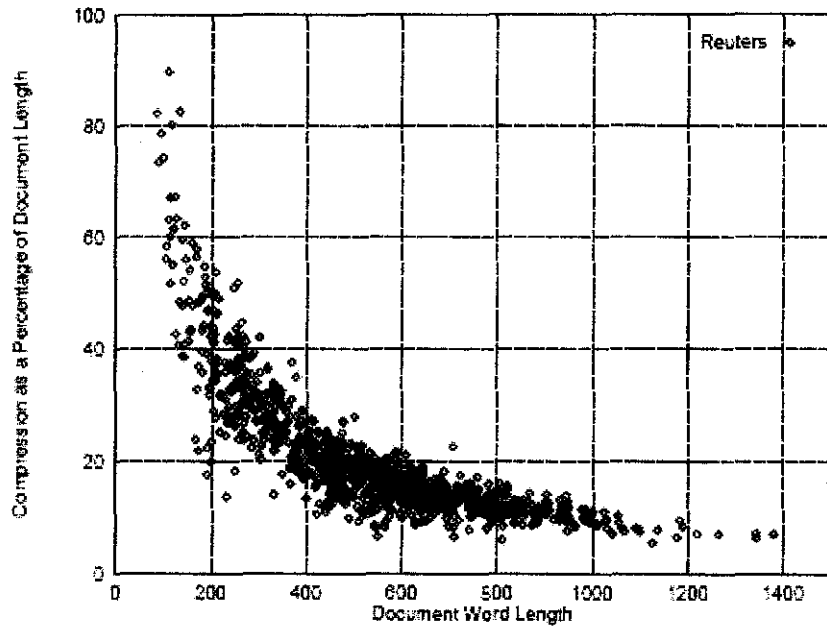
Figure 7 - Compression Ratio versus Document Word Length (Reuters)

The investigation covered 1,000 Reuter articles during the period from 10 – Nov – 2007 to 25 – Nov 1997. From the result showed in figure 6, Jade Goldstein's team inferred that "document compression ratio decreases as document word length increases."

21
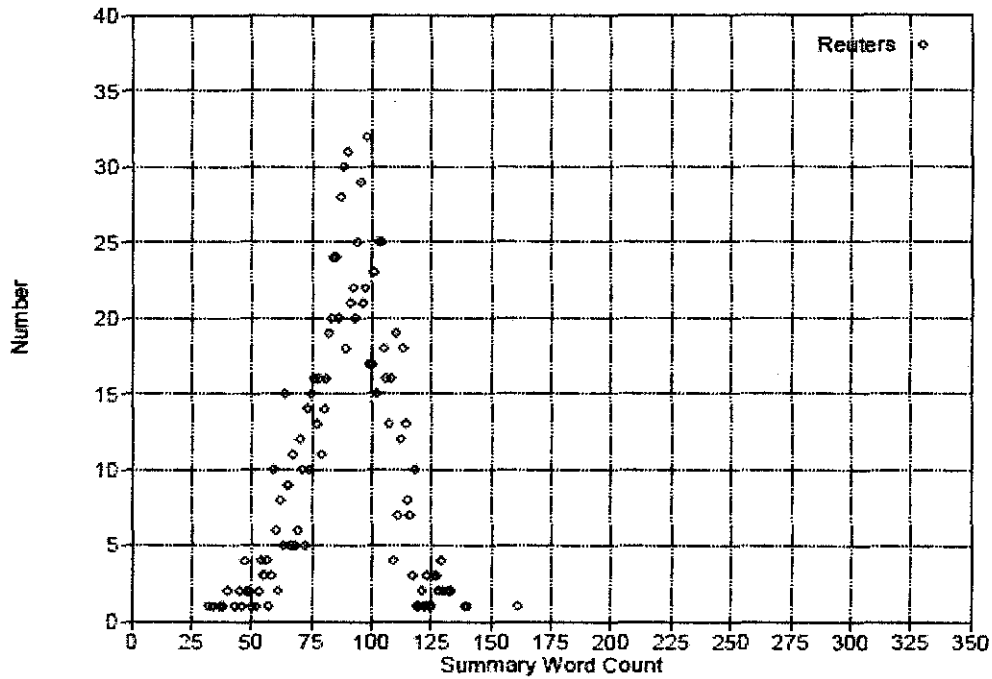
Figure 8 - Distribution of Summary Word Length (Reuters)

As the result displayed in Figure 7, no matter how long the original text is, its summary mostly has approximately 85 – 90 words.

2.4. Sentence Reduction

Hongyan Jing proposed a sentence reduction approach for text summarization systems. The algorithm Hongyan presented in the work including 5 steps:

1. Syntactic parsing: parse input sentences and develop sentence parse tree.

2. Grammar checking: identify significant components of sentence
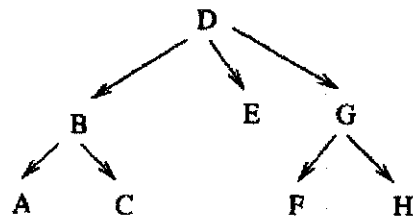
Input: A B C D E F G H

Figure 9 - Sample sentence and parse tree

22

3. Context information: find sentence resemble title

4. Corpus evidence: compare sentences reduced by human with original text

5. Final decision: reduce sentence based on previous steps' result

According to Hongyan's report, the project's average success rate is 81.3% .

CHAPTER 3

METHODOLOGY

3.1. Waterfall Development

According to literature view (Chapter2), we decided that our project would carry out the method that Mohamed Abdel group have used. Why we decided to use their scoring mechanism? It's because their formulas is the result of the combination of old and modern text summarization theories: from concepts of high frequency words (by Luhn) to ideas of using cue words, title and headings, structural indicators (by Edmundson) and then proper nouns, numeric data, etc. Their project was evaluated and proven of accuracy on 100 different English documents. According to that method, we used text features, which were referred to in their project, to score each sentence in the paragraph. Besides, the application provides a place for user to input categories of text (tag), this will be used to score sentence, too. Tags will help system to produce summary output in the same genre with text input. Hopefully, this feature would fix the problem of text-genre dependency before. Afterward, the system combine the highest score sentences then combine them orderly into summary. The number of selected sentences will vary depend on the compression rate of output that user desires to have. Besides, as regarded, our text summarization would use lexical semantic technique in which semantic is the study of meanings of the words, and lexical semantic is the study of relationship among words based on its meanings. Our system will use WorldNet database to replace words within given text with words or phrases in the database. By the end of this project, we would conduct evaluation process to investigate accuracy of developed system. Therefore, the methodology that we selected to develop this application is waterfall methodology.

This methodology involves planning, analysis, design, and implementation to be performed in a sequence.
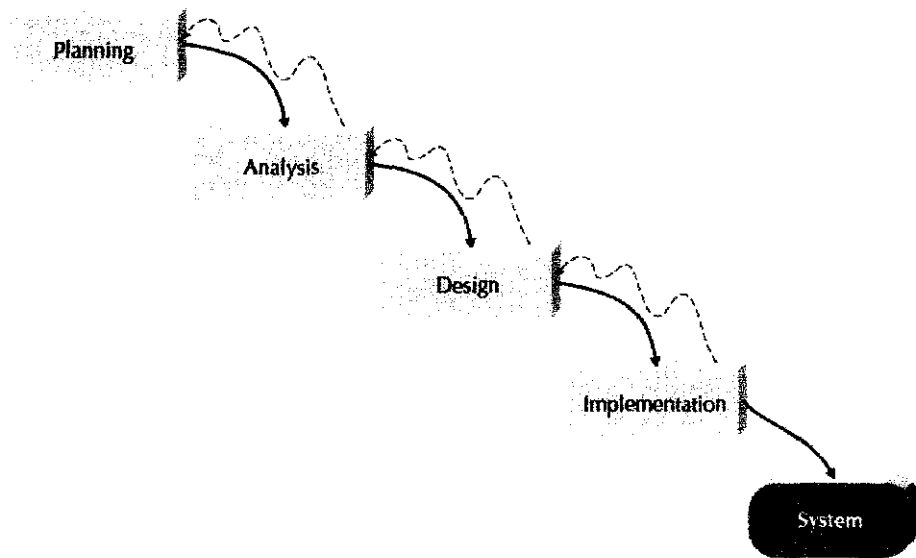
Figure 10 - A Waterfall Development-based Methodology

In our case, we have identified system requirements long before programming begins, and there would be no significant change to requirements as the project proceeds. That is why we chose waterfall methodology. Besides, this approach consumes less time so after accomplishing system development we can use few weeks to accuracy of output, and evaluate differences between human-made summary and automatic summary done by system. Finally, we want to hear feedback from end-users after using our system.

3.2. Requirements Analysis

- **Nonfunctional Requirements**
  - Operational Requirements
    - Operation System: Windows environment.
  - Performance Requirements
    - No special performance requirements are anticipated.
  - Security Requirements
    - No special security requirements are anticipated.
  - Cultural and Political Requirements
    - No special cultural and political requirements are anticipated.

- **Functional Requirements**
  - o Users can enter text that they want to summarize.
  - o Users can get summary output with compression rate that they decide.
  - o Summary output must be relevant with the genre of the text.

3.3.   Tools, mechanism, software used
In this project we will use:

- Visual Basic 2008 Express Edition: to develop application as well as
  database. This programming tools are chosen because of the following
  reasons:
  - o Cost: Microsoft supplies Visual Basic 2008 Express free of charge.
  - o Ease of use: VB.NET is an object-oriented language which is easy to
    learn, easy to use to develop complicated algorithms, and easy to
    modify when developers want to add more functions or fix the errors.
- SWOT Analysis: To analyze the Strengths and Weaknesses of the system as
  well as identify potential Opportunities and Threats to the system.
- UML diagrams: Activity, Use Case, Class, etc.
- Word Net Database: to linked words of the same category through semantic
  relations.
  - o Cost: free, because Word Net is an open-source project.
  - o Convenience: source code of Word Net is available on the website of
    Princeton University, other related projects are also published. So it's
    easier for us to work on.

CHAPTER 4

RESULTS AND DISCUSSION

4.1. The Framework of the System
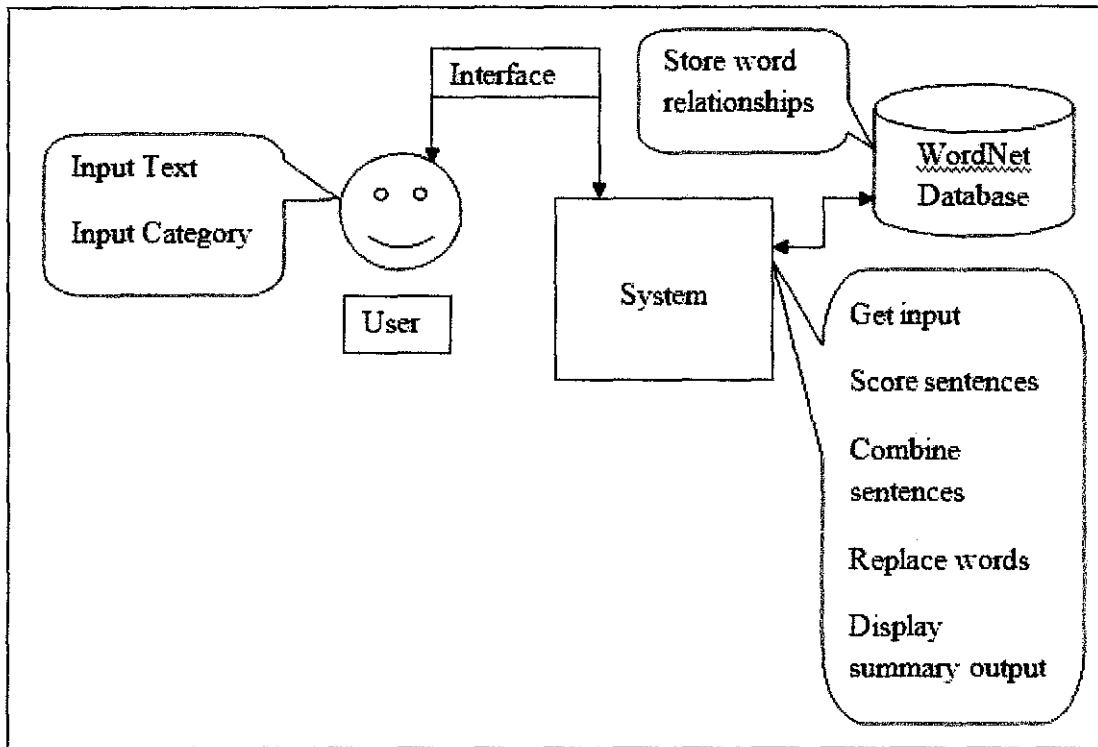
Figure 11 - Framework of the System

Where, the users include:

- Developers

- End-users who use the system to summarize text

The system will include:

- A friendly interface.

- Category function where users input category of text. This function will guide system to summarize text relevant to a proper genre. Category will be considered as a key word, and any sentences that have this key word will be added 1 in its score value. Otherwise, system will find other words within text have relationship with the category word/phrase (use Word Net database

to develop relationship among words). Sentences which have those words will increase 1 in its score value too.

- Compression rate – Users would decide what compression rate they expect summary output should be. Based on this rate, system will choose a proper number of sentences from text to put into summary draft.

- A score sentence mechanism to select key sentences to put into summary draft. Sentences within given text will be extracts and scored based on text features such as sentence position, keywords, and more.

- WordNet database to contain words and relationship among words such as synonyms, hyponyms, familiarity and more.

To emphasize one more time, our project is using semantic technique to enhance effectiveness of text summarization system. In which, semantic is the study of meaning of the words. We tend to identify lexical chains – relationship among words based on their meanings to find the key information from the given text. And here we chose WordNet database – a lexical database, this database contains trillions of words with their definition, part of speech (noun, verb, adjective, etc). Words which have same meaning will have same index number with the others. That's how we identify synset of each word. This will extend scope of keywords which helps system more easily and flexibly to find key information from sentences within text. Besides, synset provides synonyms of the words, so that system can find appropriate words to replace certain words within text. That's how semantic technique works on our system.

End-users will interact with the system through an interface by inputting **title, content, category** of text, **compression rate** percentage ($0 <$ rate <1). System will process input data: After getting text, system use scoring sentence mechanism to evaluate importance-level of sentences within text. Number of sentences selected is decided based on score of sentence, compression rate. System order selected sentences and combined into summary draft. Later, system access WordNet database to find synonyms for words within summary draft. Some words will be replaced properly before displaying summary output on the screen.

28

After extracting the whole text into array of sentences, scoring mechanism will assessed base on the following criteria (We used some certain formula proposed by Mohamed Abdel Fattah team, 2008. We chose their formulas as criteria because those are evaluated and proved with precision of results from 43-45% for compression rate = 10%, 20%, 30%. The results were tested on 100 English articles):

1. f1 = Sentence Position: We consider the first sentence and the last sentence of paragraph is the most important. Therefore if the sentence is the first or the last sentence of the paragraph, then plus 1 to score of that sentence.

2. f2 = Keyword in the sentence

$$f2 = \frac{Number\ of\ keywords\ within\ sentence}{Number\ of\ words\ within\ sentence}$$

Firstly, we extract each sentence into an array of word, and then count the number of keywords that appear within that sentence. f2 is retrieved by dividing number of keywords within sentence by number of words within that sentence. All sentences would be ranked based on f2 from highest to lower. Then we select a number of sentences based on compression ratio. For example, if compression ratio = 30%, then only 30% sentences with highest f2 are selected. Those sentences' score will be plus 1

3. f3 = sentence centrality.

$$f3 = \frac{Number\ of\ common\ keywords\ between\ sentence\ and\ others}{Number\ of\ total\ keywords\ within\ sentence}$$

At first, we find the number of common keywords between the sentence and other sentences within paragraph. Then we count the number of occurrences that common keywords appear within the sentence. Divide it by total number of keywords within that sentence, we have f3, then rank all the sentences within the paragraph from highest f3 value to one which has the lowest value. Select a

number of sentences which has highest f3 ratio based on compression ratio inputted by users, plus 1 to those sentences.

4. f4 = sentence inclusion of proper noun

$$f4 = \frac{Number\ of\ proper\ nouns\ within\ the\ sentence}{Number\ of\ words\ within\ sentence}$$

Find the number of proper nouns within each sentence. After that, we divide it by the number of total words within that sentence. Rank sentences in paragraph from those have highest f4 ratio to those which have lowest f4 ratio, and select the top highest f4 sentences, plus 1 to those sentences' score. The number of sentences selected also based on compression rate

5. f5 = sentence inclusion of numerical data

$$f5 = \frac{Number\ of\ numerical\ data\ within\ sentence}{Number\ of\ words\ within\ sentence}$$

Similar to f4, we find the number of numerical data within sentence, and then divide it by the number of words within that sentence. Again we rank the sentences top down based on f5 result of each sentence. Select those sentences which have highest f5 value, plus 1 to their scores. (Number selected is corresponding to compression rate)

6. f6 = Sentence relative length

$$f6 = \frac{Sentence\ length}{Text\ length}$$

After counting the number of words within each sentence, divide it by the number of words within text. Rank sentences from highest f6 to lower f6, then select a number = compression rate of sentences which have highest f6 value. We plus 1 to value of those sentences' scores.

7. f7 = Title resemblance

$$f7 = \frac{Keywords\ in\ the\ sentence \cap Keywords\ in\ the\ title}{Keywords\ in\ the\ the\ sentence \cup Keywords\ in\ the\ title}$$

At first, we have to find the number of common words between the sentence and the title. Then, find the number of words between sentence and title (remove redundant cases). Divide the first finding by the second one, we get f7. Repeat the ranking process, then we select sentences which have highest h7 value, and plus 1 to their scores.

In which, keywords are selected by the follow criteria:

1. Keywords are the words appeared in article's title / heading.
    Extract title into words, and add them into list of keywords.
2. User enter category which would be considered as a keyword
3. System use Word Net database to find other words within text which have relationship with certain keywords. Those words would also be considered as keywords.
4. After extracting text into words and removing redundancy, we count the number of occurrence of each word within the text. Divide that number by number of total words within text, if the quotient (p) is greater than (1 – compression rate), that word will be added into list of keywords.

$$p = \frac{Number\ of\ the\ sentence\ which\ has\ that\ words}{Number\ of\ sentences\ within\ text}$$

The score of each sentence is the sum of f1 – f7's values:

$$Score\ of\ sentence\ S(i) = f1(i) + f2(i) + f3(i) + f4(i) + f5(i) + f6(i) + f7(i)$$

After scoring sentence, a certain number of sentences with high score would be chosen. The number would depend on the compression rate that user had set:

$$No.\ of\ selected\ sentences = Compression\ rate \times No.\ of\ sentences\ in\ text$$

31

The selected sentences would be arranged into draft in a proper sequence as in given text. Refer to the synset database; replace some words in the draft with its synonyms. After replace, display output as results
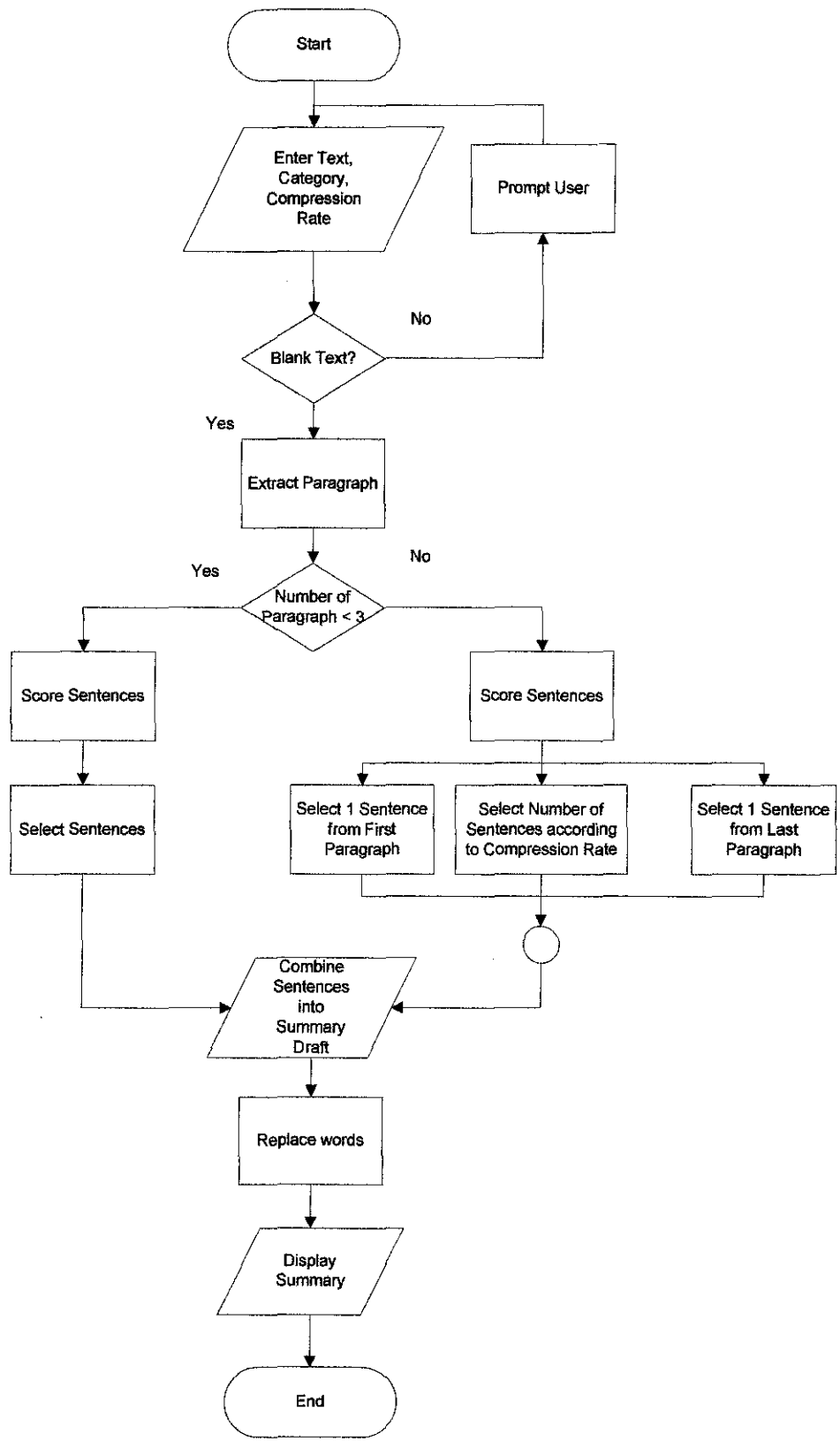
Figure 12 - System Flowchart

33
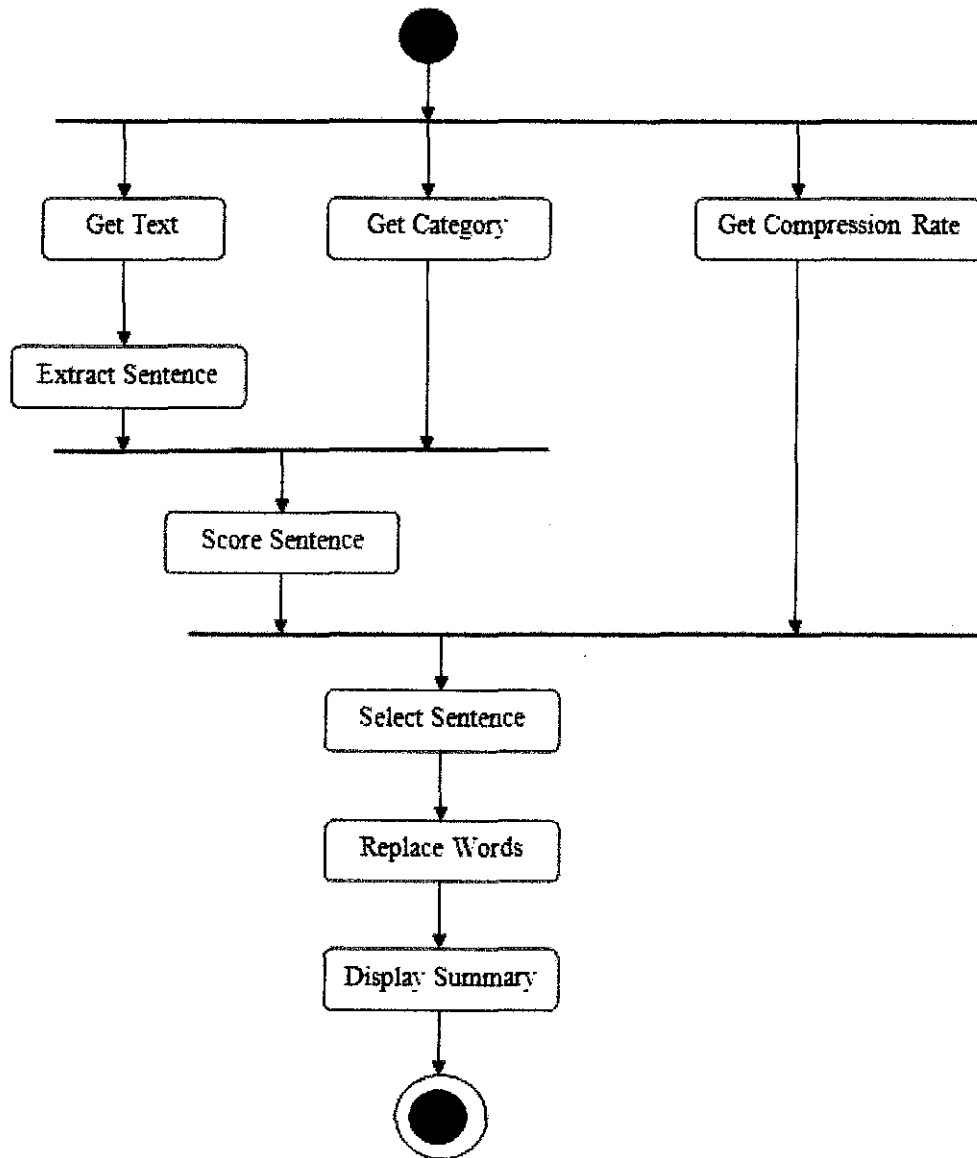
Figure 13 - Activity Diagram
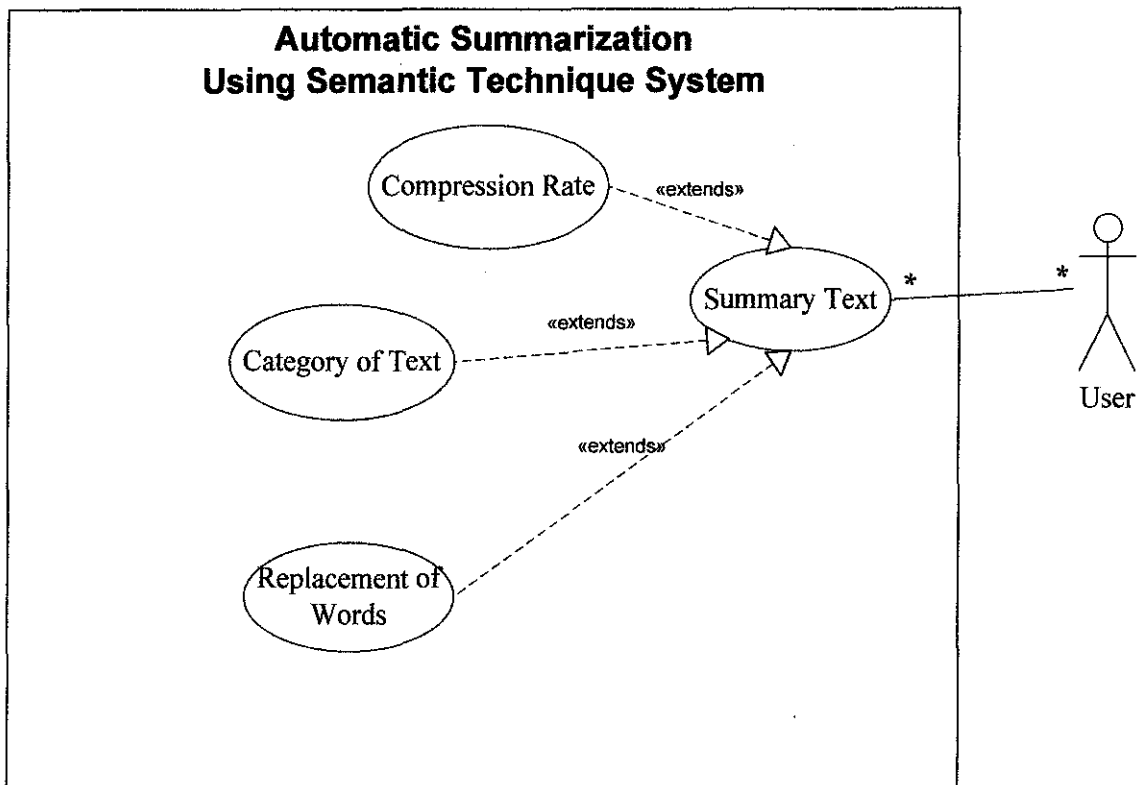
Figure 14 - Use Case Diagram

Note: The relationship between Summary Text and Compression Rate, Category of Text Replacement of Words is extends because: Compression Rate, Category are optional. If users does not enter compression rate, system will automatically select relevant sentences to put into summary draft. (As **Jade Goldstein** investigated and found out that each summary often have 85-90 words). In case user does not key in category of the text input, summary output would be the result of generic summarization process. Replacement of Words also is an extension of Summary Text since not all words from selected sentences will be replaced. Only some of them will be in order to reduce words and simplify complicated words.

Text

- Title
- Content
- Length
- Compression Rate
- Category
- No of Paragraph

1

0..*

Paragraph

- Position
- No of Sentence
- Length

1

0..*

Sentence

- Position
- Keyword
- Centrality
- Resemblance to Title
- ProperNoun
- Numeric Data
- Length

+ScoreSentence()()

0..*

1

Leads to ▶

Summary

- Sentence

+SelectSentence()()
+ArrangeSentence()()
+Update()()

1

0..*

Word

- WordType
- Plurality
- Centrality

+FindSynonym()()
+FindAntonym()()
+Replace()()

Figure 15 - Class Diagram

36

4.6. Results

Refer to Appendix B, we compared result from system with summary done by human (same compression rate), and analyzed:

	Similarity between System vs. Human 1	Similarity between System vs. Human 2
Paragraph 1	75%	50%
Paragraph 2	88%	50%
Paragraph 3	94%	94%
Paragraph 4	80%	50%
Paragraph 5	80%	83%
Paragraph 6	95%	69%
Paragraph 7	94%	93%
Paragraph 8	60%	75%
Paragraph 9	83%	83%
Paragraph 10	67%	50%
AVERAGE	**81.6%**	**73%**

Since human summary is based on human sense, so there's no fixed compression rate before summary. We had to compare between original work and human summary to identify compression rate. From that compression rate, we capture summary from our system with same title, content and compression rate.

After analyzing, we find that precise of the system work while comparing with human summary is always greater than 50%. Average for precise between system's summary and human's summary is 81.6% with the first person's work, and 73% with the second person's work. And 77.3% totally average when we compare system work with human summary.

Mostly, the system can capture key sentences from original work to put in summary. There are some differences because people who summarized changed the

grammar structure but still keep the meaning of the words, and sometimes they used their own words instead of using the word given in original text.

During implementation, we failed to link with WordNet 1.6. However, Mr. Troy Simpson, who has been studied on WordNet and achieved some result, provided us a data set which contains list of words with their corresponding part of speech, definition. Especially, words those have same meaning will have same index numbers so that we could identify the set of synonym for each current keyword and expanded scope of keywords. And we were still able to achieve the objective of this project: using semantic analysis.

Basically semantic analysis is the study of meaning of the words. For example, from this sample paragraph:

> *Today, knowledge is considered the most strategically important resource and learning the most strategically important capability for business organizations. However, many initiatives being undertaken to develop and exploit organizational knowledge are not explicitly linked to or framed by the organization's business strategy. In fact, most knowledge management initiatives are viewed primarily as information systems projects. While many managers intuitively believe that strategic advantage can come from knowing more than competitors, they are unable to explicitly articulate the link between knowledge and strategy. This article, using examples from several companies, provides a framework for making that link and for assessing an organization's competitive position regarding its intellectual resources and capabilities. It recommends that organizations perform a knowledge-based SWOT (strengths, weaknesses, opportunities, and threats) analysis, comparing their knowledge to that of their competitors and to the knowledge required to execute their own strategy. It provides a framework for describing the degree of aggressiveness of a knowledge strategy for closing strategic knowledge gaps, and concludes with several implications for competing on knowledge.*

According to the algorithm we elaborated above, system will filter text to find keywords. And we have the following list of keywords (Figure 16):
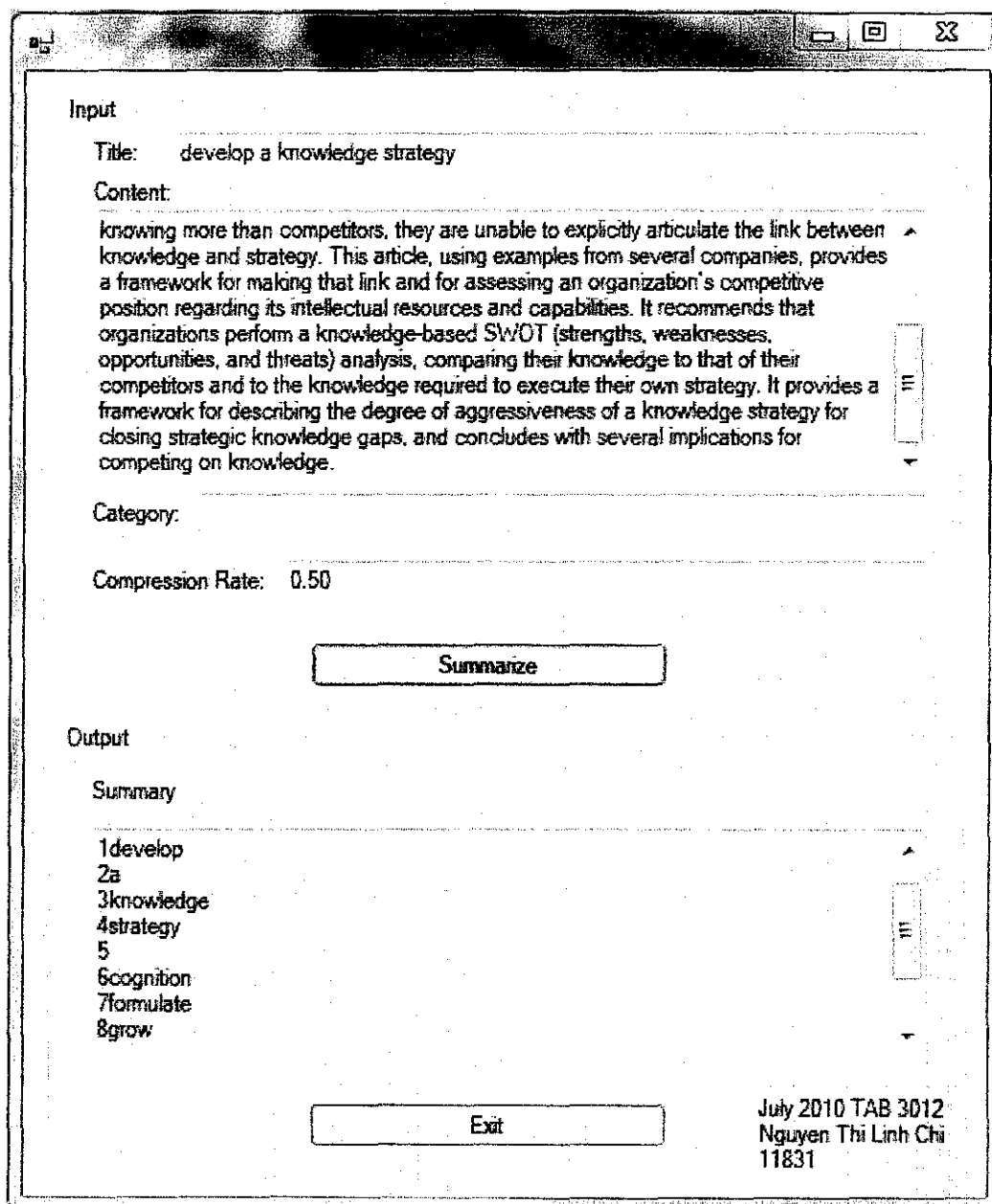
Figure 16 - Tested list of keywords

We have a cloud keyword:

Develop a knowledge strategy cognition formulate grow

In which the **knowledge** has the greatest number of occurrences – 11 times. **Strategy** was repeated 5 times. **Cognition** is the synonym of **knowledge**, however, there's no **cognition** within this text, same goes with **formulate** and **grow** – synonym of **develop**. Identifying the relationship among words helps system to identify key information more effectively. The following is result achieved from the above sample paragraph:
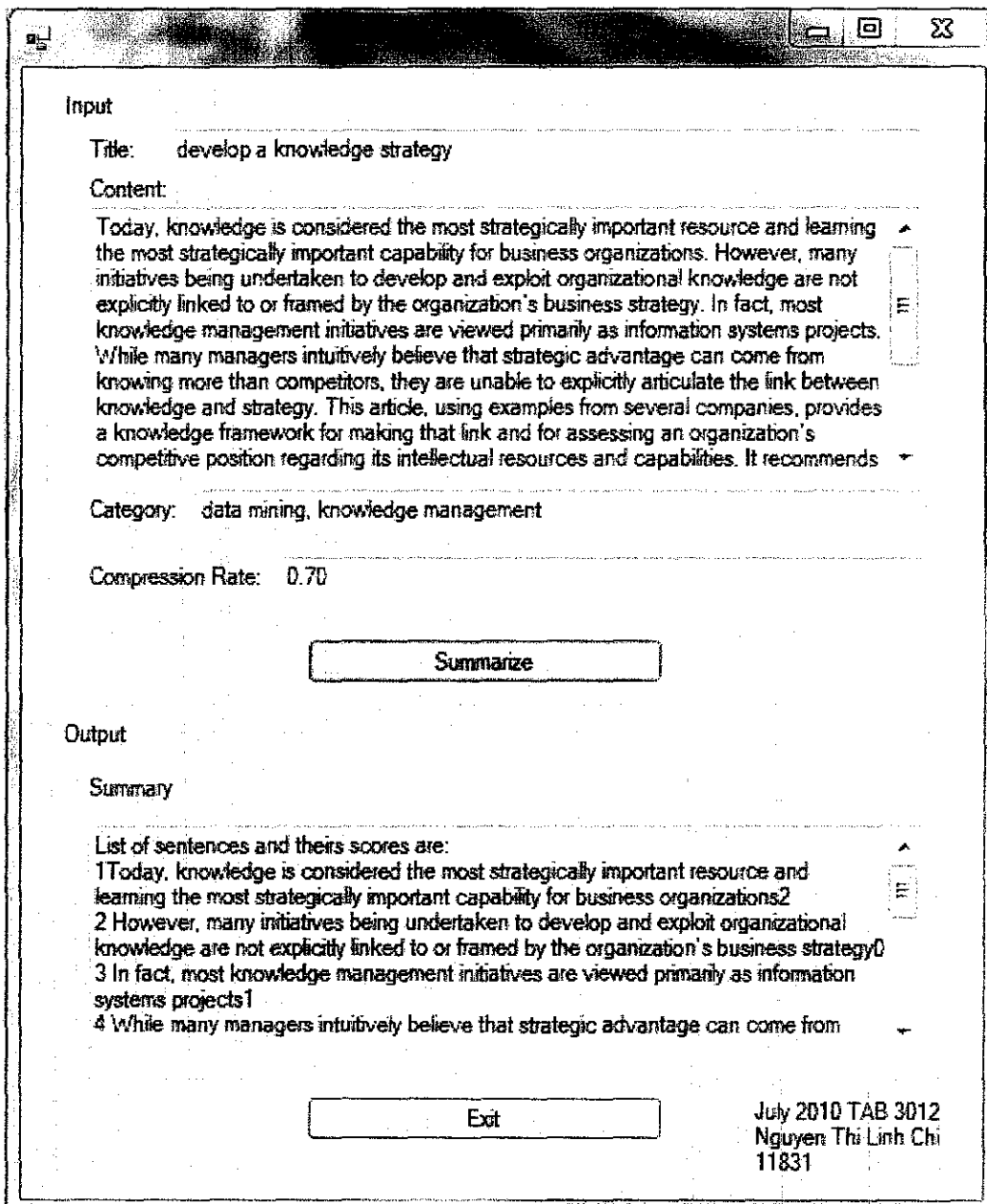
Figure 17 - Scoring sentences result

1. Summary done by other system (summary.com):

Today, knowledge is considered the most strategically important resource and learning the most strategically important capability for business organizations. It provides a framework for describing the degree of aggressiveness of a knowledge strategy for closing strategic knowledge gaps, and concludes with several implications for competing on knowledge. However, many initiatives being undertaken to develop and exploit organizational knowledge are not explicitly linked to or framed by the organization's business strategy.

2. Summary done by human (done by an English teacher in Vietnam):

This article introduced knowledge as an important resource for business organizations. And many initiatives were conducted to develop and exploit organizational knowledge. It recommends that organizations perform a knowledge-based SWOT analysis, comparing their knowledge to that of their competitors and to the knowledge required to execute their own strategy. It provides a framework for describing the degree of aggressiveness of a knowledge strategy for closing strategic knowledge gaps, and concludes with several implications for competing on knowledge.

3. Compared with the output get from system:

Today, knowledge is considered the most strategically important resource and learning the most strategically important capacity for business organizations. However, many initiatives being undertaken to develop and exploit organizational knowledge are not explicitly linked to or framed by the organization's business strategy. It provides a model for describing the stage of aggressiveness of a knowledge strategy for closing strategic knowledge gaps, and concludes with various implications for competing on knowledge.

41

From the result from human summary, our system summary and other system summary (summary.com), we compared and realized that our system's scoring mechanism helped to find 3 over 4 key sentences from the given paragraph to put into summary. In our system output, some words are replaced by its synonym such as *framework* replaced by *model, degree* replaced by *stage, several* replaced by *various.* We expect our system can do more in future work: increase the accuracy of finding key sentence; and not only replacing words, it will also help reduce unnecessary words such as *very, about,* and reduce phrases into one or two words such as *the most important* into *important.*

Figure 18 - Without synset database (compress 50%)

Text Summarization □ ▣ ✕

Input

Title: 3 Prescription Drugs That Do More Harm Than Good

Content:

True story: He had been a faculty member in three departments of a major university
with an IQ north of 180. Over time, the professor lost the ability to recognize people
he'd known closely for decades and to read more than a page of text at a time. He'd
repeat the same thing over and over, not recalling he'd already said it. The diagnosis:
rapidly progressive Alzheimer's. When he went to his 50th college reunion, he wore a
sign around his neck with his name and the statement, I have Alzheimer's. Old friends
needed an explanation for why he couldn't recognize people he'd known for decades
or repeated himself endlessly throughout the night.

Category: Cholesterol, Conditions, Health & Wellness, acid reducers, angiotensin, PPI, pre

Compression Rate: 0.20

[Summarize]

Output

Summary

Summary is:
True story: He had been a faculty member in three departments of a major university
with an IQ north of 180. Old friends needed an explanation for why he couldn't
recognize people he'd known for decades or repeated himself endlessly throughout the
night.

[Exit]

July 2010 TAB 3012
Nguyen Thi Linh Chi
11831

Figure 19 - Without synset database (compress 20%)

Input

Title: develop a knowledge strategy

Content:

Today, knowledge is considered the most strategically important resource and learning
the most strategically important capability for business organizations. However, many
initiatives being undertaken to develop and exploit organizational knowledge are not
explicitly linked to or framed by the organization's business strategy. In fact, most
knowledge management initiatives are viewed primarily as information systems projects.
While many managers intuitively believe that strategic advantage can come from
knowing more than competitors, they are unable to explicitly articulate the link between
knowledge and strategy. This article, using examples from several companies, provides
a framework for making that link and for assessing an organization's competitive
position regarding its intellectual resources and capabilities. It recommends that

Category: data mining, knowledge management

Compression Rate: 0.50

| Summarize |

Output

Summary

Summary is:
Today, knowledge is considered the most strategically important resource and learning
the most strategically important capacity for business organizations. However, many
initiatives being undertaken to develop and exploit organizational knowledge are not
explicitly linked to or framed by the organization's business strategy. It provides a
model for describing the stage of aggressiveness of a knowledge strategy for closing
strategic knowledge gaps, and concludes with various implications for competing on
knowledge. .

| Exit |

July 2010 TAB 3012
Nguyen Thi Linh Chi
11831

Figure 20 - with embedded synset (compress 50%)

For our future work, we want to develop our own database so that system can easily access in order to identify keywords. Besides, we realize that when people summarize, they not only change the words inside but also the grammar structure to reduce words and to simplify meanings. Hence, we want to do further research with English grammar rules, find a way so that system can perform all the rules to change grammar structure and words but still keep the same meaning of the content, and after all, the result will be more like natural language summary.

44

CHAPTER 5

CONCLUSION AND RECOMMENDATION

In conclusion, this work introduced a new technique to summarize text effectively. There are four things that we have accomplished: First of all, we used semantic database to enhance effectiveness of summarization. Database was used to create relationship among words and keywords. From that, it extended the scope of keywords in searching important points inside the given text. Besides, synset provided within the database helped to replace words before finalizing output summary. This would be very helpful for further future work, for sentence reduction and grammar checking. Secondly, we have done sentence extraction part by using **Mohamed Abdel Fattah and Fuji Ren**'s scoring sentences formulas. Inside their project, they proposed the scoring mechanism and showed the accuracy (43-45%) of those formulas in text summarization with 3 different compression ratios: 10%, 20%, and 30%, tested on 100 different English articles. Therefore, we chose their scoring sentences formulas to apply in our system, and those formulas worked effectively to identify correctly the key sentences from the given text inputted. Thirdly, compression rate was provided in the system which offers users a choice to determine how much they expect the summary will be. Summary's compression is based on the number of sentences. Fourthly, sentence reduction was done based on compression rate, and some words are replaced by its synset elements randomly. Generally, scoring sentences and identifying keywords mechanisms works successfully. The output was tested based on comparing statistics data: Keywords are identified exactly based on high frequency words, title/heading, category input, and semantic synset. Sentences are scored accurately based on scoring mechanism. From those, key sentences are selected for summary output. However, semantic technique we used so far in this project is just used semantic database to find the synset of each word, use that synset list to identify the keywords and to replace words. In the future, we would like to utilize other functions of the semantic database such as part of speech, hyponym list, antonym list and more for grammar checking and better sentence reduction (i.e: combine 2 sentences into one shorter sentences, reduce unnecessary words within the sentences).

REFERENCES

Regina, B. and Michael, E. 1997. *Using lexical chains for text summarization.* Ben Gurion University in the Negev Beer-Sheva, 84105 Israel.

Rejhan, B.; Damir, K. and Bojan, S.. 2009. *Automatic Text Summarization.* 706.048 Information Search and Retrieval. Graz University of Technology.

Bawakid, A. and Oussalah, M. 2008. A *Semantic Summarization System: University of Birmingham at TAC2008.*

Mohamed, A.F. and Fuji, R. 2008. *Automatic Text Summarization.* World Academy of Science, Engineering and Technology 37

Jade, G.; Mark, K.; Vibhu, M. and Jaime, C. 1999. *Summarizing Text Documents: Sentence Selection and Evaluation Metrics.* Annual ACM Conference on Research and Development in Information Retrieval. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 121-128

Hongyan, J. 2000. *Sentence reduction for automatic text summarization.* Applied Natural Language Conference. Proceedings of the sixth conference on Applied natural language processing. 310-315

Heting, C. 2007. *Information Representation and Retrieval in the Digital Age, 7-8.* Asist Monograph Series

Edmundson, H. 1969. New methods in automatic extracting. *Journal of the ACM* 16 (2): 264-285.

Hearst, M. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, 9-16. Las Cruces, New Mexico: Association for Computational Linguistics.

Hirst, G., and St-Onge, D. 1998. Lexical chains as representation of context for the detection and correction of malapropisms. In Fellbaum, C., ed., *WordNet: An Electronic Lexical Database and Some of its Applications.* Cambridge, MA: The MIT Press.

Luhn, H. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2 (2).

Morris, J., and Hirst, G. 1991. Lexical cohesion computed by thesaurus relations as an indicator of the structure of the text. *Computational Linguistics* 17(1): 21-43.

Stairmand, M. 1996. A *Computational Analysis of Lexical Cohesion with Applications in Information Retrieval.* Ph.D Dissertation, Center for Computational Linguistics UMIST, Manchester.

Teufel, S., and Moens, M. 1997. Sentence extraction as a classification task. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization, 58-65.* Madrid: Association for Computational Linguistics and the European Chapter of the Association for Computational Linguistics.

Kupiec, J.; Pedersen, J.; and Chen, F. 1995. A trainable document summarizer. In *Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 68 -73. Seattle, Washington: Special Interest Group on Information Retrieval.

Josef, S.; Karel, J. 2004. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. *In Proc. ISIM '04,* 93-100.

James, P.; Peter, A.; and Sabine, B. 1993. *MIT Press.*

GANTT CHART FOR FYP 1

TASK	DURATION	W1	W2	W3	W4	W5	W6	MID TERM BREAK	W7	W8	W9	W10	W11	W12	W13	W14
LITERATURE	2w	■	■													
Select topic	2w	■														
Write proposal	1w		■													
Topic selected	1w		■													
Preliminary report	2w			■	■											
Attend seminar	1d					■										
Collect Information	5w						■		■	■	■					
Prepare questionnaire	1w						■									
Conduct interviews	1w								■							
Combine results	3w									■	■					
Research	8w			■	■	■	■		■	■	■					
Search on Internet	8w			■	■	■	■		■	■	■					
Draft functions	2w					■	■									
Finalize functions	3w								■	■	■					
Progress report	1w										■					
Attend seminar	1d											■	■	■		
Design interface	1w										■					
Interim Final Report	1w														■	
Oral presentation	1d															■

GANTT CHART FOR FYP 2

TASK	DURATION	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
Coding (without WordNet)	7w	■	■	■	■	■	■	■							
Update reporting	7w	■	■	■	■	■	■	■							
Submit progress report	1d														
Test and Evaluation	3w							■	■	■					
Pre-EDX	1d										☺				
Dissertation	1d											■			
Prepare for Viva	2w												■	■	
Viva	4d													■	
Final Disertation	1w														■

1. Paragraph 1 – Develop a knowledge strategy by Michael Zack

LeaseCo, for example, recognized this opportunity by occasionally bidding aggressively on complex, novel or unpredictable lease opportunities (e.g., leasing personal computers in 1980) to gain unique and leverageable knowledge from those experiences, while attempting to prevent its competitors from gaining that same knowledge. LeaseCo realized a double benefit over its competitors, first by investing in its strategic knowledge platform and second, by learning enough about the particular client to competitively and profitably price leases for future opportunities with the same client. Often enough mutual learning occurred between LeaseCo and its client that the client contracted with LeaseCo for future leases without going out for competitive bids. In essence, LeaseCo created a sustainable (or renewable) knowledge-based barrier to competition. Lincoln Re, as part of its "experimental underwriting" process, similarly invested in its learning by insuring strategically selected novel and difficult classes of risk at favorable rates.

Human Summary 1: (75%)	Text Summarization using Semantic Analysis Technique: (75%)
LeaseCo recognized opportunity by occasionally bidding aggressively on complex lease opportunities to gain unique knowledge from those experiences, while preventing its competitors from gaining that same knowledge. LeaseCo realized double benefits over its competitors, first by investing in its strategic knowledge platform and second, by learning enough about the particular client to competitively and profitably price leases for future opportunities with the same client. Often mutual learning occurred, client contracted with LeaseCo for future leases without going out for competitive bids. In essence, LeaseCo created a sustainable knowledge-based barrier to competition.	LeaseCo, for example, recognized this opportunity by occasionally bidding aggressively on complex, novel or unpredictable lease opportunities (e. g. , leasing personal computers in 1980) to gain unique and leverageable knowledge from those experiences, while attempting to prevent its competitors from gaining that same knowledge. LeaseCo realized a double benefit over its competitors, first by investing in its strategic knowledge platform and second, by learning enough about the particular client to competitively and profitably price leases for future opportunities with the same client. Often enough mutual learning occurred between LeaseCo and its client that the client contracted with LeaseCo for future leases without going out for competitive bids. Lincoln Re, as part of its "experimental underwriting" process, similarly invested in its learning by insuring strategically selected novel and difficult classes of risk at favorable rates.
Human Summary 2: (25%)	Text Summarization using Semantic Analysis Technique: (25%)
LeaseCo created a sustainable (or renewable) knowledge-based barrier to competition meanwhile Lincoln Re invested in its learning by insuring strategically selected novel and difficult classes of risk at favorable rates.	LeaseCo, for example, recognized this opportunity by occasionally bidding aggressively on complex, novel or unpredictable lease opportunities (e. Lincoln Re, as part of its "experimental underwriting" process, similarly invested in its learning by insuring strategically selected novel and difficult classes of risk at favorable rates.

2. Paragraph2 – Develop a knowledge strategy by Michael Zack

Long lead time explains the attraction of strategic alliances and other forms of external ventures as potentially quicker means for gaining access to knowledge. It also explains why the strategic threat from technological discontinuity tends to come from firms outside of or peripheral to an industry. New entrants often enjoy a knowledge base different than that of incumbents, and which can be applied to the products and services of the industry under attack. This has been especially evident in industries where analog products are giving way to digital equivalents. For example, Image Corp. is experiencing a significant shift from physical film substrates to digital imaging. Its knowledge base is built on the science and technology of a physical consumable packaged good. Digital imaging, on the other hand, requires knowledge of computer systems and peripherals, imaging software, electronic distribution channels, and an economic model entirely different than for consumable physical products. The strategic challenge is to develop sufficient knowledge to support a shift to those new technologies and markets before non-traditional competitors make significant inroads in those markets, while not abandoning its years of experience and knowledge about physical imaging that is supporting its core business.

Human Summary 1: 37.5%	Text Summarization using Semantic Analysis Technique: 37.5%
Long lead time explains the attraction of strategic alliances is quick to gain access to knowledge. It also explains why the strategic threat from technological discontinuity tends to come from firms outside an industry. New entrants enjoy a knowledge base different than that of incumbents which can be applied to the products and services of the industry under attack. The strategic challenge is to develop sufficient knowledge to support a shift to those new technologies and markets before non-traditional competitors make significant inroads in those markets, while not abandoning its years of experience and knowledge about physical imaging that is supporting its core business.	Long lead time explains the attraction of strategic alliances and other forms of external ventures as potentially quicker means for gaining access to knowledge. It also explains why the strategic threat from technological discontinuity tends to come from firms outside of or peripheral to an industry. New entrants often enjoy a knowledge base different than that of incumbents, and which can be applied to the products and services of the industry under attack. The strategic challenge is to develop sufficient knowledge to support a shift to those new technologies and markets before non-traditional competitors make significant inroads in those markets, while not abandoning its years of experience and knowledge about physical imaging that is supporting its core business.
Human Summary 2: 25%	**Text Summarization using Semantic Analysis Technique: 25%**
Long lead time explains the attraction of strategic alliances and other forms of external ventures as potentially quicker means for gaining access to knowledge and why the strategic threat from technological discontinuity tends to come from firms outside of or peripheral to an industry.	Long lead time explains the attraction of strategic alliances and other forms of external ventures as potentially quicker means for gaining access to knowledge. The strategic challenge is to develop sufficient noesis to support a shift to those new technologies and markets before non-traditional competitors make significant inroads in those markets, while not abandoning its years of experience and noesis about physical imaging that is supporting its core business.

54

3. Paragraph 3 – Knowledge Audit by T Paramasivan

Knowledge management (KM) focuses on 'doing the right thing' instead of 'doing things right.' Knowledge management is a framework within which the organization views all its processes as knowledge processes. In this view, all business processes involve creation, dissemination, renewal, and application of knowledge toward organizational nourishment and survival. The concept of Knowledge Management embodies a transition from the recently popular concept of 'information value chain' to a 'knowledge value chain.' Whereas the information value chain considers technological systems as key components guiding the organization's business processes, while treating humans as relatively passive processors that implement 'best practices' archived in information databases. In contrast, the knowledge value chain treats human systems as key components that engage in continuous assessment of information archived in the technological systems. In this view, the human actors do not implement 'best practices' without active inquiry. Human actors engage in an active process of sense making to continuously assess the effectiveness of 'best practices.' The underlying premise is that 'best practices' of yesterday may not be taken for granted as 'best practices' of today or tomorrow. Hence, double loop learning, unlearning and relearning processes need to be designed into the organizational business processes.

Human Summary 1: 90%	Text Summarization using Semantic Analysis Technique: 90%
Knowledge management is a framework within which the organization views all its processes as knowledge processes. In this view, all business processes involve procedures toward organizational nourishment and survival. The concept of Knowledge Management is transition from the popular concept of 'information value chain' to a 'knowledge value chain.' Whereas the information value chain considers technological systems as key components guiding the organization's business processes, while treating humans as relatively passive processors that implement 'best practices' archived in information databases. In contrast, the knowledge value chain treats human systems as key components that engage in continuous assessment of information archived in the technological systems. In this view, the human actors do not implement 'best practices' without active inquiry. Human actors engage in an active process of sense making to continuously assess the effectiveness of 'best practices.' The underlying premise is that 'best practices' of yesterday may not be taken for granted as 'best practices' of today or tomorrow. Hence, double loop learning, unlearning and relearning processes need to be designed into the organizational business processes.	Knowledge management (KM) focuses on 'doing the right thing' instead of 'doing things right. ' Knowledge management is a framework within which the organization views all its processes as knowledge processes. In this view, all business processes involve creation, dissemination, renewal, and application of knowledge toward organizational nourishment and survival. The concept of knowledge Management embodies a transition from the recently popular concept of 'information value chain' to a 'knowledge value chain. ' Whereas the information value chain considers technological systems as key components guiding the organization's business processes, while treating humans as relatively passive processors that implement 'best practices' archived in information databases. In contrast, the knowledge value chain treats human systems as key components that engage in continuous assessment of information archived in the technological systems. Human actors engage in an active process of sense making to continuously assess the effectiveness of 'best practices. ' The underlying premise is that 'best practices' of yesterday may not be taken for granted as 'best practices' of nowadays or tomorrow. Hence, double loop learning, unlearning and relearning processes need to be designed into the organizational business processes.
Human Summary 2: 30% Knowledge management (KM) focuses on 'doing the right thing' instead of 'doing things right.' It is a framework within which the organization views all its processes as knowledge processes. Whereas the information value chain considers technological systems as key components guiding the organization's business processes	Text Summarization using Semantic Analysis Technique: 30% Knowledge management (KM) focuses on 'doing the right thing' instead of 'doing things right. ' Knowledge management is a framework within which the organization views all its processes as knowledge processes. Hence, double loop learning, unlearning and relearning processes need to be designed into the organizational business process

56

4. Paragraph 4 – Portfolio Investment Management by Open University Malaysia

In the construction of the capital market line (CML), we need to have the information on the market portfolio (M). As you've learned, a stock market index can serve as a proxy for this market portfolio. A stock market index is a number that indicates the relative level of prices or value of securities in a market on a particular day compared with a base-day figure, which is usually 100 or 1000. There are three main types of index, namely price-weighted indices, value weighted indices and equally weighted indices. You should learn more by working through the following readings on the construction of stock indices in the US as well as in Malaysia. The two commonly quoted stock market indices in the US are the Dow Jones Industrial Average (DJIA) and the Standard and Poor 500 (S&P500) indices. In Malaysia, the Kuala Lumpur Composite Index (KLCI) is the most important stock market index. The main objective of constructing market indices is to measure the performance of the relevant markets. By comparing the values of a market index over time, we can see how a market is performing over different periods. Technical analysts also use market indices to forecast the up and down trends of markets. They argue that these patterns of market index movements tend to repeat themselves. This kind of analysis requires a way to measure market performance. Besides measuring market performance, however, returns on market indices may be used as benchmarks to evaluate the performance of particular portfolios and mutual funds. Finally, market indices may be used to make comparisons on the performance and riskiness of various international markets, thereby providing information that can be used for international investments. We can try to find out, for example, which market has out-performed others. In fact, there are lots of financial futures instruments attached to market indices, such as the Hang Seng Index futures (Hong Kong), the Dow Jones Industrial Average Index futures (US), and the Nikkei 225 futures (Japan).

Human Summary 1: 60%	Text Summarization using Semantic Analysis Technique: 60%
In the construction of the capital market line (CML), we need to have the information on the market portfolio (M) which serves as a proxy for this market portfolio. A stock market index is a number that indicates the relative level of prices or value of securities in a market on a particular day compared with a base-day figure, which is usually 100 or 1000. There are three main types of index, namely price-weighted indices, value weighted indices and equally weighted indices. You should learn more by working through the following readings on the construction of stock indices in the US as well as in Malaysia. The two commonly quoted stock market indices in the US are the Dow Jones Industrial Average (DJIA) and the Standard and Poor 500 (S&P500) indices. In Malaysia, the Kuala Lumpur Composite Index (KLCI) is the most important stock market index. The main objective of constructing market indices is to measure the performance of the relevant markets. Technical analysts also use market indices to forecast the up and down trends of markets and argue that these patterns of market index movements tend to repeat themselves. This kind of analysis requires a way to measure market performance and returns on market indices which may evaluate the performance of particular portfolios and mutual funds. Finally, market indices may be used to make comparisons on the performance and riskiness of various international markets, thereby providing information that can be used for international investments.	In the construction of the capital market line (CML), we need to have the information on the market portfolio (M). As you've learned, a stock market index can serve as a proxy for this market portfolio. A stock market index is a number that indicates the relative degree of prices or value of securities in a market on a particular day compared with a base-day figure, which is usually 100 or 1000. There are three main types of index, namely price-weighted indices, value weighted indices and equally weighted indices. You should learn more by working through the following readings on the construction of stock indices in the US as well as in Malaysia. The two commonly quoted stock market indices in the US are the Dow Jones Industrial Average (DJIA) and the Standard and Poor 500 (S&P500) indices. In Malaysia, the Kuala Lumpur Composite Index (KLCI) is the most crucial stock market index. The main objective of constructing market indices is to measure the performance of the relevant markets. By comparing the values of a market index over time, we can see how a market is performing over different periods. In fact, there are lots of financial futures instruments attached to market indices, such as the Hang Seng Index futures (Hong Kong), the Dow Jones Industrial Average Index futures (US), and the Nikkei 225 futures (Japan).
Human Summary 2: 25%	Text Summarization using Semantic Analysis Technique: 25%
A stock market index can serve as a proxy for this market portfolio. There are three main types of index, namely price-weighted indices, value weighted indices and equally weighted indices. The main objective of constructing market indices is to measure the performance of the relevant markets, benchmarks to evaluate the performance of particular portfolios and mutual funds and to make comparisons on the performance and	In the construction of the capital market line (CML), we need to have the information on the market portfolio (M). As you've learned, a stock market index can serve as a proxy for this market portfolio. A stock market index is a number that indicates the relative degree of prices or value of securities in a market on a particular day compared with a base-day figure, which is usually 100 or 1000. In fact, there are lots of financial

riskiness of various international markets.	futures instruments attached to market indices, such as the Hang Seng Index futures (Hong Kong), the Dow Jones Industrial Average Index futures (US), and the Nikkei 225 futures (Japan).

5. Paragraph 5 – The synergy of social network analysis and knowledge mapping: a case study by Chan and Liebowitz

One of the key advantages of knowledge maps is to increase the visibility of knowledge sources and thus facilitate and accelerate the process of locating relevant expertise or experience in an organisation. They also provide a systematic context for the retrieval of reference information (Eppler, 2001). However, knowledge maps do not provide a systematic way to access the efficiency of the knowledge flows. SNA complements such weakness by providing an important means of analysing knowledge flows systematically. SNA has made significant contributions to a variety of fields including sociology, social psychology, anthropology, epidemiology, management studies (Cross, 2004), and terrorist network studies (Kleiner, 2002). Application of the SNA technique to knowledge mapping is relatively new. SNA focuses on analysing the relationships (ties) among the employees (actors) in terms of knowledge acquisition. For example, questions such as 'who do you ask for technical advice?', are used to determine the relationships among actors. SNA is helpful in identifying the strengths and inefficiencies in knowledge flows. It makes the invisible network of relationships between people seem more visible and thus gives valuable inputs to the managers to make decisions for improving the performance of their organisations (Krackhardt and Hanson, 1993; Cross et al., 2003). Some researchers (Cross et al., 2002) noted from their experiences, the need, to be cautious about overcorrecting the networks. In-depth analysis is required first to determine whether the 'problems' revealed are real and then to develop effective 'cures'. Cross et al. (2002) introduced a set of procedures of performing SNA in organisations: picking the right group; asking the right question; designing the survey; collecting the data; analysing the data; and sharing the results.

Human Summary 1: 83%	Text Summarization using Semantic Analysis Technique: 83%
One of the key advantages of knowledge maps is to increase the visibility of knowledge sources and thus facilitate and accelerate the process of locating relevant expertise or experience in an organisation. They also provide a systematic context for the retrieval of reference information (Eppler, 2001). However, knowledge maps do not provide a systematic way to access the efficiency of the knowledge flows. SNA complements such weakness by providing an important means of analysing knowledge flows systematically. Application of the SNA technique to knowledge mapping is relatively new. SNA focuses on analysing the relationships (ties) among the employees (actors) in terms of knowledge acquisition. SNA is helpful in identifying the strengths and inefficiencies in knowledge flows. It makes the invisible network of relationships between people seem more visible and thus gives valuable inputs to the managers to make decisions for improving the performance of their organisations (Krackhardt and Hanson, 1993; Cross et al., 2003). Some researchers (Cross et al., 2002) noted from their experiences, the need, to be cautious about overcorrecting the networks. In-depth analysis is required first to determine whether the 'problems' revealed are real and then to develop effective 'cures'.	One of the key advantages of knowledge maps is to increase the visibility of knowledge sources and thus facilitate and accelerate the process of locating relevant expertise or experience in an organisation. They also provide a systematic context for the retrieval of reference information (Eppler, 2001). However, knowledge maps execute not provide a systematic way to access the efficiency of the knowledge flows. SNA complements such weakness by providing an important means of analysing knowledge flows systematically. SNA has made important contributions to a variety of fields including sociology, social psychology, anthropology, epidemiology, management studies (Cross, 2004), and terrorist network studies (Kleiner, 2002). Application of the SNA technique to cognition mapping is relatively new. SNA focuses on analysing the relationships (ties) among the employees (actors) in terms of cognition acquisition. For example, questions such as 'who execute you ask for technical advice?', are used to determine the relationships among actors. SNA is helpful in identifying the strengths and inefficiencies in cognition flows. It makes the invisible network of relationships between people seem more visible and thus gives valuable inputs to the managers to make decisions for improving the performance of their organisations (Krackhardt and Hanson, 1993; Cross et al. Some researchers (Cross et al. In-depth analysis is required first to determine whether the 'problems' revealed are real and then to develop effective 'cures'. Cross et al. (2002) introduced a set of procedures of performing SNA in organisations: picking the right group; asking the right question; designing the survey; collecting the data; analysing the data; and sharing the results.
Human Summary 2: 50%	Text Summarization using Semantic Analysis Technique: 50%
Key advantages of knowledge maps is to increase the visibility of knowledge sources and thus facilitate and accelerate the process of locating relevant expertise or experience in an organization	One of the key advantages of cognition maps is to increase the visibility of cognition sources and thus facilitate and accelerate the process of locating relevant expertise or experience in an

60

as well as provide a systematic context for the retrieval of reference information. However, it do not provide a systematic way to access the efficiency of the knowledge flows. SNA complements such weakness by providing an important means of analysing knowledge flows systematically by focusing on analysing the relationships (ties) among the employees (actors) in terms of knowledge acquisition. It makes the invisible network of relationships between people seem more visible and thus gives valuable inputs to the managers to make decisions for improving the performance of their organizations. Some researchers noted from their experiences, the need, to be cautious about overcorrecting the networks.	organisation. They also provide a systematic context for the retrieval of reference information (Eppler, 2001). However, cognition maps do not provide a systematic way to access the efficiency of the cognition flows. SNA complements such weakness by providing a crucial means of analysing cognition flows systematically. SNA has made important contributions to a variety of fields including sociology, social psychology, anthropology, epidemiology, management studies (Cross, 2004), and terrorist network studies (Kleiner, 2002). Application of the SNA technique to cognition mapping is relatively new. , 2002) noted from their experiences, the need, to be cautious about overcorrecting the networks. (2002) introduced a set of procedures of performing SNA in organisations: picking the right group; asking the right question; designing the survey; collecting the data; analysing the data; and sharing the results.

6. Paragraph 6 – Essentials of Entrepreneurship and Small Business Management by Thomas W. Zimmer, Norman M. Scarborough

These three forces—consumerism, indebtedness, and excessive taxation— are largely taken for granted by most Americans, and their casual attitude towards such powers leads to victimization. They know there must be something wrong with their impulsive spending habits, but they have not yet linked those habits to their inability to tune out the media hype that urges them to consume, at any cost. These same people long to have more money for retirement, for their children's education, for vacations, and yet they realize they're not saving anything. Unfortunately, they have not yet seen the correlation between their enslavement to credit issuers and their inability to save for the future. These people feel overwhelmed by the amount of taxes extracted from their paycheck each month, and by the way that estate and death taxes eat into their savings and investment nest eggs, but have not yet connected big government shearing with their own ignorance about the way the tax system really works. These people see tremendous financial opportunity in the world but lack the skills necessary to control these powerful forces and harness the wealth and prosperity all around them.

Human Summary 1: 67%	Text Summarization using Semantic Analysis Technique: 67%
Consumerism, indebtedness, and excessive taxation are largely taken for granted by most Americans, and their casual attitude towards such powers leads to victimization. They know there is something wrong with their impulsive spending habits, but they have not yet linked those habits to their inability to tune out the media hype that urges them to consume, at any cost. These people feel overwhelmed by the amount of taxes extracted from their paycheck each month, and by the way that estate and death taxes eat into their savings and investment nest eggs, but have not yet connected big government shearing with their own ignorance about the way the tax system really works. These people see tremendous financial opportunity in the world but lack the skills necessary to control these powerful forces and harness the wealth and prosperity all around them.	These three forces—consumerism, indebtedness, and excessive taxation— are largely taken for granted by most Americans, and their casual attitude towards such powers leads to victimization. They know there must be something wrong with their impulsive spending habits, but they have not yet linked those habits to their inability to tune out the media hype that urges them to consume, at any cost. These same people long to have more money for retirement, for their children's education, for vacations, and yet they realize they're not saving anything. Unfortunately, they have not yet seen the correlation between their enslavement to credit issuers and their inability to save for the future. These people see tremendous financial opportunity in the world but lack the skills necessary to control these powerful forces and harness the wealth and prosperity all around them.
Human Summary 2: 17% Consumerism, indebtedness, and excessive taxation are largely taken for granted by most Americans, and this leads to victimization.	**Text Summarization using Semantic Analysis Technique: 17%** These three forces—consumerism, indebtedness, and excessive taxation— are largely taken for granted by most Americans, and their casual attitude towards such powers leads to victimization.

7. Paragraph7 – Essentials of Entrepreneurship and Small Business Management by Thomas W. Zimmer, Norman M. Scarborough

One crucial concern of entrepreneurs and the potential lenders and investors who finance their companies is whether there is a real market for the proposed good or service. Every entrepreneur must therefore describe the company's target market and its characteristics. Defining the target market and its potential is one of the most important – and most challenging – parts of building a business plan. Creating a successful business depends on an entrepreneur's ability to attract real customers who are willing and able to spend real money to buy its products or services. Perhaps the worst marketing error and entrepreneur can commit is failing to define his or her target market and trying to make the business "everything to everybody". Small companies usually are much more successful when focusing on a specific market niche where they can excel at meeting customers' special needs or wants.

Human Summary 1: 83%	Text Summarization using Semantic Analysis Technique: 83%
One concern of entrepreneurs and the potential lenders and investors is whether there is a real market for the proposed goods or service. Every entrepreneur must therefore describe the company's target market and its characteristics. Defining the target market and its potential is one of the most important – and most challenging – parts of building a business plan. Creating a successful business depends on an entrepreneur's ability to attract real customers. Perhaps the worst marketing error and entrepreneur can commit is failing to define his or her target market. Small companies usually are much more successful when focusing on a specific market niche where they can excel at meeting customers' special needs or wants.	One significant concern of entrepreneurs and the potential lenders and investors who finance their companies is whether there is a real market for the proposed good or service. Every entrepreneur must therefore describe the company's target market and its characteristics. Defining the target market and its potential is one of the most significant – and most challenging – parts of building a business plan. Creating a successful business depends on an entrepreneur's ability to attract real customers who are willing and able to spend real money to buy its products or services. Perhaps the worst marketing error and entrepreneur can commit is failing to define his or her target market and trying to make the business "everything to everybody". Small companies usually are much more successful when focusing on a specific market niche where they can excel at meeting customers' special needs or wants.
Human Summary 2: 40%	Text Summarization using Semantic Analysis Technique: 40%
Concern of entrepreneurs and the potential lenders and investors who finance their companies is whether there is a real market for the proposed good or service. They must therefore describe the company's target market and its characteristics.	One crucial concern of entrepreneurs and the potential lenders and investors who finance their companies is whether there is a real market for the proposed good or service. Every entrepreneur must therefore describe the company's target market and its characteristics.

.

8. Paragraph 8 – Essentials of Entrepreneurship and Small Business Management by Thomas W. Zimmer, Norman M. Scarborough

While renovating their top-floor apartment in Stockholm, Sweden, civil engineers Hakan and Annika Olsson came up with a unique idea for creating high-quality modular penthouses that could be manufactured in factories and installed atop existing flat-roof buildings. When the couple moved to London, they purchased aerial photographs of the city and marked all of the flat-roof buildings in red ink. "We knew we had a good business idea when the whole picture was red," says Hakan. After conducting more research and building a business plan, the Olssons launched First Penthouse, a company specializing in rooftop development. Their business model adds value both for tenants, who get ritzy penthouse living quarters where none existed before, and for landlords, whose property values are enhanced by the addition of the modular penthouses. First Penthouse offers the convenience of one-day installation of its penthouses and guarantees no disturbances to existing residents. Like most entrepreneurs, the Olssons had to overcome obstacles, including banks that were hesitant to extend credit "because the idea was so new," says Hakan. (To get the capital they needed, the Olssons used angel financing, when they convinced a wealthy friend to put up most of the $400,000 they needed to create and install the first penthouse. To convince balking regulators, the Olssons agreed to use special "quiet" tools and to place soundproof mats over the roofs on which they worked. Sales of the company's penthouses are growing, and the Olssons are planning to take their concept into other urban markets around the world, including New York City.

Human Summary 1: 75%	Text Summarization using Semantic Analysis Technique: 75%
Hakan and Annika Olsson came up with a unique idea for creating high-quality modular penthouses that could be manufactured in factories and installed atop existing flat-roof buildings. After conducting more research and building a business plan, the Olssons launched First Penthouse, a company specializing in rooftop development. Their business model adds value both for tenants, who get ritzy penthouse living quarters where none existed before, and for landlords, whose property values are enhanced by the addition of the modular penthouses. The Olssons agreed to use special "quiet" tools and to place soundproof mats over the roofs on which they worked. Sales of the company's penthouses are growing, and the Olssons are planning to take their concept into other urban markets around the world, including New York City.	Hakan and Annika Olsson came up with a unique idea for creating high-quality modular penthouses that could be manufactured in factories and installed atop existing flat-roof buildings. After conducting more research and building a business plan, the Olssons launched First Penthouse, a company specializing in rooftop development. They convinced a wealthy friend to put up most of the $400,000 to create and install the first penthouse. The Olssons agreed to use special "quiet" tools and to place soundproof mats over the roofs on which they worked. Sales of the company's penthouses are growing, and the Olssons are planning to take their concept into other urban markets around the world, including New York City.
Human Summary 2: 50%	Text Summarization using Semantic Analysis Technique: 50%
Civil engineers,Hakan and Annika Olsson came up with an idea for creating high-quality modular penthouses that could be manufactured in factories and installed atop existing flat-roof buildings while renovating their top-floor apartment in Stockholm. After conducting more research and building a business plan, the Olssons launched First Penthouse, a company specializing in rooftop development. They managed to convince a wealthy friend to put up most of the $400,000 as their capitals after failed to get sources from the bank. The company sales are growing, and they are planning to take their concept into other urban markets around the world.	Hakan and Annika Olsson came up with a unique idea for creating high-quality modular penthouses that could be manufactured in factories and installed atop existing flat-roof buildings. After conducting more research and building a business plan, the Olssons launched First Penthouse, a company specializing in rooftop development. Their business model adds value both for tenants, who get ritzy penthouse living quarters where none existed before, and for landlords, whose property values are enhanced by the addition of the modular penthouses. Sales of the company's penthouses are growing, and the Olssons are planning to take their concept into other urban markets around the world, including New York City.

9. Paragraph 9 – The Dipmeter Advisor System: A case study in commercial expert system development by Reid G. Smith and James D. Baker

Oil-well logs are made by lowering tools into the borehole and recording measurements made by the tools as they are raised to the surface. The resulting logs are sequences of values indexed by depth. Logging tools measure a variety of petrophysical properties. The dipmeter tool in particular measures the conductivity of rock in a number of directions around the borehole. Variations in conductivity can be correlated and combined with measurements of the inclination and orientation of the tool to estimate the magnitude and azimuth of the dip or tilt of various formation layers penetrated by the borehole Because the dipmeter tool has high resolution in the vertical direction (0.1-0.2 in.), it provides the petroleum geologist with detailed information on relatively fine-structured sedimentary beds. This type of information is invaluable in defining hydrocarbon reservoir structure and designing methods to drain such reservoirs. Knowledge of the dip variations as a function of depth in the vicinity of the borehole does not in itself identify geologic features. However, when combined with knowledge of local geology and rock properties measured by other logs (e.g., lithology [sand, shale ...]), the characteristic dip patterns (signatures) of geologic events in the depositional sequence can be interpreted. The right channel of Figure 2 is an interval of a dipmeter log. Dip estimates are shown as tadpoles. Dip magnitude increases to the right of the graph, and the down dip direction is indicated by the tail on each tadpole. The vertical axis is depth. (Hollow tadpoles indicate lower confidence dip estimates than solid tadpoles.) The left channel is a gamma ray log. (It measures natural gamma radiation in the formation-a rudimentary lithology indicator.) From this localized data, a skilled interpreter is often able to make comprehensive deductions about the geological history of deposition, the composition and structure of the beds, and the optimum locations for future wells.

Human Summary 1: 45%

Oil-well logs are made by lowering tools into the borehole and recording measurements made by the tools as they are raised to the surface. The resulting logs are sequences of values indexed by depth. The *dipmeter* tool measures the conductivity of rock in a number of directions around the borehole. Variations in conductivity can be correlated and combined with measurements of the inclination and orientation of the tool to estimate the magnitude and azimuth of the *dip* or tilt of various formation layers penetrated by the borehole. Knowledge of the dip variations as a function of depth in the vicinity of the borehole does not in itself identify geologic features. However, when combined with knowledge of local geology and rock properties measured by other logs, the characteristic dip patterns of geologic events in the depositional sequence can be interpreted.

Human Summary 2: 45%

Oil-well logs are made by lowering tools into the borehole and recording measurements made by the tools as they are raised to the surface. The resulting logs are sequences of values indexed by depth. The dipmeter tool measures the conductivity of rock in a number of directions around the borehole. It is invaluable in defining hydrocarbon reservoir structure and designing methods to drain such reservoirs hydrocarbon reservoir structure and designing methods to drain such reservoirs. Knowledge of the dip variations as a function of depth in the vicinity of the borehole does not in itself identify geologic features. However, when combined with knowledge of local geology and rock properties measured by other logs the characteristic dip patterns *(signatures)* of geologic events in the depositional sequence can be interpreted.From interval of a dipmeter log graph, a skilled interpreter is able to make comprehensive deductions about the geological history of deposition, the composition and structure of the beds, and the optimum locations for future wells

Text Summarization using Semantic Analysis Technique: 45%

Oil-well logs are made by lowering tools into the borehole and recording measurements made by the tools as they are raised to the surface. The resulting logs are sequences of values indexed by depth. Logging tools measure a variety of petrophysical properties. The dipmeter tool in particular measures the conductivity of rock in a number of directions around the borehole. Variations in conductivity can be correlated and combined with measurements of the inclination and orientation of the tool to estimate the magnitude and azimuth of the dip or tilt of various formation layers penetrated by the borehole Because the dipmeter tool has high resolution in the vertical direction (0. 1-0. 2 in.), it provides the petroleum geologist with detailed information on relatively fine-structured sedimentary beds. This type of information is invaluable in defining hydrocarbon reservoir structure and designing methods to drain such reservoirs. Knowledge of the dip variations as a function of depth in the vicinity of the borehole does not in itself identify geologic features.) From this localized data, a skilled interpreter is often able to make comprehensive deductions about the geological history of deposition, the composition and structure of the beds, and the optimum locations for future wells.

10. Paragraph 10 – What is sustainable development? From http://www.lbl.gov

Sustainable development is a new term that grew out of the conservation/environmental movement of the 1970's. While the conservation/environmental movement asked questions about preserving the Earth's resources, sustainable development includes questions about how human decisions affect the Earth's environment. At this moment, sustainable development means different things to different people/groups. The most widely held definition is that of the Brundtland Commission Report of 1987 which stated we must " meet the needs fo the present without compromising the ability of future generations to meet their own needs". In other words, when people make decisions about how to use the Earth's resources such as forests , water, minerals, gems, wildlife, etc., they must take into account not only how much of these resources they are using, what processess they used to get these resources., and who has access to these resources. Are enough resources going to be left for your grandchildren to use and will the environment be left as you know it today?

Human Summary 1: 50%	Text Summarization using Semantic Analysis Technique: 50%
Sustainable development is a new term that grew out of the conservation/environmental movement of the 1970's. While the movement asked questions about preserving the Earth's resources. However now most widely held definition is that of the Brundtland Commission Report of 1987 which stated we must rules and regulations.	Sustainable development is a new term that grew out of the conservation/environmental movement of the 1970's. While the conservation/environmental movement asked questions about preserving the Earth's resources, sustainable development includes questions about how human decisions affect the Earth's environment. At this moment, sustainable development means different things to different people/groups. , and who has access to these resources.
Human Summary 2: 33% Sustainable development includes questions about how human decisions affect the Earth's environment. The most widely held definition is they must meet the needs for the present without compromising the ability of future generations to meet their own needs.	Text Summarization using Semantic Analysis Technique: 33% Sustainable development is a new term that grew out of the conservation/environmental movement of the 1970's. While the conservation/environmental movement asked questions about preserving the Earth's resources, sustainable development includes questions about how human decisions affect the Earth's environment. , and who has access to these resources.