IJETAE
EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013)

Comparative Analysis of Automatic Summarization Systems for English Language - Neats, Letsum, Information Delivery System for Mobile Commerce and Microsoft Word

Jasmeen[1], Jagroop Kaur[2]

[1,2]*University College of Engineering, Punjabi University, Patiala*

***Abstract*— Automatic summarization is the creation of a shortened version of a text by a computer program. The product of this procedure still contains the most important points of the original text. The phenomenon of information overload has meant that access to coherent and correctly-developed summaries is vital. In this paper the comparison of the four systems of the automatic summarization is held and the systems are: NeATS, LetSum, Information delivery system and Microsoft word. A basic and detailed comparison study is carried out to endow these systems, aiming to investigate how automatic summarization is held.**

***Keywords*— Multi document summarization, table style summary, Document summarization, Mobile commerce, Fractal summarization.**

I. INTRODUCTION

Automatic Summarization: Automatic summarization [1] involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original. Key phrase extraction. A piece of text is given such as a journal article, and you must produce a list of keywords or key phrases that capture the primary topics discussed in the text. Like key phrase extraction, document summarization hopes to identify the essence of a text. The only real difference is that now we are dealing with larger text units—whole sentences instead of words and phrases.

NeATS: NeATS is an extraction based multi document summarization system. It is a system to take out the important portion of text from the document. We outline the NeATS system and describe how it performs content selection, filtering, and presentation. NeATS is among the best performers in the large scale summarization evaluation DUC-2001 [2], the Document Understanding Conference (DUC) sponsored by the National Institute of Standards and Technology(NIST) started in 2001 in the United States. NeATS generates summaries in three stages: content selection, filtering, and presentation.

LetSum: LetSum (Legal text Summarizer), a prototype system, which determines the thematic structure of a judgment in four themes Introduction, Context, Juridical Analysis and Conclusion. Then it identifies the relevant sentences for each theme. The approach to produce summary is based on the identification of thematic structure to find the argumentative themes of judgment. ROUGE [3] software is used to compare the original summary with the model summary. We extract the relevant sentences for each and present them as a table-style summary.

Information Delivery System for Mobile Commerce: The advance of mobile network creates business opportunities and provides value-added services to users. Access to the Internet through mobile phones and other handheld devices is growing significantly in recent years. With a fast paced economy, organizations need to make decisions as fast as possible, they must gain competitive advantage by having access to the most current and accurate information available. There are many shortcomings associated with handheld devices although the development of handheld devices is fast in the recent years. The shortcomings include limited screen size with low resolution, low bandwidth, and low memory capacity. It is impossible to search and visualize the critical information on a small screen with an intolerable slow downloading speed using handheld devices. Automatic summarization summarizes a document for users to preview its major content. Users may determine if the information fits their needs by reading the summary instead of browsing the whole document one by one. Summarization techniques have been applied to delivery of information on handheld devices. The document summarization on handheld devices must make use of tree view or hierarchical display [4].

Microsoft Word: In MS-Word, the sentences that contain words used frequently in the document are given a higher score and assumed as the most important sentence. If these sentences occur in the heading, they too can be included in the summary. Since the sentences can be of any size. This condition is unacceptable.

The solution to this problem which we have used is to divide by the number of words, hence we find out the rank of sentences per word. An important pre-processing which we have done is that we have stemmed each word before calculating the TF. There can be words like summary, summarize, summarizer. But the root word is always summary.

II. NEATS

NeATS is a multi-document summarization [5] system that attempts to extract relevant or interesting portions from a set of documents about some topic and present them in coherent order., NeATS generates summaries in three stages: content selection, filtering, and presentation.

A. *Content Selection*

The goal of content selection [6] is to identify the important concepts mentioned in a document collection. For example: AA flight 11. NeATS computes the likelihood ratio λ to identify key concepts in unigrams, bigrams, and trigrams. Clusters are formed through strict lexical connection e.g. Milan and Kucan are grouped as "Milan Kucan" since "Milan Kucan" is a key bigram concept. Each sentence in the document set is then ranked e.g. , a sentence containing "Milan Kucan" has a higher score than sentence contains only either Milan or Kucan.

B. *Content Filtering*

NeATS uses three different filters:

1) *Sentence Position:* Sentence position has been used as a good important content filter. It was also used as a baseline in a preliminary multi-document. We apply a simple sentence filter that only retains the lead 10 sentences.

2) *Stigma Words:* Some sentences start with conjunctions, verb, quotation marks, and pronouns usually cause discontinuity in summaries. We simply reduce the scores of these sentences to avoid including them in short summaries.

3) *Maximum Marginal Relevancy:* A sentence is added to the summary if and only if its content has less than X percent overlap with the summary. The overlap ratio is computed using simple stemmed word overlap and the threshold X is set empirically.

C. *Content Presentation*

We face two problems: definite noun phrases and events spread along an extended timeline. We describe these problems and the solutions:

1) *A Buddy System of Paired Sentences:* A sentence has a higher score (34.60) than other sentence (32.20) and should be included in the shorter summary (size="50").If we select 1st sentence without including sentence 2nd, the definite noun phrase seems to come without any context. To remedy this problem, we introduce a buddy system to improve cohesion and coherence. Each sentence is paired with a suitable introductory sentence unless it is already an introductory sentence. This assumes lead sentences provide introduction and context information about what is coming next.

2) *Time Annotation and Sequence:* One main problem in multi-document summarization is that documents in a collection might span an extended time period e.g. in topic "Slovenia Secession from Yugoslavia" contains 11 documents dated from 1988 to 1994, from 5 different sources. In multi document summarization, a date expression such as Monday occurring in two different documents might mean the same date or different dates. If no absolute time references are given, the summary might mislead the reader to think that all the events mentioned in the summary sentences occurred in a single week. Therefore, time disambiguation and normalization are very important in multi document summarization. We then order the summary sentences in their chronological order. Each sentence is marked with its publication date and a reference date is inserted after every date expression.

III. LETSUM, LEGAL TEXT SUMMARIZING SYSTEM

The approach to produce summary is based on the identification of thematic structure to find the argumentative themes of judgment. This approach is a result of analysis in which we compared model summaries written by humans with the texts of the original judgments by using ROUGE [3] software. We extract the relevant sentences for each and present them as a table-style summary.

A. *Components of LetSum*

Pre-Processing: It splits the input judgment into main units. First the body of the text of the decision are identified. Some keywords like Reasons for order, Reasons for judgment [7] and order separate the basic data, placed in the head of document. The features used for the end of the decision are the date and place of hearing, name and signature of the judge. Then the document is divided into: section titles, paragraphs, sentences and tokens.

Thematic segmentation: It is based on the specific knowledge of the legal field. According to our analysis, the texts of jurisprudence have a thematic structure, independently of the category of judgment.

Introduction: It describes the situation before the court and answers these questions: who? Did what? To whom? It includes application for judicial review, application to review a decision.

Context: It explains the facts in chronological order, Section titles are Facts, Background, Factual background and Agreed statement of the facts.

Juridical Analysis: It describes the comments of the judge and finding of facts, and the application of the law to the facts as found.

Conclusion: It expresses the disposition which is the final part of a decision containing the information about what is decided by the court. The motion is dismissed, application must be granted.

Filtering identifies parts of the text which can be eliminated, without losing relevant information for the summary. Citation units occupy large volume in the text, up to 30% and content is less important for the summary so we remove citations inside blocks of thematic segments. In the case of eliminating a citation of legislation, we save the reference of the citation in decision data. The identification of citations is based on two types of markers: direct and indirect. A direct marker is one of the linguistic indicators that we classified into three classes: verbs, concepts and complementary indications. The indirect citations are the neighbouring units of a quoted phrase.

Selection builds a list of the best candidate units for each structural level of the summary. LetSum computes a score for each sentence in the judgment based on heuristic functions related to the following information: position of the paragraphs in the document, position of the paragraphs in the thematic segment, position of the sentences in the paragraph, distribution of the words in document. For example, the phrase application is dismissed that can be considered in the conclusion. At the end of this stage, the passages with the highest resulting scores are sorted to determine the most relevant ones.

Production controls the size of the summary and displays the selected sentences in tabular format. In the Introduction segment, units with the highest score are kept within 10%, Context segment occupy 25%, Juridical Analysis occupy 60% and Conclusion occupy 5% of the summary.

IV. AN INFORMATION DELIVERY SYSTEM FOR MOBILE COMMERCE

The advance of mobile network creates business opportunities and provides value-added services to users. Access to the Internet through mobile phones and other handheld devices is growing significantly in recent years. The shortcomings include limited screen size with low resolution, low bandwidth, and low memory capacity.

It is impossible to search and visualize the critical information on a small screen with an intolerable slow downloading speed using handheld devices. However, it is believed that the document summarization on handheld devices must make use of tree view or hierarchical display.

A. Document delivery architecture on handheld devices

It uses three tier architecture. There are two major categories of wireless handheld devices, namely WAP-enabled mobile phones and wireless PDAs.

B. Fractal Summarization

Advance summarization techniques take the document structure into consideration to compute the probability of a sentence to be included in the summary. The sentences according to the document structure from the top level to the low level until sufficient information has been extracted [9]. Fractal summarization generates a brief skeleton of summary at the first stage, and the details of the summary on different levels of the document are generated on demands of users. Such interactive summarization reduces the computation load. Fractal summarization is developed based on the fractal theory. In fractal summarization, the important information is captured from the source text by exploring the hierarchical structure and salient features of the document. A document consists of chapters. A chapter consists of sections. A section may consist of subsections. A section or subsection consists of paragraphs. A paragraph consists of sentences. A sentence consists of terms. A term consists of words. A word consists of characters. A document structure can be considered as a fractal structure. At the lower abstraction level of a document, more specific information can be obtained. The smallest unit in a document is character.

Fractal Summarization Model

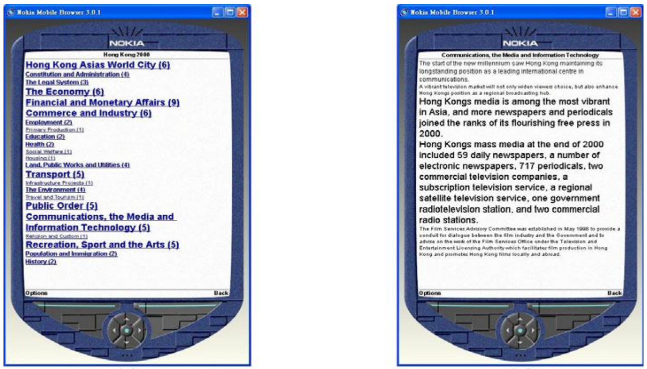
```
i.) Choose a Compression Ratio.
ii.) Choose a Threshold Value.
iii.) Calculate the Sentence Number Quota of the summary.
iv.) Divide the document into range-blocks.
v.) Transform the document into fractal tree.
vi.) Set the current node to the root of the fractal tree.
vii.) Repeat
        a. For each child node under current node, calculate fractal value of child node.
        b. Allocate Quota to child nodes in proportion to fractal values.
        c. For each child nodes,
        If the quota is less than threshold value Select the sentences in the range-block by extraction
        Else
        Set the current node to the child node.
        Repeat Step a, b, c
```

viii.) Until the entire child nodes under current node are processed

The compression ratio of summarization is defined as the ratio of number of sentences in the summary to the number of sentences in the source document.

C. *Visualization of fractal Summarization on hand held devices*

WML is the mark up language supported by wireless handheld devices. The basic unit of a WML file is a deck; each deck must contain one or more cards. The card element defines the content displayed to users, and the card cannot be nested. Each card links to another card within or across decks. Nodes on the fractal tree of fractal summarization model are converted into cards, and anchor links are utilized to implement the tree structure. Given a card of a summary node, there may be a lot of sentences or child nodes. Fisheye View [10] is a visualization technique to enlarge the focus of interest and diminish the information that is less important. There are totally 23 chapters in the annual report, 8 of them are in large font, which means that they are more important, and the rest are in normal font or small font according to their importance to the report. The number inside the parentheses indicates the number of sentences under the node that are extracted as part of the summary.


Nokia Mobile Browser 3.0.1
NOKIA
Hong Kong 2000
Hong Kong Asias World City (6)
Constitution and Administration (4)
The Legal System (3)
The Economy (6)
Financial and Monetary Affairs (9)
Commerce and Industry (6)
Employment (2)
Primary Production (1)
Education (2)
Health (2)
Social Welfare (1)
Housing (1)
Land, Public Works and Utilities (4)
Transport (5)
Infrastructure Projects (1)
The Environment (4)
Travel and Tourism (1)
Public Order (5)
Communications, the Media and Information Technology (5)
Religion and Custom (1)
Recreation, Sport and the Arts (5)
Population and Immigration (2)
History (2)
Options
Back
Nokia Mobile Browser 3.0.1
NOKIA
Communications, the Media and Information Technology
Options
Back

The start of the new millennium saw Hong Kong maintaining its longstanding position as a leading international centre in communications.

A vibrant television market will not only widen viewers choice, but also enhance Hong Kongs position as a regional broadcasting hub.

Hong Kongs media is among the most vibrant in Asia, and more newspapers and periodicals joined the ranks of its flourishing free press in 2000.

Hong Kongs mass media at the end of 2000 included 59 daily newspapers, a number of electronic newspapers, 717 periodicals, two commercial television companies, a subscription television service, a regional satellite television service, one government radiotelevision station, and two commercial radio stations.

The Film Services Advisory Committee was established in May 1998 to provide a conduit for dialogue between the film industry and the Government and to advise on the work of the Film Services Office under the Television and Entertainment Licensing Authority which facilitates film production in Hong Kong and promotes Hong Kong films locally and abroad.

(a) (b)

Figure 1 Fractal Summarization on Mobile Phones

If the user wants to explore a particular node, user can click the anchor link Figure 1 (a) and the handheld device sends the request to the WAP gateway, and the gateway then decides whether to deliver another menu or the summary of the node to the user depends on its fractal value and quota allocated. Figure 1 (b) shows the summary of a particular chapter.

V. MICROSOFT WORD

AutoSummarize analyses a document and then assigns a score to each sentence.

AutoSummarize uses the scoring system to extract the key points and assemble them automatically.

Term Frequency: According to MS Word, Sentences that contain words used frequently in the document are given a higher score and assumed as the most important sentence. The words which occur in the heading and too occur in the body of document that words contain higher score. But just adding the scores will not be correct. Since the sentences can be of any size. This condition is unacceptable. The solution to this problem which we have used is to divide by the number of words, hence we find out the rank of sentences per word. An important pre-processing which we have done is that we have stemmed each word before calculating the TF. This is very important because on a input document on summarization there can be words like summary, summarize, summarizer. But the root word is always summary. Hence stemming increases the TF of the word summary. For stemming software called WordNet is used.

$$TFW_i = \sum_{j=1}^{n} (\partial_{i,j}) w_j \qquad (1)$$

where $\partial_{i,j} = 1$ if jth term exists in sentence 'i',

otherwise $\partial_{i.\, i} = 0$.

For example, in test document we had 247 sentences. In the drop down menu, there are options only for 10 Sentence summaries and above. The percentage has to be manually entered, so in this case if I want 5 sentences, I'd enter 1%.

Cue Word: Cue Phrases are the categories of phrases like {the most important', 'Hence we conclude',' in this paper we show 'etc….} which automatically implies that those sentences that contains these cue-phrases should contain higher ranks. The solution to this problem which we have used is to check for any cue-phrases occurring in the document and if the sentences contain any of the phrases then assign a very high rank.

Position-Based: This model deals with the fact that the start sentence is an important sentence of the document in most of the cases. So assign higher rank to the initial sentence.

Algorithm [11]:

Input: A text file with .txt or .rtf extension and the Percentage of the text user want to extract in resultant summary.

1. Read the text file in .txt or .rtf format.
2. Split the text file into individual tokens.
3. Removing the stop words to filter the text.(e.g. as, ago, after, below etc)

4. Extract the Parts Of Speech of each word by connecting to the Word Net dictionary using Java Word Net Library (JWNL) (e.g. verb, pronouns, nouns etc)
5. Removing the redundancy in the text using cosine Similarity between the sentences
6. Extracting the nouns in each sentence and giving a weight For each sentence as :
   Wtn= (No. of nouns in sentence) / (total no. of words in sentence)
7. Assign an Attention factor to the sentences which appear in bold, italic, underlined or any combination of these.
   Afs= (count of special effect term*a_value) / (Total no. of special effect terms in document)
   Where a_value is taken as follows: for bold, italic, Underlined, a_value=1 for bold-italic, italic-underlined, bold- underlined, a_value=2, for bold-italic-underlined, a_value=3
8. Calculate the total weight of the each sentence as:
   Wts=Wtn+Afs
9. Rank the sentences according the weight of the sentence Wts
10. Finally, extract the higher ranked sentences including the First sentence of the input text in order to find the required Summary.

Output: A relevant summarized text which is shorter than the original text.

VI. Comparison Of The Systems

Table I
COMPARISON OF SYSTEMS BY THEIR PARAMETERS

Parameters	**Systems**			
	NeATS	**LetSum**	**Information Delivery System for Mobile Commerce**	**Microsoft Word**
1.Summarization Type	Multi document Summarization	Table Style Summarization	Fractal Summarization	Single Document Summarization
2. Scoring	Give higher score to bigram sentence than to unigram sentence	Score each sentence based on heuristic functions	Scoring is held according to prefactal structure of document	Give higher score to the sentences that occur most frequently
3.Relevance	A sentence is added to the summary if its content has less than X percent overlap with the summary	Relevance of sentence is taken by the sentence position and word position	The sentence held at the root of the fractal summarization model whose fractal value is largest, taken as relevance sentence	The sentences at the heading and the highest score sentences are taken as the relevant sentence
4.Evaluation Procedure	Two system generated baseline summaries are created, NIST assessors compare this with other assessors Summary	Two steps of Evaluation-evaluation of the modules of the system and global evaluation of produced summaries	System utilizes the fractal value to compute the sentence quota of each range-block by sharing the quota of parent node	The percentage of the words of original document included in summary and the type of summary is entered manually

NeATS perform extraction by giving higher weight to the bigram sentences than unigram and remove verbs, citation, pronouns etc. from the document and a sentence is added to the summary if and only if its content has less than X percent overlap with the summary, X percent is the threshold value which is set empirically. LetSum perform extraction by removing the citations and before removing that should be stored in the decision data, direct and indirect markers are used to find the citations.

It perform selection of relevant concepts according to heuristic functions e.g. position of sentence, position of word. Information delivery system for mobile commerce use three tier architecture for WAP enabled mobile phones and PDAs. It performs fractal summarization which generates a brief skeleton of summary at the first stage, and the details of the summary on different levels of the document are generated on demands of users.

Microsoft Word extract document which occur most frequently in the document and give that the highest score and the sentence contain cue phrases contain higher score and it takes the start sentence as the important sentence and included in the summary.

VII. Conclusion

From the above analysis and basic study of the four systems of the automatic summarization [12], the extraction of the relevant sentences from the document is held. NeATS deliberately used simple methods guided by a few principles: Extracting important concepts based on reliable statistics. Filter sentences by their positions and stigma words. Reduce redundancy using MMR. Present summary sentences in their chronological order with time annotations. The generation of the summary in LetSum is done in four steps: thematic segmentation to detect the legal document structure in four themes Introduction, Context, Juridical analysis and Conclusion, filtering to eliminate unimportant quotations and noises, selection of the candidate units and production of table style summary. The presentation of the summary is in a tabular form. There are many shortcomings of the handheld devices, such as limited resolution and narrow bandwidth. In order to overcome the shortcomings, fractal summarization proposed in this paper. The fractal summarization creates a summary in hierarchical tree structure and presents the summary to the handheld devices through cards in WML. In Microsoft Word summarization of the document is based on three models: term frequency, cue word, position Based.

REFERENCES

[1] Inderjeet Mani. 2001. Automatic Text Summarization. John Benjamins Publishing Company.

[2] DUC. 2001. The Document Understanding Workshop 2001. http://www-nlpir.nist.gov/projects/duc/2001.html.

[3] Chin Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out Workshop held in conjunction with ACL'2004, pages 74–81, Barcelona, Spain, 25–26 July 2004.

[4] Mani, Recent development in text summarization, The Proceedings of the tenth International Conference on Information and Knowledge Management (CIKM'01), ACM Press, Atlanta, GA, USA, 2001 (Nov.), pp. 529– 531.

[5] McKeown, K., R. Barzilay, D. Evans, V. Hatzivassiloglou, M-Y Kan, B, Schiffman, and S. Teufel 2001. Columbia Multi- Document Summarization: Approach and Evaluation. DUC-01 Workshop on Text Summarization. New Orleans, LA.

[6] Goldstein, J., M. Kantrowitz, V. Mittal, and J. Carbonell. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR-99), Berkeley, CA, 121–128.

[7] Louise Mailhot. 1998. Decisions, Decisions: a handbook for judicial writing. Editions Yvon Blais, Qu´ebec, Canada.

[8] Filipe Borges, Raoul Borges, and Daniele Bourcier. Artificial neural networks and legal categorization. In The 16th Annual Conference on Legal Knowledge and Information Systems (JURIX'03), page 187, The Netherlands, 11 and 12 December 2003.

[9] J. Feder, Fractals, Plenum, New York, 1988.

[10] G.W. Furnas, Generalized fisheye views, Proceedings of the SIGCHI Conference on Human Factors in Computing System (CHI '86), ACM SIGCHI Bulletin, vol. 17 (4), 1986 (Apr.), pp. 16– 23.

[11] D.K. Harman, Ranking algorithms, in: W.B. Frakes, R. Baeza Yates (Eds.), Information d, Retrieval: Data Structures and Algorithms, Prentice-Hall, 1992, pp. 363–392 (Ch. 14).

[12] Inderjeet Mani. 2001. Automatic Summarization. John Benjamins Publisher. Dipanjan Das, Andr_e F.T. Martins. 2007. A Survey on Automatic Text Summarization.