# Text Summarization within the Latent Semantic Analysis Framework: Comparative Study

Rasha Mohammed Badry
Faculty of Computers and Information, Fayoum University
Fayoum, Egypt

Ahmed Sharaf Eldin
Faculty of Computers and Information, Helwan University
Cairo, Egypt

Doaa Saad Elzanfally
Faculty of Computers and Information, Helwan University
Cairo, Egypt

## ABSTRACT
It is very difficult for human beings to manually summarize large documents of text. Text summarization solves this problem. Nowadays, Text summarization systems are among the most attractive research areas. Text summarization (TS) is used to provide a shorter version of the original text and keeping the overall meaning. There are various methods that aim to find out well-formed summaries. One of the most commonly used methods is the Latent Semantic Analysis (LSA). In this review, we present a comparative study among almost algorithms based on Latent Semantic Analysis (LSA) approach.

## General Terms
Natural Language Processing (NLP).

## Keywords
Text Summarization, Latent Semantic Analysis, SVD, Sentence Extraction.

## 1. INTRODUCTION
We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

Text summarization is considered one of the most important applications. Text summarization systems produce concise information from the source document, and then the user can easily determine the more relevant documents without reading the whole document. Text summarization systems are useful for many systems such as search engine, and for many people like researchers. They help them to do their jobs more efficiently.

Automatic text summarization (TS) is a process of generating a summary that contains important concepts and sentences of an original document. It is also a process that results in a decrease of the document length.

Text summarization (TS) methods can be classified into extractive and abstractive summarization method [4,6,8]. An extractive summarization method involves collecting and selecting important sentences from the original document to generate summaries into shorter form. The sentence is extracted based on statistical and linguistic features. An abstractive summarization method involves understanding the original document and generating new sentences from the given document into shorter form. The new sentences represent the most important information from the original document. An abstractive summarization method is a more complex task where it is similar to human summarization [7]. Most of ATS systems are extractive summarization systems.

In extractive summarization systems, the important sentences are selected from the original text. Various approaches are used to determine the important sentences[3,6]. One of these approaches that is used in the summarization systems is based on semantic oriented analysis such as lexical chains. Lately, Latent Semantic Analysis (LSA) is used to determine the important sentences and successful results are obtained [9]. LSA will be discussed in more details in section2.

There are various algorithms that use LSA for text summarization. In this paper, we present the existing algorithms that use different LSA approaches[2,9,10].

## 2. Latent Semantic Analysis (LSA)
Latent Semantic Analysis (LSA) is used in many applications (e.g. information retrieval, document categorization, information filtering, and text summarization ). LSA is a method based on statistical calculations to extract and represent the contextual meaning of words and the similarity of sentences[5,10,12]. It is an unsupervised method of deriving vector space semantic representation from a large corpus of data[3], which doesn't need any training or external knowledge. LSA uses context of input document and extracts information such as (1) which words are used together. and (2) which common words are seen in different sentences. we can conclude that if the number of common words between sentences is high, it means that the sentences are more semantically related [1,2].LSA is based on mathematical technique which is named singular value decomposition (SVD) [1,9]. SVD is mathematical matrix decomposition techniques to (1) identify patterns in the relationships between the terms and sentences contained in an unstructured collection of text. And (2) determine the similarity of meaning of words and sentences.

LSA has three main steps. These steps are as follows [1,2,10]:

i. The creation of input matrix: the text (input document) is represented as a matrix. Each row represents the word and each column represents the sentence. The cell value represents the importance of the word. There are many different approaches to fill the cell values such as the frequency of the words in sentences.
ii. Singular Value Decomposition (SVD): singular value decomposition is a mathematical method

applied to the input matrix. SVD is used to identify patterns in the relationships between the terms and sentences. SVD as a mathematical equation can be represented as an m×n matrix (M). M is formed as

$$M = U \Sigma V^T \quad \text{..... (1)}$$

Where U is an m×n matrix which represents the original rows as vectors of extracted values ,

Σ is an n×n rectangular diagonal matrix with nonnegative real numbers on the diagonal representing the scaling values,

and $V^T$ (the conjugate transpose of V) is an n×n real or complex unitary matrix which represents the original columns as vectors of extracted values .

iii. Sentence Selection: after applying the SVD, its result is used to select the sentences to generate the summary. There are many methods and algorithms to select the sentences. These algorithms will be explained in the following sections

LSA has many properties that make it widely applicable to many problems as follows:

1. LSA is a global algorithm that has the ability to collect all trends and patterns from all documents and all words
2. LSA provides the ability to retrieve documents based on words and vice versa. Where LSA is used to map the documents and words to the same concept space.
3. The concept space contains fewer dimensions where these dimensions contain the most information and least noise.

LSA has several limitations that must be considered when deciding whether to use LSA. Some of these are:

1. LSA is difficult to handle the polysemy . Polysemy means that the words with multiple meanings depending on the context. In other words, the same word with different meanings has the same concept and this will cause a big problem.
2. LSA depends on SVD. SVD has some disadvantages which are (1) SVD is time consuming and (2) when new documents are added, their calculations are very hard to be performed.

Since SVD is a very complex algorithm, the performance is decreased.

# 3. GONG AND LIU'S APPROACH (2001)

In [Gong and Liu,2001], one of the main studies in LSA that is utilized for text summarization. The steps of Gong and Liu's approach[9,12] are representing the input document in matrix and doing calculations of SVD. SVD creates a($V^T$) matrix. The order of the row in the created $V^T$ matrix represents the significance of the concept. The cell values represent the cognation between the sentence and the concept. finally, the sentence that is more cognate to the concept has a high cell value. The number of sentences that will be in the summary is given as a parameter. Example: Three sentences are given as an input to LSA:

d0: "The man walked the dog"

d1: "The man took the dog to the park"

d2: "The dog went to the park"

After performing the SVD calculations, the resulting $V^T$ matrix is in Table 1:

**Table 1. $V^T$ matrix**

| $V^T$ **matrix (r=2)** | | | |
|---|---|---|---|
| | **Sent0** | **Sent1** | **Sent2** |
| Con0 | 0,457 | ***0,728*** | 0,510 |
| Con1 | -0,770 | 0,037 | 0,637 |

In Table 1, sentence with highest value is selected. So, concept (con0) is selected and then sentence (sen1) is selected.

There are some disadvantages of Gong and Liu approach that must be considered when deciding whether to use this approach [1,10].

1. If the number of sentences that will be considered in the summary are large, then there are some extracted sentences in the engendered summary may be less important.
2. There are some important concepts that contain highly cognate sentences, but only one sentence is selected from each concept.
3. The third disadvantage is that the same significance level is postulated for all the selected concepts.

# 4. STEINBERGER AND JEZEK'S APPROACH (2004)

Steinberger and Jezek's approach is called a lengthy strategy. This approach [10] has the same first two steps as the approach of Gong and Liu which are representing the input document in matrix and doing calculations on SVD. The Lengthy strategy is different in the way of sentence selection. The length of the sentence vector is used for sentence selection. The sentence length is represented by the row of V matrix and is calculated as follows:

$$Length = \sqrt{\sum_{j=1}^{n} V_{ij} * \sum_{jj}} \quad \text{..... (2)}$$

The dimension (n) of new space is given as a parameter. If the indexes of the concepts are less than or equal to the given dimension, these concepts are used to calculate the length. Also to get the most important concepts, ∑ matrix is used as a multiplication parameter. The sentences with the highest length value are selected to be part of the summary.

Using the example given in Gong and Liu approach, the length values are calculated and the dimension size is two. In Table 2, sentence with highest length value is selected. So,sent1 is selected first to be part of the summary.

**Table 2. Length values**

| **Length values** | |
|---|---|
| Sent0 | 1,043 |
| Sent1 | ***1,929*** |

| | |
|---|---|
| Sent2 | 1,889 |

Steinberger and Jezek's approach try to avoid the limitations of Gong and Liu's approach. In the first disadvantage Gong and Liu's approach, all the extracted sentences are related to all the important concepts. In the second disadvantage of Gong and Liu's approach, more than one sentence are selected from each concept.

# 5. MURRAY, RENALS AND CARLETTA'S APPROACH (2005)

The approach of Murray, Renals and Carletta try to avoid the disadvantages of Gong and Liu's approach. The approach of Murray, Renals and Carletta's starts after representing the input document in matrix and doing calculations of SVD. In the sentence selection step, $V^T$ matrix and $\sum$ matrices are used . This approach has two main functions [11] which are the ability (1) to select more than one sentence from the topmost important concept and (2) to determine how many sentences will be amassed from each concept using Σ matrix . The number of sentences are calculated by getting percentage of the related singular value over the sum of all singular values for each concept.

Using the example given in Gong and Liu approach. In the $V^T$ matrix, sentences with higher values are selected from each row.

**Table 3: $V^T$ matrix**

| $V^T$ **matrix (r=2)** | | | |
|---|---|---|---|
| | **Sent0** | **Sent1** | **Sent2** |
| Con0 | 0,457 | ***0,728*** | **0,510** |
| Con1 | -0,770 | 0,037 | 0,637 |

In Table 3, sentence with higher value is selected. So, from concept (con0) sentences (sen1) and (sen2) are selected to be part of the summary. Two sentences are selected for demonstration purpose.

# 6. OZSOY'S APPROACH (2010)

The approach of ozsoy is one of the main studies that commonly used the Latent Semantic Analysis (LSA). In the sentence selection step, different algorithms use different approaches to extract the important sentences. (Ozsoy 2010) proposes two new methods named cross and topic methods.

## 6.1 Cross Method

Cross method is an extension of Steinberger and Jezek approach [10]. Cross method adds a preprocessing step between the SVD calculations step and sentence selection step. And then the VT matrix is used for sentence selection. The preprocessing step try to remove the sentences that are not one of the most important sentences for each concept. For each sentence, the average value is calculated. Then, the cell value is set to zero if its value is less than or equal to the average. Consider the same example of the previous approaches. Consider the same example of the previous approaches.

**Table 4: $V^T$ matrix after preprocessing**

| $V^T$ **matrix (r=2)** | | | | |
|---|---|---|---|---|
| | **Sent0** | **Sent1** | **Sent2** | **Avg** |
| Con0 | ~~0,457~~ | *0,728* | ~~0,510~~ | .565 |
| Con1 | ~~-0,770~~ | 0,037 | 0,637 | -0.021 |

In Table4, the average value for each concept is calculated. Then, the cell values which are less than or equal to the average are set to zero.

After preprocessing, the total length of each sentence vector is calculated. Then, the sentence with the longest vectors are selected to be included in the summary. In Table5, sen1 is selected to be part of the summary since it has the highest length value.

**Table 5: $V^T$ matrix and length values**

| $V^T$ **matrix (r=2)** | | | |
|---|---|---|---|
| | **Sent0** | **Sent1** | **Sent2** |
| Con0 | 0 | *0,728* | 0 |
| Con1 | 0 | 0,037 | 0,637 |
| Length | 0 | **0.765** | 0.637 |

## 6.2 Topic Method

Topic method is similar to cross method. It is based on discovering the main-concepts and sub-concepts. The extracted concepts from the SVD calculations are called topics of the input document. These topics can be sub-topics, and then the sentences are collected from the main topics. As in cross method, a preprocessing step is added between the SVD calculations step and sentence selection step.

Consider the same example of the previous approaches. In Table6, the average value for each concept is calculated. Then, the cell values which are less than or equal to the average are set to zero.

**Table6: $V^T$ matrix after preprocessing**

| $V^T$ **matrix (r=2)** | | | | |
|---|---|---|---|---|
| | **Sent0** | **Sent1** | **Sent2** | **Avg** |
| **Con0** | ~~0,457~~ | *0,728* | ~~0,510~~ | .565 |
| **Con1** | ~~-0,770~~ | 0,037 | 0,637 | -0.021 |

After preprocessing, the concept x concept matrix is created to find out the main topics. The concept that has common sentences is determined, and new cell values are set to the total of common sentence scores. In Table7, an example of concept x concept matrix is given.

**Table7: New concept x concept matrix**

| | **Con0** | **Con1** |
|---|---|---|
| **Con0** | 1.456 | *0.765* |

| | | |
|---|---|---|
| **Con1** | 0.765 | 1.348 |

After the creation of concept x concept matrix, the strength value is calculated for each concept. The cumulative of the cell values for each row of the concept x concept matrix is used to calculate he strength value. Then, the concept with the highest strength value is selected as the main topic of the input document. In Table 8, the strength values are calculated. So, con0 is selected to be the main topic since it has the highest strength value.

**Table8: Strength values**

| | **Strength** |
|---|---|
| **Con0** | **2.221** |
| **Con1** | 2.113 |

After calculating strength values step, the Gong and Liu's approach [9] is followed to collect the sentences from the preprocessed $V^T$. In able9, sen1 with the highest value is selected from con0. Where one sentence is selected from each concept.

**Table 9. $V^T$ matrix after preprocessing**

| $V^T$ **matrix (r=2)** | | | |
|---|---|---|---|
| | **Sent0** | **Sent1** | **Sent2** |
| Con0 | 0 | ***0,728*** | 0 |
| Con1 | 0 | 0,037 | 0,637 |

## 7. COMPARATIVE SUMMARY AMONG THE LSA BASED SUMMARIZATION ALGORITHMS

Table 10 presents the important comparative parameters to distinguish among various algorithms that use different LSA approaches .

**Table 10. Comparative summary among the LSA based summarization Algorithms**

| **Algorithm with LSA approach, Year** | **Algorithm Name** | **Inputs** | **Extracted Sentences** | **The Main Steps** | **Features for Sentence Selection** | **Outputs** |
|---|---|---|---|---|---|---|
| Gong and Liu's Approach, (2001) | Gong and Liu's Method | Single-document | One sentence/ important concept | 1. The creation of input matrix<br>2. Singular Value Decomposition (SVD) Calculations<br>3. Sentence Selection | **It is based on**<br>1. Matrix ($V^T$) | Extracts |
| Steinberger and Iezek's Approach, (2004) | Lengthy Method | Single-document | More than one sentence/ important concept | 1. The creation of input matrix<br>2. Singular Value Decomposition (SVD) Calculations<br>3. Sentence Selection | **It is based on**<br>1.Mmatrix ($V^T$)<br>2.The length of the sentence vector | Extracts |
| Murray, Renals and Carletta's approach, (2005) | Murray, Renals and Carletta's Method | Single-document | More than one sentence/ important concept | 1. The creation of input matrix<br>2. Singular Value Decomposition (SVD) Calculations<br>3. Sentence Selection | **It is based on**<br>1.$V^T$ matrix<br>2.$\sum$ matrices | Extracts |
| Ozsoy's approach, (2010) | Cross Method | Single-document | More than one sentence/ important concept | 1. The creation of input matrix<br>2. Preprocessing<br>3.Singular Value Decomposition (SVD) Calculations<br>4. Sentence Selection | **It is based on**<br>1.Matrix ($V^T$)<br>2. The average value of each sentence<br>3. The total length of each sentence vector | Extracts |
| | Topic Method | Single-document | More than one sentence/ important concept | 1. The creation of input matrix<br>2. Preprocessing 3.Singular Value Decomposition (SVD) Calculations<br>4. Sentence Selection | **It is based on**<br>1. Matrix ($V^T$)<br>2. The creation of concept x concept matrix<br>3. The strength values of each concept<br>4. Discovering the main-concepts and sub-concepts | Extracts |

## 8. CONCLUSION

There is a need to develop efficient and accurate summarization systems because of the rate of information growth. This review emphasizes approaches to summarization using semantic oriented analysis in order to determine the important sentences. Lately, an algebraic method known as Latent Semantic Analysis (LSA) is used in the determination of the important sentences, and successful results are obtained. In this review, a distinction has been made among the LSA based summarization algorithms.

## 9. FUTURE WORK

We can focus in future to get other methods such as graph based approaches that will be used with Latent Semantic Analysis (LSA) to enhance and improve the performance of the summarization system.

## 10. ACKNOWLEDGMENTS

I would like to express my greatest appreciations to my supervisor Prof. Dr. Ahmed Sharaf Eldin and my co-supervisor. Dr. Doaa Elzanfaly, for their encouragement, guidance and valuable advices.

## 11. REFERENCES

[1] Ozsoy.M.G. Text Summarization Using Latent Semantic Analysis, M.sc thesis, Middle East Technical University, 2011.

[2] Ozsoy.M.G., Cicekli.I.,and Alpaslan.F.N.2010. Text Summarization of Turkish Texts using Latent Semantic Analysis Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 869–87.

[3] Hovy, E.H. and Lin, C.Y.1999. Automated Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. Cambridge: MIT Press, pp. 81–94.

[4] Gupta,V., and Lehal.G.S. 2010. A Survey of Text Summarization Extractive Techniques. JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3.

[5] Landauer, T. K., Foltz, P. W., and Laham, D. 1998. Introduction to Latent Semantic Analysis.Discourse Processes, 25, 259-284.

[6] Hahn.U, and Mani.I. 2000.The challenges of automatic summarization. Computer 33: 29-36.

[7] Lin.C.Y.2004. ROUGE: a Package for Automatic Evaluation of Summaries. Workshop on Text Summarization Branches Out (WAS 2004). 25-26.

[8] Das.D., and Martins.A.F.T. 2007. A Survey on Automatic Text Summarization. Literature survey for Language and Statistics II, Carnegie Mellon University.

[9] Gong.Y., and Liu.X. 2001. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of ACM SIGIR. New Orleans, USA.

[10] Steinberger, J. and Jezek, K. 2004. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. Proceedings of ISIM '04, pages 93-100.

[11] Murray, G., Renals, S. and Carletta, J. 2005. Extractive summarization of meeting recordings. Proceedings of the 9th European Conference on Speech Communication and Technology.

[12] Steinberger, J. and Jezek, K. 2009. Evaluation Measures for Text Summarization. Proceedings of Computing and Informatics, Vol 28, pages 251-275