Article

JIS

Journal of Information Science
1–13

DOI: 10.1177/0165551511408848
jis.sagepub.com
SAGE

Text summarization using Latent Semantic Analysis

Makbule Gulcin Ozsoy and Ferda Nur Alpaslan
Department of Computer Engineering, Middle East Technical University, Turkey

Ilyas Cicekli
Department of Computer Engineering, Hacettepe University, Turkey

Abstract
Text summarization solves the problem of presenting the information needed by a user in a compact form. There are different approaches to create well formed summaries. One of the newest methods is the Latent Semantic Analysis (LSA). In this paper, different LSA-based summarization algorithms are explained, two of which are proposed by the authors of this paper. The algorithms are evaluated on Turkish and English documents, and their performances are compared using their ROUGE scores. One of our algorithms produces the best scores and both algorithms perform equally well on Turkish and English document sets.

Keywords
information retrieval; Latent Semantic Analysis; text summarization

1. Introduction

The growth in electronically available documents makes it difficult to obtain the necessary information related to the needs of a user. Text summarization systems extract brief information from a given document. By using the summary produced, a user can decide if a document is related to his/her needs without reading the whole document.

In the studies of Radev et al. [1] and Das and Martins [2], the aspects of a summary are defined as follows: first, a summary can be created using single or multiple documents. Second, a summary contains all necessary information and it does not include redundant information. Third, a summary is short, at least shorter than half of the original document.

Text summarization systems can be categorized as extractive or abstractive according to the way the summary is created. Also they can be categorized according to the number of input documents used, as single-documents or multi-document summarizations. Another categorization method is based on the purpose of summary, and defined as generic or query-based summarization. In literature, there are different approaches based on supervised or unsupervised techniques for the summarization systems.

The first studies on document summarization started in the late 1950s, and were based on surface level information. Later, statistical approaches, more semantic-oriented analysis such as lexical chains and algebraic-based methods such as Latent Semantic Analysis (LSA) were developed for text summarization.

In this paper, we present a generic extractive text summarization system based on LSA. We applied the known and proposed LSA-based text summarization approaches to Turkish and English texts. The proposed approaches were presented in Ozsoy et al. [3]. One of the main contributions of this paper is the evaluation of these approaches in different languages.

The rest of the paper is organized as follows. Section 2 presents the related work in single-document summarization. Section 3 explains the LSA approach in detail. The existing algorithms that use different LSA approaches are presented

Corresponding author:
Makbule Gulcin Ozsoy, Department of Computer Engineering, Middle East Technical University, Ankara, Turkey.
Email: e1395383@ceng.metu.edu.tr

[4, 5, 6], and two new algorithms that belong to the writers of this paper are also explained in Section 3. Section 4 presents the evaluation results of these algorithms, and Section 5 presents the concluding remarks.

2. Related work

Text summarization is an active research area of natural language processing. Text summarization systems can be categorized according to how and why they are created, and what kind of approach is used for the creation of summaries. Detailed information related to the categorization of text summarization systems is given in Section 2.1. Detailed information about different summarization methods in literature is given in Section 2.2.

2.1. Categorization of text summarization systems

Summaries can have different forms [7]. Extractive summarization systems extract important text units (such as sentences, paragraphs etc.) from the input document. The abstractive summarization approach is similar to the way that human summarizers first understand the main concepts of a document, and then generate new sentences which are not seen in the original document. Since abstractive summarization approach is more complex than the extractive summarization, most automatic text summarization systems are extractive.

Another categorization of summarization systems is based on whether they use single or multiple documents [8]. While a single document is used for generating the summary in a single-document summarization system, multiple documents on the same subject are used for the generation of a single summary in multi-document summarization systems.

Summarization systems can also be categorized as generic and query-based. In generic summarization systems, main topics are used to create the summary while in query-based summarization systems, topics that are related to the answer of a question are used for the construction of the summary.

Document summarization systems can also be categorized based on the technique they use, namely as supervised and unsupervised, as mentioned in Patil and Brazdil [9]. Supervised techniques use data sets that are labelled by human annotators. Unsupervised approaches do not use annotated data but linguistic and statistical information obtained from the document itself. There are other categorization systems for document summarization, similar to those explained in Jezek and Steinberger [10].

2.2. Text summarization approaches in literature

There are various text summarization approaches in literature. Most of them are based on extraction of important sentences from the input text. The first study on summarization, which was conducted by in 1958 [11], was based on frequency of the words in a document. After this study other approaches arose, based on simple features like terms from keywords/key phrases, terms from user queries, frequency of words, and position of words/sentences are proposed. The algorithms belonging to Baxendale [12] and Edmundson [13] are examples of the approaches based on simple features.

Statistical methods are another approach for summarization. The SUMMARIST project [8] is a well known text summarization project that uses statistical approach. In this project, concept relevance information extracted from dictionaries and WordNet is used together with natural language-processing methods. Another summarization application based on statistics belongs to Kupiec et al. [14] where a Bayesian classifier is used for sentence extraction.

Text connectivity is another approach for dealing with problems of referencing to the already mentioned parts of a document. Lexical chains method [15, 16] is a well known algorithm that uses text connectivity. In this approach, semantic relations of words are extracted using dictionaries and WordNet. Lexical chains are constructed and used for extracting important sentences in a document, using semantic relations.

There are graph-based summarization approaches for text summarization. As stated in Jezek and Steinberger [10], the well known graph- based algorithms HITS [17] and Google's PageRank [18], were developed to understand the structure of the Web. These methods are then used in text summarization, where nodes represent the sentences, and the edges represent the similarity among the sentences. TextRank [19] and Cluster LexRank [20] are two methods that use graph-based approach for document summarization.

There are also text summarization algorithms based on machine learning. These algorithms use techniques like Naïve-Bayes, Decision Trees, Hidden Markov Model, Log-linear Models, and Neural Networks. More detailed information related to machine learning based text summarization approaches can be found in Das and Martins [2].

In recent years, algebraic methods such as Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF) and Semi-discrete Matrix Decomposition (SDD) [21, 22, 23] have been used for document summarization. Among these algorithms the best known is LSA, which is based on singular value decomposition (SVD). Similarity among sentences and similarity among words are extracted in this algorithm. Other than summarization, the LSA algorithm is also used for document clustering and information filtering.

3. Text summarization using LSA

The algorithms in the literature that use Latent Semantic Analysis (LSA) for text summarization perform differently. In this section, information on LSA will be given and these approaches will then be discussed in more detail.

3.1. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is an algebraic-statistical method which extracts hidden semantic structures of words and sentences. It is an unsupervised approach which does not need any training or external knowledge. LSA uses context of the input document and extracts information such as which words are used together and which common words are seen in different sentences. High number of common words among sentences indicates that the sentences are semantically related. Meaning of a sentence is decided using the word it contains, and meaning of words are decided using the sentences that contains the word. Singular Value Decomposition (SVD), an algebraic method, is used to find out the interrelations between sentences and words. Besides having the capability of modelling relationships among words and sentences, SVD has the capability of noise reduction, which helps to improve accuracy. In order to see how LSA can represent the meaning of words and sentences the following example is given:

Example 1: 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 calculations we get the resulting figure, Figure 1:

From Figure 1, we can see that d1 is more related to d2 than d0; and the word 'walked' is related to the word 'man' but not so much related to the word 'park'. These kinds of analysis can be made by using LSA and input data, without any external knowledge.

The summarization algorithms that are based on LSA method usually contain three main steps.

3.1.1 Step 1.

- *Input matrix creation*: an input document needs to be represented in a way that enables a computer to understand and perform calculations on it. This representation is usually a matrix representation where columns are sentences and rows are words/phrases. The cells are used to represent the importance of words in sentences. Different approaches can be used for filling out the cell values. Since all words are not seen in all sentences, most of the time the created matrix is sparse.

The way in which an input matrix is created is very important for summarization, since it affects the resulting matrices calculated with SVD. As already mentioned, SVD is a complex algorithm and its complexity increases in size of input matrix which degrades the performance. In order to reduce the matrix size, rows of the matrix, i.e. the words, can be reduced by approaches like removing stop words, using the roots of words only, using phrases instead of words and so on. Also, cell values of matrix can change the results of SVD. There are different approaches to fill out the cell values. These approaches are as follows.

- *Frequency of word*: the cell is filled in with the frequency of the word in the sentence.
- *Binary representation*: the cell is filled in with 0/1 depending on the existence of a word in the sentence.
- *Tf-Idf (Term Frequency-Inverse Document Frequency)*: the cell is filled in with tf-idf value of the word. Higher tf-idf value means that the word is more frequent in the sentence but less frequent in the whole document. The higher value indicates that the word is much more representative for that sentence than others.

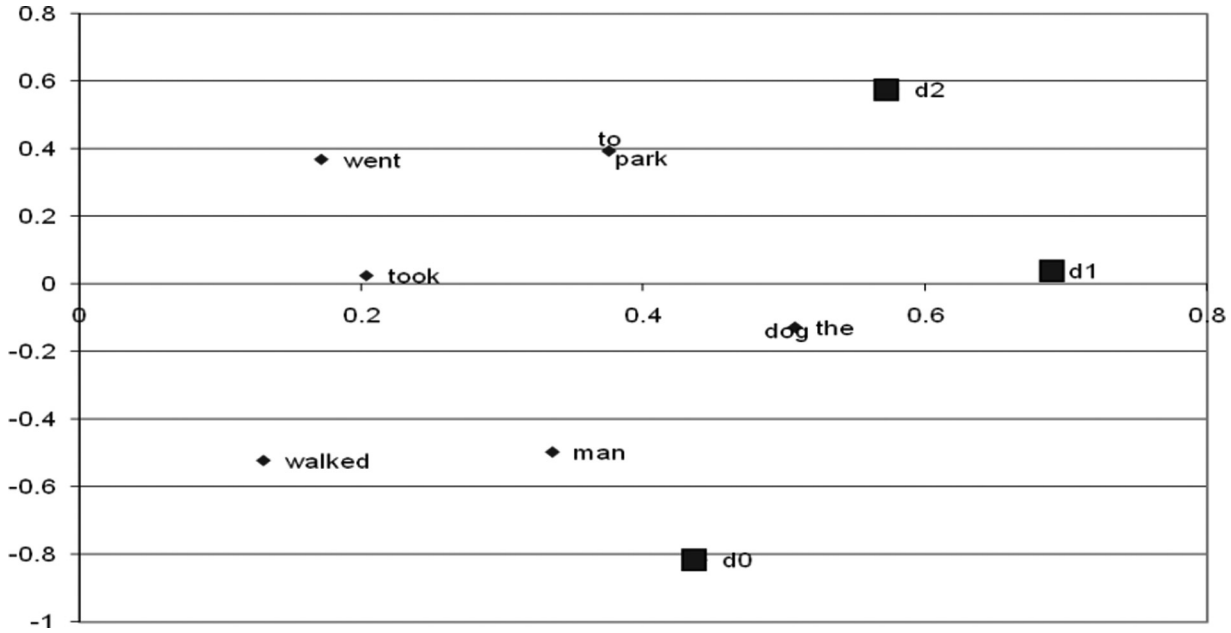

Figure 1. LSA can represent the meaning of words and sentences.

- *Log entropy:* the cell is filled in with log-entropy value of the word, which gives information on how informative the word is in the sentence.
- *Root type*: the cell is filled in with frequency of the word if its root type is a noun, otherwise the cell value is set to 0.
- *Modified Tf-Idf:* this approach is proposed in Ozsoy et al. [3], in order to eliminate noise from the input matrix. The cell values are set to tf-idf scores first, and then the words that have scores less than or equal to the average of the row is set to 0.

3.1.2 Step 2. Singular Value Decomposition (SVD): SVD is an algebraic method that can model relationships among words/phrases and sentences. In this method, the given input matrix A is decomposed into three new matrices as follows:

$A = U \sum V^T$
A: Input matrix (m × n)
U: Words × Extracted Concepts (m × n)
$\sum$: Scaling values, diagonal descending matrix (n × n)
V: Sentences × Extracted Concepts (n × n)

3.1.3 Step 3. Sentence selection: using the results of SVD different algorithms are used to select important sentences. The details of these algorithms are described in Section 3.2.

LSA has several limitations. The first one is that it does not use the information about word order, syntactic relations, and morphologies. This kind of information can be necessary for finding out the meaning of words and texts. The second limitation is that it uses no world knowledge, but just the information that exists in input document. The third limitation is related to the performance of the algorithm. With larger and more inhomogeneous data the performance decreases sharply. The decrease in performance is caused by SVD which is a very complex algorithm.

3.2. Sentence selection approaches

There are various approaches for selection of the sentences while creating summaries using LSA. In this section, five of them will be explained.

V^T matrix (k = 2)			
	Sent0	Sent1	Sent2
Con0	0.457	0.728	0.510
Con1	-0.770	0.037	0.637

Figure 2. V^T matrix. From each row, sentence with highest score is chosen until predefined number of sentences are collected.

3.2.1 Gong and Liu (2001). The algorithm of Gong and Liu [4] is one of the main studies conducted in LSA-based text summarization. After representing the input document in the matrix and calculating SVD values, V^T matrix, the matrix of extracted *concepts* × *sentences*, is used for selecting the important sentences. In V^T matrix, row order indicates the importance of the concepts, such that the first row represents the most important concept extracted. The cell values of this matrix show the relation between the sentence and the concept. A higher cell value indicates that the sentence is more related to the concept.

In the approach of Gong and Liu, one sentence is chosen from the most important concept, and then a second sentence is chosen from the second most important concept until a predefined number of sentences are collected. The number of sentences to be collected is given as a parameter.

In the Example 1, three sentences were given, and the SVD calculations were performed accordingly. The resulting V^T matrix having rank set to two is given in Figure 2. In this figure, first, the concept *con0* is chosen, and then the sentence *sent1* is chosen, since it has the highest cell value in that row.

The approach of Gong and Liu has some disadvantages that are defined by Steinberger and Jezek [5]. The first disadvantage is that the number of sentences to be collected is the same with the reduced dimension. If the given predefined number is large, sentences from less significant concepts are chosen. The second disadvantage is related to choosing only one sentence from each concept. Some concepts, especially important ones, can contain sentences that are highly related to the concept, but do not have the highest cell value. The last disadvantage is that all chosen concepts are assumed to be in the same importance level, which may not be true.

3.2.2 Steinberger and Jezek (2004). The approach of Steinberger and Jezek [5] starts with input matrix creation and SVD calculation. The next step is the sentence selection step which differs from the approach of Gong and Liu. The approach of Steinberger and Jezek uses both V and $\sum$ matrixes for sentence selection.

In this approach, length of each sentence vector, represented by the row of V matrix, is used for sentence selection. The length of the sentence *i* is calculated using the concepts whose indexes are less than or equal to the given dimension. $\sum$ matrix is used as a multiplication parameter in order to give more emphasis on the most important concepts. The sentence with the highest length value is chosen to be a part of the resulting summary.

Using the results of Example 1, calculated length values are given in Figure 3. The dimension size is two for this example. Since the sentence *sent1* has the highest length, it is extracted first as a part of the summary.

The main purpose of this algorithm is to create a better summary by getting rid of disadvantages of the Gong and Liu summarization algorithm. In the Steinberger and Jezek approach sentences that are related to all important concepts are chosen, while allowing collection of more than one sentence from an important concept.

3.2.3 Murray et al. (2005). The first two steps of the LSA algorithm are executed before sentence selection step, as in the previous algorithms. In this approach [6], V^T and $\sum$ matrices are used for sentence selection.

More than one sentence can be collected from the topmost important concepts, placed in first rows of the V^T matrix. Decision of how many sentences will be collected from each concept is made by using $\sum$ matrix. The value is decided by getting percentage of the related singular value over the sum of all singular values, for each concept.

In Figure 4, the V^T matrix of Example 1 is given. From the calculations of $\sum$ matrix, it is observed that one sentence collection from the first row is enough, but for demonstration purposes two sentences will be collected from the Figure 4. So, from *con0* the sentences *sent1* and *sent2* are selected as a part of the summary.

The approach of Murray et al. solves Gong and Liu's problem of selecting a single sentence from each concept, especially when the concept is very important. In this approach, more than one sentence can be chosen even if they do not have the highest cell value in the row of the related concept. Also, the reduced dimension does not have to be the same as the number of sentences in the resulting summary.

Length scores	
Sent0	1.043
Sent1	1.929
Sent2	1.889

Figure 3. Length scores. Sentence with highest length score is chosen.

V^T matrix (k = 2)			
	Sent0	Sent1	Sent2
Con0	0.457	0.728	0.510
Con1	-0.770	0.037	0.637

Figure 4. V^T matrix. From each row, sentence with the highest score is chosen until all predefined sentences are collected.

V^T matrix (k = 2)				
	Sent0	Sent1	Sent2	Avg.
Con0	~~0.457~~ 0	0.728	~~0.510~~ 0	0.565
Con1	~~-0.770~~ 0	0.037	0.637	-0.021
Length	0	0.765	0.637	

Figure 5. V^T matrix after pre-processing.

3.2.4 Cross method. Cross method is an extension to the approach of Steinberger and Jezek [5], and proposed in Ozsoy et al. [3]. In this approach, input matrix creation and SVD calculation steps are executed as in the other approaches and the V^T matrix is used for sentence selection purposes. Between the SVD calculation step and the sentence selection step, there exists a pre-processing step. The aim of the pre-processing step is to remove the overall effect of sentences that are related to the concept somehow, but not the core sentence for that concept. For each concept, which is represented by the rows of the V^T matrix, the average sentence score is calculated. Then the cell values which are less than or equal to the average score are set to zero. This process of setting the cell values to the zero whose score is less than average removes less related sentences while keeping more related ones for that concept.

After pre-processing, the steps of Steinberger and Jezek approach are followed with a modification. In our Cross approach, the total length of each sentence vector, which is represented by a column of the V^T matrix, is calculated. While calculating the length score the parameter of number of concepts to be used is given by the user; if it is not given, all of the extracted concepts are used. Then, the longest sentence vectors are collected as a part of the resulting summary.

In Figure 5, an example V^T matrix is given after the pre-processing is executed. For the pre-processing step; first the average score for each concept is calculated, and then the cell values less than this average are set to zero. Finally, length scores are calculated by adding up the concept scores with values after the pre-processing step. In this example matrix, *sen1* has the highest length score, so it has been chosen to be part of the summary.

3.2.5 Topic method. The topic method is proposed by Ozsoy et al. [3]. It performs a pre-processing step, which is followed by the sentence selection. In both of these steps the V^T matrix is used. The main idea in topic method is to find out the main-concepts and sub-concepts. The resulting concepts extracted from SVD calculations are known to be the topics of the input document. But these topics can be sub-topics of other extracted topics as well. In this approach, after deciding the main topics which may be a group of subtopics, the sentences are collected from the main topics as a part of the summary.

The pre-processing step of this approach starts with a similar way of the pre-processing step of Cross approach. First, average sentence score is calculated for each concept using the row of V^T matrix. Then the cell values less than this score are set to zero. This step removes sentences that are not highly related to the concept, leaving only the most important sentences related to that concept. In Figure 6, an example V^T matrix after pre-processing is given.

V^T matrix (k = 2)				
	Sent0	Sent1	Sent2	Avg.
Con0	~~0.457~~ 0	0.728	~~0.510~~ 0	0.565
Con1	~~-0.770~~ 0	0.037	0.637	-0.021

Figure 6. V^T matrix after pre-processing.

	Con0	Con1	Strength
Con0	1.456	0.765	2.221
Con1	0.765	1.348	2.113

Figure 7. New *concept x concept* matrix.

After the first step of pre-processing comes the step of finding out main topics. For this step, a *concept* × *concept* matrix is created by finding out the concepts that have common sentences. The common sentences are the ones that have cell values other than zero in both concepts that are considered. Then the new cell values of the *concept* × *concept* matrix are set to the total of common sentence scores. In Figure 7, the *concept* × *concept* matrix based on the V^T is given. After the creation of the *concept* × *concept* matrix, the strength of each concept is calculated. For each concept, the strength value is computed by getting the cumulative cell values for each row of the *concept* × *concept* matrix. The concept with the highest strength value is chosen as the main topic of the input document. A higher strength value indicates that the concept is much more related to the other concepts, and it is one of the main topics of the input text. In Figure 7, calculated strength values can be seen. Since *con0* has the highest strength value, it is chosen to be the main topic.

After these steps, the sentences are collected from the pre-processed V^T matrix following the approach of Gong and Liu. As explained earlier, a single sentence is collected from each concept until pre-defined numbers of sentences are collected. In the topic method, instead of the topmost concepts of V^T matrix, the chosen main concepts are used for sentence selection. In Figure 6, *sen1* is chosen from *con0*, since that sentence has the highest cell value.

4. Evaluation

Evaluation of summaries is an active research area in natural language processing. There exist different methods for the evaluation, such as using human evaluators, using precision/recall values, or using ROUGE scores. Evaluation of the LSA-based summarization approaches are conducted on Turkish and English datasets in this paper. Different LSA approaches are executed using different input matrix creation methods. In order to make the resulting matrix smaller in size, stemming and stop word removal methods are used. All the summaries created have length of 10% of the input document. The evaluations are based on the ROUGE evaluation measurements.

The ROUGE evaluation approach [24] is based on n-gram co-occurrence, longest common subsequence and weighted longest common subsequence between the ideal summary and the extracted summary. The n-gram based ROUGE score, ROUGE-N, is based on comparing n-grams in the ideal summaries and the reference summary. The longest common subsequence (LCS) based ROUGE score, ROUGE-L, is based on the idea that longer LCS value between the ideal and extracted summary sentences indicate that the sentences are more similar. Another ROUGE measure, namely ROUGE-W, is based on using weights to penalize non-consecutive subsequence matches. ROUGE-S score, on the other hand, is based on ordered pairs of words that are common among the ideal and the extracted summaries. One can reach detailed information on ROUGE evaluation system in Lin [25]. In this paper ROUGE-1, ROUGE-2, ROUGE-3, and ROUGE-L results are obtained, but discussions are made using only ROUGE-L results. Other ROUGE results behave in a similar manner as ROUGE-L score.

In Section 4.1, information about the Turkish datasets and evaluation results for them are given. In Section 4.2, information about the English datasets and evaluation results are presented. In Section 4.3, the results of LSA based summarization approaches are compared against other summarization approaches. Lastly in 4.4, analysis of the evaluation results is given.

Table 1. ROUGE-L f-measure scores for data_set_1

	G&L	S&J	MRC	Cross	Topic
freq	0,236	0,250	0,244	0,302	0,244
binary	0,272	0,275	0,274	0,313	0,274
tf-idf	0,200	0,218	0,213	0,304	0,213
logent	0,230	0,250	0,235	0,302	0,235
root	0,283	0,282	0,289	0,320	0,289
mod. tf-idf	0,195	0,221	0,223	0,290	0,223

4.1. Evaluation results for Turkish datasets

Four different sets of Turkish documents are used for the evaluation of summarization approaches. The first two sets of articles are scientific articles, related to different areas such as medicine, sociology, and psychology. Each dataset contains 50 articles. The articles in the second dataset are longer than the articles in the first set. The evaluation is done by using abstracts of the input documents, whose sentences may not match with the sentences of the input document.

These datasets are already used in Ozsoy et al. [3]. The ROUGE-L f-score values of the first data set (data_set_1) can be found in the Table 1 and the ROUGE-L f-score values of data_set_2 can be found in the Table 2. From the results, it has been observed that Cross method has the highest score among the LSA based summarization approaches. Topic method has achieved better results than the approach of Gong and Liu [4], and has achieved same results as the approach of Murray et al. [6].

The third dataset is composed of news texts in Turkish. It contains 120 texts, and the number of sentences is not as high as in scientific papers. The evaluation results of this dataset can be found in Table 3. The results of the news dataset show that the Cross method does not work well with shorter documents. Topic method results are nearly the same as the results of the Gong and Liu approach [4]

The fourth dataset in Turkish is a new dataset, which is used for the first time. The dataset is composed of scientific articles in Turkish, related to different areas of science, such as sociology, biology, computer science and so on. The number of documents in this dataset is 153. The ROUGE-L f-score results can be seen in Table 4. The results of the new dataset of Turkish scientific papers show that Cross method works better than all other approaches. The results of Topic method are nearly same as Murray et al. approach [6], as observed in the first two datasets.

4.2. Evaluation results for English datasets

The datasets that are used for the evaluation of the LSA-based summarization approaches are Duc2002, Duc2004, and Summac datasets. All datasets are used for single-document summarization. There were tasks defined in the Duc datasets that limit the output summary size. The given output sizes are very short in Duc datasets, e.g. 100 words or less, which can limit the quality of the extracted summaries. Instead, we preferred to extact longer summaries and determined the length of summaries as 10% of the length of original document.

The first dataset is the Duc2002 dataset, which defines three different tasks. Since Task-1 is related to single document summarization, we have chosen this task. In this task, nearly 600 newspaper texts were given as input. The ROUGE-L f-score results are given in Table 5. The results show that the Cross approach achieves the best result. Topic approach achieved nearly the same results as Murray et al. [6], and both have similar results to Gong and Liu approach [4].

The second dataset is the Duc2004 dataset. This dataset defines five different tasks and in this paper Task1 is chosen in order to evaluate the summarization results. For this task, 50 sets of document clusters with nearly 10 documents each are given as input. The input documents are collected from the AP newswire and *The New York Times* newswire. The aim is to create very short summaries of 75 bytes each for each document. As in the Duc2002 dataset, instead of limiting the output result to predefined bytes, the output summary length is set to 10% of original document. The results for Duc2004-Task1 are given in Table 6. As observed in Duc2002-Task1, the Cross approach achieved the best results among other approaches and the Topic approach achieved similar results with Murray et al. [6].

The third dataset is Summac Dataset which contains 183 documents. All documents are scientific articles about computer science collected from ACL sponsored conferences. The comparison of extracted summaries is done against the abstracts of the given input articles. The ROUGE-L f-scores results in Table 7 show that Cross approach gets the best

Table 2. ROUGE-L f-measure scores for data_set_2

	G&L	S&J	MRC	Cross	Topic
freq	0,256	0,251	0,259	0,264	0,259
binary	0,191	0,220	0,189	0,274	0,189
tf-idf	0,230	0,235	0,227	0,266	0,227
logent	0,267	0,245	0,268	0,267	0,268
root	0,194	0,222	0,197	0,263	0,197
mod. tf-idf	0,234	0,239	0,232	0,268	0,232

Table 3. ROUGE-L f-measure scores for the news data set (data_set_3)

	G&L	S&J	MRC	Cross	Topic
freq	0.347	0.341	0.358	0.422	0.347
binary	0.479	0.437	0.465	0.432	0.465
tf-idf	0.303	0.295	0.313	0.389	0.305
logent	0.350	0.316	0.359	0.225	0.350
root	0.501	0.449	0.470	0.454	0.471
mod. tf-idf	0.319	0.291	0.339	0.387	0.320

Table 4. ROUGE-L f-measure scores for the new dataset of Turkish scientific articles data_set_4

	G&L	S&J	MRC	Cross	Topic
freq	0.219	0.192	0.218	0.225	0.217
binary	0.134	0.155	0.133	0.219	0.133
tf-idf	0.212	0.203	0.205	0.231	0.205
logent	0.232	0.191	0.232	0.229	0.231
root	0.232	0.191	0.232	0.229	0.231
mod. tf-idf	0.211	0.202	0.203	0.230	0.202

result when using scientific papers. Topic method produces the same results as the approach of Murray et al. [6] and similar results to the approach of Gong and Liu [4].

4.3. Comparison against other summarization approaches

Evaluation of the LSA-based summarization systems in this paper was carried out using Turkish and English datasets. In order to evaluate the results ROUGE scores were used. As stated in Steinberger [26], different ROUGE scores gives different correlation with human judgment and strongest correlation is observed with ROUGE-1 and ROUGE-L scores.

The datasets used in English are common for the evaluation of the summarization systems. The first dataset used is the Duc2002 dataset. The ROUGE-L results for lexical chains based summarization systems were obtained from Ercan [27]. The evaluation results of studies belonging to Wan et al. [28] and the algorithms named SentenceRank [19] and MutualRank [29] were obtained from Wan et al. The result for TextRank, which is based on graphs, was collected from Mihalcea [30]. The results for SumGraph, MEAD [31], and LexRank [32] were obtained from Patil and Brazdil [9]. The results for LSA-based summarization algorithms were gathered from the evaluation results stated in Section 4.2. The best results obtained from these approaches were used for comparison. All results are shown on the Table 8.

Another dataset that we have used for evaluating the results for English documents is the Duc2004 dataset. There are different tasks defined over this dataset, and we have chosen Task1-creation of a very short summary of a single document. As in the Duc2002 dataset, we gathered the results of different algorithms using different resources. The results for lexical chains-based approaches were obtained from Ercan [28]. The results for a machine learning-based algorithm (Dublin) were obtained from the study of Doran et al. [33]. The last two algorithms are TF ad Hybrid algorithms whose evaluation results were collected from the study of Wang et al. [34]. Another machine learning-based algorithm is

Table 5. ROUGE-L scores for Duc2002-Task1

	G&L	S&J	MRC	Cross	Topic
F-measure scores					
freq.	0.093	0.095	0.098	0.179	0.097
binary	0.234	0.224	0.230	0.196	0.230
tf-idf	0.094	0.098	0.109	0.177	0.108
logent.	0.103	0.098	0.104	0.176	0.104
root	0.103	0.098	0.104	0.176	0.104
mod. tf-idf	0.094	0.098	0.119	0.177	0.118
Precision scores					
freq	0.323	0.287	0.315	0.324	0.315
binary	0.304	0.299	0.301	0.323	0.301
tf-idf	0.312	0.295	0.304	0.327	0.303
logent	0.356	0.307	0.347	0.333	0.347
root	0.356	0.307	0.347	0.333	0.347
mod. tf-idf	0.312	0.294	0.307	0.331	0.307

Table 6. ROUGE-L scores for Duc2004-Task1

	G&L	S&J	MRC	Cross	Topic
F-measure scores					
freq.	0.087	0.070	0.082	0.102	0.080
binary	0.103	0.094	0.102	0.097	0.101
tf-idf	0.072	0.064	0.074	0.093	0.071
logent	0.088	0.068	0.082	0.101	0.081
root	0.088	0.068	0.082	0.101	0.081
mod. tf-idf	0.075	0.063	0.080	0.098	0.074
Precision scores					
freq	0.078	0.064	0.074	0.072	0.072
binary	0.064	0.058	0.063	0.065	0.063
tf-idf	0.063	0.057	0.063	0.064	0.061
log ent	0.081	0.065	0.075	0.071	0.075
root	0.081	0.065	0.075	0.071	0.075
mod. tf-idf	0.063	0.057	0.064	0.068	0.060

Table 7. ROUGE-L f-measure scores for Summac dataset

	G&L	S&J	MRC	Cross	Topic
freq	0.161	0.131	0.156	0.177	0.156
binary	0.118	0.136	0.119	0.177	0.119
tf-idf	0.129	0.126	0.121	0.182	0.121
logent	0.180	0.138	0.180	0.182	0.180
root	0.180	0.138	0.180	0.182	0.180
mod. tf-idf	0.127	0.126	0.122	0.187	0.122

LAKE, improved by D'Avanzo et al. [35]. In Table 9, a comparison of different algorithms which performed Duc2004-task1 can be seen.

4.4. Analysis of evaluation results

The evaluation of LSA based summarization systems were performed on multiple dataset that are in Turkish and in English whose results can be seen in Section 4.1 and Section 4.2. The evaluation of LSA-based summarization systems on Turkish documents was carried out using four different datasets. It has been observed that Cross method works better

Table 8. Comparison of precision scores on Duc2002 dataset

	ROUGE-L
ROUGE-L precision scores	
Barzilay	0.309
Ercan	0.285
Gong-Liu	0.356
Steinberger-Jezek	0.307
Murray et al.	0.347
Cross	0.333
Topic	0.347
	ROUGE-1
ROUGE-1 precision scores	
Wan-WordNet	0.473
Wan-Corpus	0.472
SentenceRank	0.462
MutualRank	0.438
TextRank-HITS	0.502
TextRank-PageRank	0.500
SumGraph	0.484
MEAD	0.472
LexRank	0.469
Gong-Liu	0.432
Steinberger-Jezek	0.428
Murray et al.	0.428
Cross	0.453
Topic	0.428

than other LSA based approaches. The results for the Topic method are usually the same as the results of the approach of Murray et al. [6]. Also, it is observed that the Cross method does not perform well in shorter documents.

The evaluation of approaches on English documents is done using three different datasets. Among different LSA-based summarization approaches, the Cross method performed better than the others. It is observed that the performance of Topic method is nearly the same as the approach of Gong and Liu [4]. Also, as in the Turkish datasets, it is observed that the performance of the LSA-based approaches is lower for shorter documents.

In both Turkish and English datasets, the evaluation of the approaches is done using different input matrix creation methods. Different summarization algorithms performed differently for each input matrix creation approach. It is observed that the Cross approach is not affected from the different methods of input matrix creation and that it performed nearly equally well in all approaches.

The input matrix creation approaches used in this paper are all well known methods in literature. The modified tf-idf approach is a newer approach which is proposed by Ozsoy et al. [3]. As stated in that study, it has been observed that this new approach does not produce better results when the input document is short, because every word/sentence in shorter document carries information, and removing some of them may remove important amount of this information.

Comparison of LSA-based summarization systems against other summarization approaches was explained in Section 4.3. It has been observed that LSA-based algorithms do not perform as successfully as machine learning-based algorithms or other algorithms that use outer information. But attention should be given to the fact that LSA-based algorithms use information in the input document only. LSA-based algorithms are unsupervised and do not need any outer information to extract semantic information that exists in the document. Another concern related to LSA-based algorithms is that they do not perform well while creating shorter summaries. LSA based summarization approaches are extractive approaches, and as stated in the paper of Das and Martins [2], there is a claim of Witbrock and Mittal [36] which states that extractive summarization methods are not very efficient when creating very short summaries. This observation was also made by other researchers who stated that finding a single or a few sentences that give the main idea of a text is very difficult [34].

5. Conclusion

Finding out the information related to the needs of a user among large number of documents is a problem that has become obvious with the growth of text-based resources. In order to solve this problem, text summarization methods

Table 9. Comparison of precision scores on Duc2004 dataset.

	ROUGE-L
ROUGE-L precision scores	
Barzilay	0.155
Ercan	0.170
Dublin	0.176
TF	0.171
Hybrid	0.176
LAKE	0.156
Gong-Liu	0.081
Steinberger-Jezek	0.065
Murray et al.	0.075
Cross	0.072
Topic	0.075
	ROUGE-1
ROUGE-1 precision scores	
Barzilay	0.178
Ercan	0.195
Dublin	0.219
TF	0.244
Hybrid	0.219
LAKE	0.188
Gong-Liu	0.090
Steinberger-Jezek	0.071
Murray et al.	0.083
Cross	0.085
Topic	0.083

are proposed and evaluated. The research on summarization started with the extraction of simple features and went on to use different methods, such as lexical chains, statistical approaches, graph-based approaches, and algebraic solutions. One of the algebraic-statistical approaches is the Latent Semantic Analysis method. In this paper, text summarization methods based on Latent Semantic Analysis have been explained.

All approaches were evaluated on Turkish and English datasets and a comparison of the results against other text summarization approaches was carried out. ROUGE scores were used for the evaluation of the results. The results show that among LSA-based approaches the Cross method performs better than all other approaches. Another important result of this approach is that it is not affected by different input matrix creation methods. Also, it is observed that the Cross and Topic methods, which are proposed by the writers of this paper, perform equally well on both Turkish and English datasets. This work shows that the Cross and Topic methods can be used in any language for summarization purposes.

In future, ideas that are used in other methods, such as graph-based approaches, will be used together with the proposed approaches to improve the performance of summarization.

References

[1] Radev DR, Hovy E, McKeown K. Introduction to the special issue on summarization. *Computational Linguistics* 2002; 28(4): 399–408.

[2] Das D, Martins AFT. *A Survey on automatic text summarization*. Unpublished manuscript, Literature survey for Language and Statistics II, Carnegie Mellon University, 2007.

[3] Ozsoy MG, Cicekli I, Alpaslan FN. Text summarization of Turkish texts using Latent Semantic Analysis. In: *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)* 2010: 869–876.

[4] Gong Y, Liu X. Generic text summarization using relevance measure and latent semantic analysis. In: *Proceedings of SIGIR '01* 2001.

[5] Steinberger J, Jezek K. Using Latent Semantic Analysis in text summarization and summary evaluation. In: *Proceedings of ISIM '04* 2004: 93–100.

[6] Murray G, Renals S, Carletta J. Extractive summarization of meeting recordings. In: *Proceedings of the 9th European conference on speech communication and technology* 2005.

[7] Hahn U, Mani I. The challenges of automatic summarization. *Computer* 2000; 33: 29–36.

[8] Hovy E, Lin CY. Automated text summarization in SUMMARIST. In: Mani I, Maybury MT (eds) *Advances in automatic text summarization*. Cambridge, MA: MIT Press, 1999.
[9] Patil K, Brazdil P. Sumgraph: text summarization using centrality in the pathfinder network. *IADIS International Journal on Computer Science and Information Systems* 2007; 2: 18–32.
[10] Jezek K, Steinberger J. Automatic Text Summarization: the state of the art 2007 and new challenges. *Znalosti* 2008: 1–12.
[11] Luhn HP. The automatic creation of literature abstracts. *IBM Journal of Research Development* 1958; 2(2): 159–165.
[12] Baxendale P. Machine-made index for technical literature: an experiment. *IBM Journal of Research Development* 1958; 2(4): 354–361.
[13] Edmundson HP. New methods in automatic extracting. *Journal of the ACM* 1969; 16(2): 264–285.
[14] Kupiec J, Pedersen JO, Chen F. A trainable document summarizer. *Research and Development in Information Retrieval* 1995: 68–73.
[15] Barzilay R, Elhadad M. Using lexical chains for text summarization. In: *Proceedings of the ACL/EACL '97 workshop on intelligent scalable text summarization* 1997: 10–17.
[16] Ercan G, Cicekli I. Lexical cohesion based topic modeling for summarization. In: *Proceedings of 9th international conference on intelligent text processing and computational linguistics (CICLing-2008)* 2008: 582–592.
[17] Kleinberg JM. Authoritative sources in a hyper-linked environment. *Journal of the ACM* 1999; 46(5): 604–632.
[18] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 1998; 30: 1–7.
[19] Mihalcea R, Tarau P. Text-rank: bringing order into texts. In: *Proceeding of the conference on empirical methods in natural language processing* 2004: 404–411.
[20] Qazvinian V, Radev DR. Scientific paper summarization using citation summary networks. In: *Proceedings of COLING 2008* 2008: 689–696.
[21] Landauer TK, Foltz PW, Laham D. An introduction to Latent Semantic Analysis. *Discourse Processes* 1998; 25: 259–284.
[22] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999; 401: 788–791.
[23] Kolda TG, O'Leary DP. A semi discrete matrix decomposition for latent semantic indexing information retrieval. *ACM Transactions on Information Systems* 1998; 16(4): 322–346.
[24] Lin CY, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: *Proceedings of the 2003 conference, North American chapter of the Association for Computational Linguistics on human language technology (HLT-NAACL-2003)* 2003: 71–78.
[25] Lin CY. ROUGE: a package for automatic evaluation of summaries. In: *Proceedings of the workshop on text summarization branches out (WAS 2004)* 2004.
[26] Steinberger J. *Text summarization within the LSA framework*. Doctoral Thesis, 2007.
[27] Ercan G. *Automated text summarization and keyphrase extraction*. MSc Thesis, 2006.
[28] Wan X, Yang J, Xiao J. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In: *Proceedings of the 45th annual meeting of the association of computational linguistics* 2007: 552–559.
[29] Zha HY. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In: *Proceedings of SIGIR 2002* 2002: 113–120.
[30] Mihalcea R. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: *Proceedings of the 42st annual meeting of the Association for Computational Linguistics* 2004: 170–173.
[31] Radev D, Blair-Goldstein S, Zang Z. Experiments in single and multi-document summarization using MEAD. In: *Proceedings of the document understanding conference* 2001.
[32] Erkan G, Radev DR. LexRank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)* 2004; 22: 457–479.
[33] Doran W, Stokes N, Newman E, Dunnion J, Carthy J, Toolan F. News story gisting at University College Dublin. In: *Proceedings of the document understanding conference* 2004.
[34] Wang R, Dunnion J, Carthy J. Machine learning approach to augmenting news head-line generation. In: *Proceedings of the international joint conference on natural language processing* 2005.
[35] D'Avanzo E, Magnini B, Vallin A. Keyphrase extraction for summarization purposes: the LAKE System at DUC-2004. In: *DUC Work-shop, human language technology conference/North American chapter of the Association for Computational Linguistics Annual Meeting* 2004.
[36] Witbrock MJ, Mittal VO. Ultra-summarization (poster abstract): a statistical approach to generating highly condensed non-extractive summaries. In: *Proceedings of SIGIR '99* 1999: 315–316.