



# Genetic Algorithm and Latent Semantic Analysis based Documents Summarization Technique

Imen Tanfourri<sup>1</sup> <sup>a</sup> and Fethi Jarray<sup>1,2</sup> <sup>b</sup>

<sup>1</sup>*LIMITIC Laboratory, UTM University, Tunisia*

<sup>2</sup>*Higher Institute of Computer Science of Medenine, Tunisia*

**Keywords:** Natural Language Processing, Genetic Algorithm, Latent Semantic Analysis, Topic Modeling, Single Document Summarization.

**Abstract:** Automatic text summarization (ATS) is the process of generating or extracting a shorter text of the original document while preserving relevant and important information. Nowadays, it is a hot research topic in natural language processing with various applications, including social networks and the healthcare domain. The task of summarizing can be divided into two categories, extractive and abstractive. In this paper, we are concerned with extractive summarization for a single Arabic document. In this contribution, we propose a combination of semantic and combinatorial methods to summarize a document by clustering its content through topic modeling techniques and subsequently generating an extractive summary for each of the identified topics using genetic algorithms. This approach ensures that the final summary covers all important topics in the document. We achieve state-of-the-art performance on the common Arabic summarization benchmark datasets. The obtained results show the effectiveness of combining genetic algorithms (GA) and latent semantic analysis (LSA) for document summarization.

## 1 INTRODUCTION


Due to the large amount of data currently in digital form, it is important to develop a system that can automatically shorten longer texts to reduce reading time and accelerate the research process for an information. For this, automatic text summarization become a popular problem in Natural Language Processing field. The goal of the text summarization system is to generate a summary that contains all the important information in the original text without redundancy. Automatic text summarization has many applications in information retrieval, information extraction and sentiment analysis (Chouikhi et al., 2021). There are two main approaches to automatic text summarization: extractive and abstractive. Extractive approach, in which systems extract important sentences from the input text and combine them to make a final summary. However, the abstractive approach generates the summary with sentences that are different from the original ones. According to the number of documents to be summarized, there is single and multi-


document summarization. A single-document system generates a summary of just one document. In contrast, a multi-document system generates a summary of various documents that discuss the same topic. In this paper, we propose an extractive single document summarization system for Arabic texts.

This paper is organized as follows. The next section presents a short survey of recent approaches for Arabic text summarization. Section 3 describes our proposed methodology for the text summarization task. Section 4 explains experiment settings, dataset, and evaluation results. Section 5 concludes and discusses future work.

## 2 RELATED WORKS

Most previous work in automatic text summarization can be classified into four approaches: 1) Statistical approach, 2) Machine learning approach, 3) Semantic approach and 4) Meta-heuristic approach. Let's give more details about these approaches.

<sup>a</sup>  <https://orcid.org/0000-0003-2504-5267>

<sup>b</sup>  <https://orcid.org/0000-0003-2007-2645>

## 2.1 Statistical Approach

These methods depend on statistical features of sentences such as the frequency of the term, the frequency of the inverse document, the position of the sentence, and many other characteristics to extract important sentences and words from the source text. Al-Hashemi (Al-Hashemi, 2010) proposed a statistical-based system for automatic text summarization based on some features such as TF-IDF to extract keyphrases to select important sentences for summary. Haboush (Haboush et al., 2012) proposed a model using a technique of word root clustering. This model adopts cluster weight of word roots instead of the word weight itself, using TF-IDF on clustered word roots.

## 2.2 Machine Learning Approach

These methods convert the summarization problem to a supervised classification problem at the sentence level. There are several techniques used in the machine learning approach, including Nave Bayesian (NB), support vector machine (SVM), adaptive boosting (AdaBoost), and hidden Markov model (HMM). Boudabous et al. (Boudabous et al., 2010) present an automatic summarization method of Arabic documents based on the SVM algorithm. Belkebir and Guessoum (Belkebir and Guessoum, 2015) have proposed an extractive machine learning-based approach to Arabic text summarization that includes training of two classifiers AdaBoost and SVM to predict whether a sentence is a summary sentence or not. Abu Nada et al. (Abu Nada et al., 2020) proposed an extractive Arabic text summarizer using AraBERT model to summarize the Arabic document by evaluating and extracting the most important sentences at this document.

## 2.3 Semantic Approach

Latent Semantic Analysis (LSA) is an unsupervised learning algorithm that is frequently used for extractive text summarization. Extract semantically significant sentences by applying singular value decomposition (SVD) to the matrix of term-document frequency. Froud et al. (Froud et al., 2013) use LSA model on Arabic documents to solve the problems of noisy information and improve the clustering performance. AL-Khawaldeh and Samawi (AL-Khawaldeh and Samawi, 2015) used lexical cohesion and text entailment-based segmentation as scoring measures to prevent redundant and less important sentences from being generated in the summary. In addition,

many other semantic-based techniques are used for automatic text summarization, like Semantic Role Labelling (SRL) and Explicit Semantic Analysis (ESA). Mohamed and Oussalah (Mohamed and Oussalah, 2019) propose a semantic based extractive text summarization system based on SRL and ESA in which the SRL is used to achieve sentence semantic parsing whose word tokens are represented as a vector of weighted Wikipedia concepts using ESA method.

## 2.4 Meta-heuristic Approach

These methods convert the summarization problem to an optimization problem, there are several techniques in meta-heuristic based approach. Al-Abdallah and Al-Taani (Al-Abdallah and Al-Taani, 2017) propose the use of Particle Swarm Optimization (PSO) algorithm to extract the summary of single Arabic documents. Jaradat (Jaradat, 2015) incorporate the Harmony Search (HS) algorithm with the summarization process to obtain the summary of Arabic documents. Mosa (Mosa et al., 2017) used ant colony optimization algorithm, coming with a mechanism of local search, to produce a summary of short texts. (Tanfour et al., 2021) used Genetic Algorithm (GA) to select the most important sentences that will be present in our final summary using a specific fitness function that will be maximized to produce a near-optimal summary, using the EASC corpus and the ROUGE toolkit to evaluate the proposed approach. Our contribution falls into the latter two categories.

## 3 PROPOSED METHODOLOGY

Our approach for ATS is to combine GA and LSA techniques to extract a summary that covers all the topics detected in the document. We use the LSA technique to detect topics from the original text. Then, we extract pertinent sentences to create the summary using the GA method, while ensuring that each topic detected by the LSA technique is covered by at least one sentence present in the summary. Our contribution is based on centralized DL techniques (Boughorbel et al., 2018).

Our approach consists of three main steps. text pre-processing, topic detection, and GA optimization.

### 3.1 Pre-processing Step

Pre-processing step includes sentence and word segmentation, stop word removal, and stemming. First, the text is divided into sentences using punctuation marks. Then, sentences are separated into words or

tokens based on white space. The stop-word removal step removes all insignificant words that appear frequently in the document. The stemming step is the process of extracting the root of each word. We use the Tashaphyne stemmer (Zerrouki, 2010) for the Arabic language.

### 3.2 Topic Modeling

Topic modeling is a type of statistical model used to discover hidden topics that occur in a set of documents through machine learning (Mohammed and Alaugby, 2020). This technique analyze huge amount of unlabeled texts to indicate the hidden relationships between items as well as topics. There are several techniques used to get topic models like (Latent Semantic Analysis PLSA, Probabilistic Latent Semantic Analysis LDA etc ...). In this study, we use LSA technique combined with genetic algorithm. Let's give a brief refresher on LSA and GA.

**Latent Semantic Analysis LSA:** is a technique to create vector-based representations of texts which are claimed to extract their semantic content, it extends the vector-based approach by using Singular Value Decomposition (SVD) to reconfigure the data (Dumais, 2004).

The first step extracts  $k$  topics from all the text. Generates a document-term matrix  $A$  of shape  $m \times n$  in which each row represents a document and each column represents a word which has a score. For calculating the scores, LSA uses term frequency-Inverse document frequency TF-IDF to assign a weight for term  $j$  in document  $i$  according to the following formula:

$$W_{i,j} = tf_{i,j} \times \log \frac{N}{df_j} \quad (1)$$

Then, to reduce the dimensions of the above document term matrix to  $k$  (number of desired topics) dimensions, LSA uses truncated Singular Value Decomposition. SVD is a technique in linear algebra that factorizes any matrix  $A$  into the product of 3 separate matrices.

$$A = USV^T \quad (2)$$

Where  $U$  matrix rows represent document vectors expressed in terms of topics,  $V$  matrix rows represent term vectors expressed in terms of topics, and  $S$  is a diagonal matrix that has diagonal elements as singular values of  $A$ .

### 3.3 Genetic Algorithm

Genetic algorithm is an evolutionary metaheuristic for solving discrete optimization problems such as ex-

tractive text summarization. It includes five principal steps: initialization, evaluation using the fitness function, selection, crossover, and mutation.

#### 3.3.1 Initialization

We start by randomly generating an initial population  $P$  composed of  $ps$  chromosomes (individuals); for the automatic text summarization problem, each chromosome presents a summary. It is composed of  $N$  binary genes ( $N$  is the number of sentences of in the document), each gene represents a sentence in the document and can take the value 1 if the sentence is included in the summary or 0 otherwise respecting a maximum length constraint of the summary.

#### 3.3.2 Fitness Function

The fitness function determines how fit an individual is; It gives a fitness score to each individual that represents the probability that an individual will be selected for reproduction. In this work, our aim is to ensure that all the sentences in the summary cover all the topics detected from the original text. The score of a topic is the sum of the similarities between the topic and the sentences of the summary. finally, the fitness of a summary is the sum of the topics scores.

$$Fitness(S) = \sum_{j \in T} score(T_j) \quad (3)$$

where  $T$  is the set of topics obtained by LSA and  $S$  is the summary, i.e. the set of selected sentences.

$$score(T_j) = \sum_{i \in S} score(S_i, T_j) \quad (4)$$

Where  $score(S_i, T_j)$  represent the similarity between sentence  $S_i$  and each topic  $T_j$ . In this research, we adopt the cosine similarity measure, i.e.

$$score(S_i, T_j) = cosine(S_i, T_j) \quad (5)$$

#### 3.3.3 Selection

The selection phase allows one to select the fittest individuals from the current population and let them pass their genes to the next generation.

There are many methods to select individuals from a population. In this work, the father is selected by **Tournament selection**, it consists of selecting  $k$  individuals from the population at random and choosing the best one of these individuals to become a parent. The mother is chosen by **Roulette wheel selection**, where each individual receives a portion of the wheel that is proportional to its fitness function.

### 3.3.4 Crossover

Crossover step allows the creation of new individuals; it exchanges information between two individuals that are selected using the selection operator to generate two other new offspring.

### 3.3.5 Mutation

Mutation operator deals with the change of parts of one solution randomly, which increases the diversity of the population, and it is usually used with a low probability.

## 3.4 Summary Generation

After we execute the genetic algorithm, as a result, we obtain a solution vector  $X_s$  according to the maximum fitness values of the final population. Then, we decode the gene of the winning chromosome that has value equal to 1 to obtain the respective sentence of the document.

## 4 EXPERIMENTAL RESULTS

For the experimental dataset we used ESAC (El-Haj et al., 2010) corpus which contains 153 Arabic articles on different topics (Health, Politic, Education, environment, Art and Music, Science and Technology, Religion, Sport, Tourism and Finance) which are collected from Wikipedia and Arabic newspapers. Each article in this corpus has five different reference summaries, each one is generated by a different human.

### 4.1 Evaluation Metric

We evaluate the quality of our summarizer by ROUGE metric, which stands for Recall-Oriented Understudy for Gisting Evaluation. It assesses the quality of a summary by comparing it to other golden summaries created by humans.

$$ROUGE - N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-grams \in S} Count_{match}(N - gram)}{\sum_{S \in Summ_{ref}} \sum_{N-grams \in S} Count(N - gram)} \quad (6)$$

Where:

- N is the length of n-grams i.e. ROUGE-1: uni-gram metric and ROUGE-2: Bi-gram metric.
- $Count_{match}$  represents the maximum number of the matching N-grams between the reference summary ( $Summ_{ref}$ ) and the generated summary,

- $Count_{N-gram}$  is the total number of n-grams in the reference summary.

To evaluate how accurate our machine generated summaries are, we compute the Precision, Recall and F-measure for these metrics.

ROUGE precision refers to how many candidate summary words are relevant. It is calculated according to the following equation:

$$P = \frac{|grams_{ref} \cap grams_{gen}|}{grams_{gen}} \quad (7)$$

ROUGE recall refers to how many words of the candidate summary are extracted from reference summary. Recall is calculated according to the following equation:

$$R = \frac{|grams_{ref} \cap grams_{gen}|}{grams_{ref}} \quad (8)$$

Where  $grams_{gen}$  are grams of generated summary and  $grams_{ref}$  are the grams of reference summary.

F measure provides the complete information that recall and precision provide separately using the equation below:

$$F_{score} = \frac{2PR}{P+R} \quad (9)$$

### 4.2 Experimental Results

The parameters of our system are: population size **ps**, mutation probability **pm**, number of chromosomes selected by elitism **e**, number of iterations **iter**, number of topics detected **t**, and the weight of the fitness function  $\alpha$ . For evaluation experiments, we used EASC corpus with ps = 100, iter = 1500,  $\alpha=0.6$ , t=10, pm = 0.1, and e = 2.

Results of our proposed method are compared against results of other Arabic summarization related systems. Table 1 lists recall, precision and F-score of these systems.

Table 1: Comparison of the proposed method with other related works and systems.

System	Recall	Precision	F-score
<b>Proposed Method</b>	<b>0.658</b>	0.396	<b>0.455</b>
GA method	0.454	0.262	0.303
Al-Rdaideh	0.465	0.376	0.422
Al-Abdallah	0.449	<b>0.482</b>	0.454

GA method (Tanfour et al., 2021) is based on genetic algorithm to select pertinent sentences to create the final summary, Al-Radaideh (Al-Radaideh and Bataineh, 2018) represented a hybrid approach using domain knowledge and genetic algorithm to extract important points of Arabic documents, and Al-Abdallah (Al-Abdallah and Al-Taani, 2017) is an automatic Arabic text summarization system based on

### Particle Swarm Optimization.

From Table 1, it appears that the proposed method outperforms the other approaches in terms of the recall, precision, and F-score metrics. This is due to the fact that we hardly impose a threshold coverage of each topic. However, it is less performant than Al-Abdallah (Al-Abdallah and Al-Taani, 2017) approach in terms of precision since precision and recall are inversely related.

The advantages of genetic algorithms are the ability to deal with complex combinatorial problems and parallelism, but the main limitations of them are local convergence and the identification of the fitness function.

## 5 CONCLUSION

In this paper, we have addressed the problem of extracting a summary from Arabic texts. We have proposed a combination of latent semantic analysis and a genetic algorithm for this task. Our approach is two-fold. Firstly, we cluster the documents into topics by LSA. Secondly, we apply a Genetic algorithm based optimizer to select the best sentences while ensuring a threshold cover for each topic. The experimental results show that our proposed method achieves state-of-the-art performance in the Arabic document summarization task. A direct extension of our contribution will be the design of a more convenient fitness function to overcome the aforementioned limitations of GA.

## REFERENCES

- Abu Nada, A. M., Alajrami, E., Al-Saqqa, A. A., and Abu-Naser, S. S. (2020). Arabic text summarization using arabert model using extractive text summarization approach.
- Al-Abdallah, R. Z. and Al-Taani, A. T. (2017). Arabic single-document text summarization using particle swarm optimization algorithm. *Procedia Computer Science*, 117:30–37.
- Al-Hashemi, R. (2010). Text summarization extraction system (tses) using extracted keywords. *Int. Arab. J. e Technol.*, 1(4):164–168.
- AL-Khawaldeh, F. T. and Samawi, V. W. (2015). Lexical cohesion and entailment based segmentation for arabic text summarization (lceas). *World of Computer Science & Information Technology Journal*, 5(3).
- Al-Radaideh, Q. A. and Bataineh, D. Q. (2018). A hybrid approach for arabic text summarization using domain knowledge and genetic algorithms. *Cognitive Computation*, 10(4):651–669.
- Belkebir, R. and Guessoum, A. (2015). A supervised approach to arabic text summarization using adaboost. In *New contributions in information systems and technologies*, pages 227–236. Springer.
- Boudabous, M. M., Maaloul, M. H., and Belguith, L. H. (2010). Digital learning for summarizing arabic documents. In *International Conference on Natural Language Processing*, pages 79–84. Springer.
- Boughorbel, S., Jarray, F., Venugopal, N., and Elhadi, H. (2018). Alternating loss correction for preterm-birth prediction from ehr data with noisy labels. *arXiv preprint arXiv:1811.09782*.
- Chouikhi, H., Chniter, H., and Jarray, F. (2021). Arabic sentiment analysis using bert model. In *International Conference on Computational Collective Intelligence*, pages 621–632. Springer.
- Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.
- El-Haj, M., Kruschwitz, U., and Fox, C. (2010). Using mechanical turk to create a corpus of arabic summaries.
- Froud, H., Lachkar, A., and Ouatik, S. A. (2013). Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering. *arXiv preprint arXiv:1302.1612*.
- Haboush, A., Al-Zoubi, M., Momani, A., and Tarazi, M. (2012). Arabic text summarization model using clustering techniques. *World of Computer Science and Information Technology Journal (WCSIT) ISSN*, pages 2221–0741.
- Jaradat, Y. A. (2015). *Arabic Single-Document Text Summarization Based on Harmony Search*. PhD thesis, Yarmouk University.
- Mohamed, M. and Oussalah, M. (2019). Srl-esa-textsum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing & Management*, 56(4):1356–1372.
- Mohammed, S. H. and Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science*, 19(1):353–362.
- Mosa, M. A., Hamouda, A., and Marei, M. (2017). Graph coloring and aco based summarization for social networks. *Expert Systems with Applications*, 74:115–126.
- Tanfouri, I., Tlik, G., and Jarray, F. (2021). An automatic arabic text summarization system based on genetic algorithms. *Procedia Computer Science*, 189:195–202.
- Zerrouki, T. (2010). Tashaphyne, arabic light stemmer.