*Computers, Materials & Continua*
DOI:10.32604/cmc.2021.014361
***Article***

Tech Science Press

# Automatic Persian Text Summarization Using Linguistic Features from Text Structure Analysis

**Ebrahim Heidary[1], Hamïd Parvïn[2,3,4,*], Samad Nejatian[5,6], Karamollah Bagherifard[1,6] and Vahideh Rezaie[6,7]**

[1]Department of Computer Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran
[2]Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam
[3]Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Vietnam
[4]Department of Computer Science, Nourabad Mamasani Branch, Islamic Azad University, Mamasani, Iran
[5]Department of Electrical Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran
[6]Young Researchers and Elite Club, Yasooj Branch, Islamic Azad University, Yasooj, Iran
[7]Department of Mathematics, Yasooj Branch, Islamic Azad University, Yasooj, Iran
[*]Corresponding Author: Hamid Parvin. Email: parvin@iust.ac.ir

Received: 16 September 2020; Accepted: 03 February 2021

**Abstract:** With the remarkable growth of textual data sources in recent years, easy, fast, and accurate text processing has become a challenge with significant payoffs. Automatic text summarization is the process of compressing text documents into shorter summaries for easier review of its core contents, which must be done without losing important features and information. This paper introduces a new hybrid method for extractive text summarization with feature selection based on text structure. The major advantage of the proposed summarization method over previous systems is the modeling of text structure and relationship between entities in the input text, which improves the sentence feature selection process and leads to the generation of unambiguous, concise, consistent, and coherent summaries. The paper also presents the results of the evaluation of the proposed method based on precision and recall criteria. It is shown that the method produces summaries consisting of chains of sentences with the aforementioned characteristics from the original text.

**Keywords:** Natural language processing; extractive summarization; linguistic feature; text structure analysis

## 1 Introduction

With the massive volume of digital text data generated every day, fast and accurate retrieval of valuable information from texts has become an increasingly worthwhile challenge. One approach to tackling this challenge is automatic text summarization.

Text summarization is the process of revising a text to make it shorter than the original document while retaining important words and sentences and without losing its main content and information [1]. Automatic text summarization means using machine or computer-based tools to produce a useful summary. Although the first automatic text summarization solutions were introduced in the 1950s [2,3], summarization has long been and still is one of the main challenges of natural language processing. Computer-generated summaries are often different from those produced by humans, because it is very difficult for machines to gain a deep understanding of the content of a text based on its syntactic and semantic structure as humans do [4].

Summarization systems can be classified based on the type of input, output, purpose, language, and method of summarization (See Fig. 1). In terms of the type of input document, summarization systems are divided into two groups: single-document and multi-document. Depending on whether the input is simple text, news article, science article, etc., the purpose of a summarization system could be to produce up-to-date information, run queries, or inform users about a particular subject. The purpose of summarization is often an important determinant of the method of summarization [5]. Summarization methods can be classified into two categories: extractive and abstractive.

Extractive summarization involves selecting a set of sentences or phrases in the text based on the scores they earn according to a given criterion and copying them into the summary without any change. Abstractive summarization means producing a brief interpretation of the original text. In this method, sentences of the summary may not necessarily be written in the same way as in the original text. Summarization systems can also be classified based on whether they are built to produce educational or informative outputs.

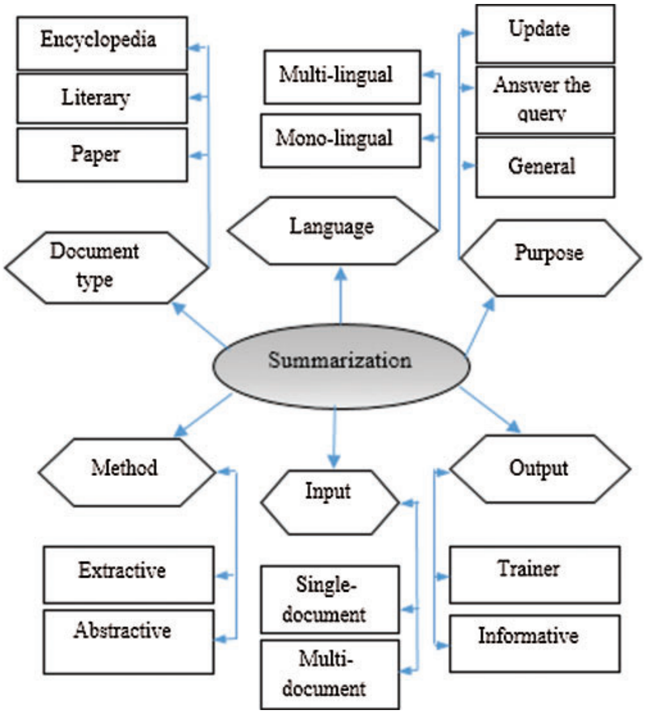

**Figure 1:** Classification of summarizers [6]

The majority of existing automatic summarization systems use the extractive summarization method. Extractive summarization can be done with three approaches: statistical approach, linguistic approach, and combined approach [4]:

**Statistical approach:** In this approach, summarization depends on the statistical distribution of features of interest and quantitative characteristics of the text. This approach involves using information retrieval and classification techniques without trying to understand the whole document. In this method, an information retrieval method analyzes the position, length, and frequency of words and sentences in the document and a classifier determines which sentences could be part of the summary based on a set of instances on which it is trained. In this method, the sentences of the original text are extracted without attention to the meaning of the words.

**Linguistic approach:** In this approach, the computer needs to have a deep knowledge of the language it is processing to the extent that it can analyze and interpret sentences and then decide which phrases should be included in the summary. In this method, the relationships between words and phrases in the document are identified through part-of-speech tagging, grammatical analysis, lexical analysis, and extraction of meaningful phrases. The parameters of these processes could be sign words, features, nouns, and verbs. While the statistical approach tends to be more computationally efficient, the linguistic approach often produces better summaries as it factors in the semantic relationships in the original text.

**Combined approach:** This approach involves using a combination of both statistical and linguistic methods to generate more concise and meaningful summaries.

Statistical summarization methods only use statistical features, which while making them quite simple and agile, also makes them more susceptible to incoherence and inconsistency in generated summaries.

The combined use of multiple extractive summarization techniques can indeed be very effective in improving the quality of produced summaries. In this study, the combined approach to summarization is used to produce unambiguous, concise, consistent, and coherent summaries based on the linguistic features taken from the text structure analysis, modeling of the text structure and the relationship between its entities, and an improved single-document feature selection process.

In the rest of this article, the second section provides a review of text summarization methods and systems developed for Persian and other languages, the third section describes the proposed method, the fourth section presents and discusses the results of implementing the method, and the fifth section presents the conclusions and offers a few suggestions for future works.

## 2 Review of Literature

The concept of automatic text summarization was first introduced by Luhn in 1958, in the sense of determining the distribution of words in sentences and identifying the keywords of a document. Since then, a variety of summarization methods have been developed based on different approaches and for different purposes. However, most of these methods can be described as improvements upon previous techniques.

### *2.1 Summarization Methods for Persian Texts*

In [2], researchers have proposed a method for improving the quality of automatic single-document Persian text summarization. In this method, first, a combination of natural language processing methods and correlation graphs is used to compute an importance factor for each grammatical unit in the entire text with the help of the Term Frequency-Inverse Document Frequency (TF-IDF) method. Then, a sentence feature vector is constructed based on four features: the degree of similarity of each sentence to the title, the degree of similarity of each sentence to keywords, the degree of similarity of two sentences with each other, and the position of each sentence. Finally, the constructed similarity graphs are used to identify and select the sentences with the highest external similarity and the lowest internal similarity for inclusion in the summary text.

The automatic Persian text summarization system presented in [6] has introduced a new method for summarizing Persian news texts using the knowledge contained in FarsiNet. In this method, sentences are clustered based on how similar or related they are into three categories of similar, related, and co-occurring. The sentences to be included in the summary are then selected accordingly to reach the lowest possible redundancy and the highest possible relevance. In this method, the use of co-occurring clusters reduces the ambiguity of the summary text. This method scores the sentences based on eight features: sentence length, paragraph position, demonstrative pronouns, title words, keywords, word weight, specific nouns, and general nouns. In [7], researchers at the Ferdowsi University of Mashhad introduced the Ijaz system for single-document summarization of Persian news texts. This system uses a set of factors including the degree of similarity with the context, stop words, sentence length, sentence position, important phrases, pronouns, demonstrative pronouns and determining phrases in sentences, marked phrases, and similarity to title to rate the sentences in terms of importance. It then uses the linear combination of these factors to compute a final score for each sentence. The PSO-based extractive Persian text summarization method of [8] extracts all sentences of the input document and identifies the candidate words of each sentence and then generates and stores a context vector for each candidate. Using this vector, it quantifies the similarity of each two sentences and stores it in a similarity matrix. Finally, it uses a clustering algorithm to classify the sentences and picks the most important sentence within each class based on the calculated scores. In [9], a neural network technique was used to investigate the parameters that may influence the performance of summarization systems.

In the summarization method of this study, first, the paragraphs are scored and then a neural network is used to compute sentence-based scores only for those paragraphs that score higher than a certain threshold level. The neural network used in this study is a three-layered feed-forward neural network with 9 input nodes, 6 hidden nodes, and 1 output node. The output of this network determines whether or not a sentence should be reflected in the summary text. According to the author, this approach reduces the processing volume, increases the speed of summary generation, and improves the performance of existing systems. In [10], an automatic Persian text summarization system has been developed based on Latent Semantic Analysis (LSA). Using a rich vocabulary, this system performs stemming with higher precision and higher efficiency than similar systems. Thanks to this rich vocabulary, the LSA process of this system can properly determine the set of synonyms and lexical chains and the semantic links between words and sentences. In [11], researchers have developed an extractive Persian text summarization method based on an anthropological approach. Inspired by the human way of thinking, this method involves creating a summary by building a chain of sentences that are more correlated and related to each other, rather than just the most important sentences. In this method, the input text is divided into paragraphs and the text of each paragraph is divided into a set of sentences. Once the relationships between all sentences in each paragraph are determined, the system attempts to find a chain of sentences with the strongest connection to each other. The resulting chain will be the summary of the original text. The FarsiSum automatic Persian text summarizer system introduced in [12] is a modified version of the SweSum summarizer for Swedish [13]. This system receives the input texts in HTML format. In the graph-based Persian text summarization algorithm of [14], the graph theory is used to choose which sentences of the input document should be a candidate for inclusion in the summary. In this algorithm, the nodes and edges of a graph constructed for the text are weighted based on different criteria and then the final weight of each sentence is determined by combining these values. The final weight reflects the importance of the sentence and the likelihood that it will appear in the final summary. The FarsiSum summarizer is a statistical method and gives higher scores to first sentences. The Persian text summarizer proposed in [15] is an extractive summarization method operating based on the graph theory and lexical chains.

### *2.2 Summarization Methods for Other Documents*

In [16], the vector space model has been used to develop an abstractive automatic summarization system for online debate texts. This system consists of three modules: point extraction, point curation, and summary generation. Point extraction is done by dependency parsing and analyzing syntactic structure. After selecting the topic points and the points that could be suitable for the summary, shorter points are generated by smaller indirect points. In [17], an extractive summarization method has been developed for Arabic texts. This method uses a combination of semantic information extracted from the Arabic word net and rhetorical structure theory (RST), which is one of the most widely used theories in natural language processing. In this method, a combination of linguistic selection methods and sentence feature selection methods is used to improve the quality of Arabic text summarization. The proposed RST-based method first generates an initial summary and then uses the score of each sentence in this summary to investigate the similarity of sentences with the main title and subheadings. The automatic Indonesian text summarization system of [18] uses a combination of sentence scoring and decision tree for summary generation. In this system, the C4.5 algorithm is used to select the sentences of interest. Then, a sentence scoring method is used to weight each sentence based on 8 features, including TF-IDF, uppercase letters, proper nouns, cue phrases, numerical data, sentence length, sentence position, and similarity to title. Next, a decision tree model is generated based on the training data, and finally, the resulting rules are used to determine important sentences and generate the summary accordingly. In [4], an extractive summarization method based on a combined statistical-linguistic approach has been proposed for Indian texts.

This summarization system consists of three main stages: preprocessing, sentence feature extraction, and genetic algorithm (GA) for ranking sentences based on optimized feature weights. Each sentence is represented by a sentence feature vector. For each sentence, the statistical-linguistic features are examined and a score is produced based on the weight of features in that sentence. The results are then used to rank the sentences. Sentence features can take values between zero and one. In the GA, the fittest chromosome is selected after a certain number of generations, and then the distance between each sentence score and the fittest chromosome is measured by the Euclidean distance formula. Sentences are then sorted in ascending order of this distance. Finally, a summary is produced by extracting a certain number of highest ranked sentences from the document depending on the intended degree of summarization. The extractive summarization method proposed in [19] uses a Hidden Markov Model (HMM) part-of-speech tagging technique for summary generation. Part-of-Speech (POS) tagging is an automatic machine learning process for tagging each word in a sentence based on verbs, nouns, adjectives, and other typical components of natural languages. In the method of [20], summarization is performed by feature selection based on GA and a probabilistic technique. This method considers five features:

similarity to title, sentence length, sentence position, numbers, and pseudowords. Depending on the number of features used, each chromosome is made of up to five genes, each representing a binary format feature. The system proposed in [21,22], which is called QUESTS, is an integrated query system for generating extractive summaries from a set of documents. This system draws an integrated graph of the relationships between the sentences of all input documents and then uses the found relationships to derive multiple subgraphs from the main graph. These subgraphs consist of sentences that are higher related to the query and to each other. The system then ranks the subgraphs based on a scoring model and selects the highest-ranked subgraph that is most relevant to the query for inclusion in the summary.

## 3 Proposed Method

This paper presents a new combined extractive single-document text summarization method based on text structure. The proposed summarization process consists of three stages: preprocessing, feature selection, and summary generation. The architecture of the proposed method is described in the following Fig. 2.

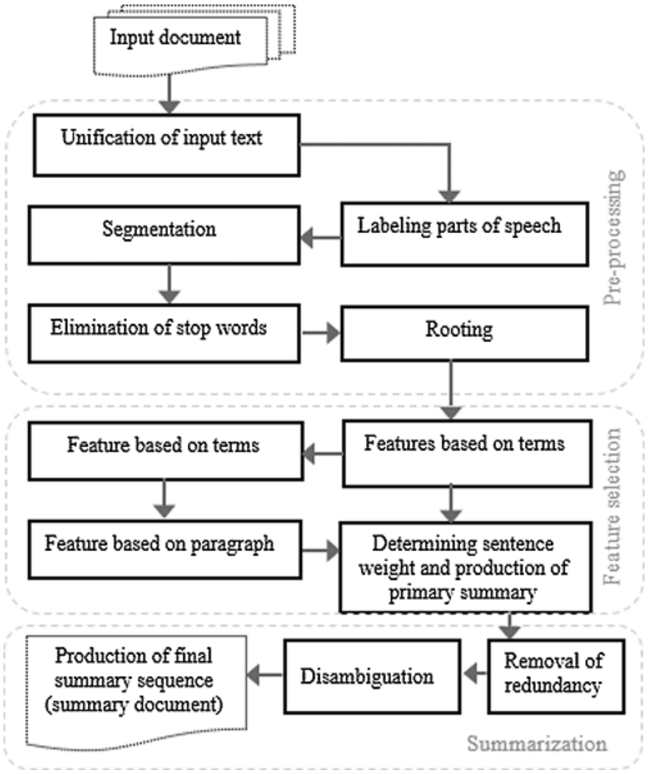
Input document
Unification of input text
Labeling parts of speech
Segmentation
Elimination of stop words
Rooting
Pre-processing
Feature based on terms
Features based on terms
Feature based on paragraph
Determining sentence weight and production of primary summary
Feature selection
Production of final summary sequence (summary document)
Disambiguation
Removal of redundancy
Summarization

**Figure 2:** Architecture of the proposed summarization method

The idea used in the proposed method and the improvement made in the feature selection stage based on the text structure greatly reduce issues such as inconsistency, ambiguity, and redundancy in the summary text.

### *3.1 Preprocessing*

Before starting the summarization process, it is necessary to convert the original text into a single form that is suitable for the summarization operation. The pre-processing stage generally consists of normalization, tokenization, POS tagging, stemming/lemmatization, and stop word removal, and is done more or less the same way in all languages. In this study, pre-processing is performed with the help of the ParsiPardaz tool, which has been developed by the Iranian National Cyberspace Research Institute (Telecommunication Research Center of Iran) for the Persian language. This tool has shown to have 98% accuracy in tagging and 100% accuracy in normalization [11].

#### *3.1.1 Normalization*

The first step in pre-processing is to standardize the input text. One of the common problems in Persian texts is the variety of ways that some words and letters can be written. For example, all words that contain “ی” or “ي” should be changed into a single format. This normalization is done at the start of the pre-processing stage to avoid problems in the subsequent stages.

#### *3.1.2 POS Tagging*

After the normalization stage, the role of words (e.g. noun, verb, adjective, conjunction, etc.) in sentences should be determined for use in the next stages.

#### *3.1.3 Tokenization*

This step involves identifying and separating the words and sentences of the input text based on the characters that signify the end of a word or sentence. In this step, conjunctions are also treated as the boundaries of sentences.

#### *3.1.4 Stop Word Removal*

Stop words are the words that are commonly used in texts but are not descriptive, do not depend on the subject, and also do not have a semantic role. Examples include conjunctions, linking verbs, pronouns, adpositions, and adverbs.

#### *3.1.5 Stemming*

After removing the stop words, the stem of the words that have remained in the text must be determined.

## 4 Feature Selection

In this stage, the input text undergoes a feature analysis. Since the most important words in the text that arrives at this stage are nouns and adjectives, here, the main features of the text are weighted based on the features identified for important terms, sentences, and paragraphs. In the feature selection stage, the main features of the input document are scored in the following three phases.

TBF → SBF → PBF = WFS

**TBF:** Term-based Features

**SBF:** Sentence-based Features

**PBF:** Paragraph-based Features

**WFS:** Weight Features Selection Sentence

### *4.1 Term-Based Features*

Each sentence in the text consists of a number of terms that play an important role in the meaning of that sentence. Each of these terms has a number of features that can affect the importance of that sentence.

#### *4.1.1 TS-ISF*

This parameter indicates the importance of a term in a sentence. TF-ISF of each term is calculated as follows Eqs. (1) and (2):

$$\mathrm{ISF_w} = \log \frac{\mathrm{N}}{\mathrm{n_w}} \tag{1}$$

$$\mathrm{TF_{w \cdot fs}} = \text{The number of times the word wis repeated in the sentence s} \tag{2}$$

In the above equation, N is the total number of sentences in the input document and $n_w$ is the number of sentences that contain the word w. Accordingly, the importance of the words of the sentence s is calculated as follows Eq. (3):

$$\mathrm{TBF_!} = \mathrm{Max_w 1} \rightarrow \ \mathrm{Ns}\{\mathrm{TF_w} \cdot \mathrm{ISF_w}\} \tag{3}$$

In this equation, Ns is the number of words in the sentence s.

#### *4.1.2 Positive and Negative Terms*

Here, positive terms refer to terms and phrases like “therefore,” “concluded” “in conclusion,” “in the end,” “found,” etc. the presence of which signifies the importance of the sentence. In contrast, negative terms refer to terms and phrases like “stated,” “in other words,” “as an example,” etc. that are indicative of the insignificance of the sentence. Based on this definition, sentences are scored as follows Eq. (4):

$$\mathrm{TBF_2} = \begin{cases} 1, & \text{If exist pos word} \\ 0, & \text{If not exist pos and neg word} \\ -1, & \text{If exist negetive word} \end{cases} \tag{4}$$

#### *4.1.3 Similarity to Title Words and Keywords*

If a sentence contains a word of the title or a keyword, it is more likely to be more important than other sentences. Keywords are usually specified by the user. Therefore, the effect of similarity to title words and keywords is calculated by the following Eq. (5):

$$\mathrm{TBF_3} = \frac{|\mathrm{Sentwords_s} \cap \mathrm{Titlewords}|}{|\mathrm{Sentwords_s}| + |\mathrm{Titlewords}| - |\mathrm{Sentwords_s} \cap \mathrm{Titlewords}|} \tag{5}$$

#### *4.1.4 Numerical Values*

Numerical values within sentences usually contain important information. Thus, the effect of this feature in the sentence is quantified by the following Eq. (6):

$$\mathrm{TBF}_4 = \frac{|\mathrm{Sentwords}_\mathrm{s}\ \text{is numerical data}|}{|\mathrm{N}_\mathrm{s}|} \tag{6}$$

#### *4.1.5 Emphasized Words*

Sometimes, one or more words in a sentence are emphasized by the use of quotation marks, italic, bold or underlined formatting, etc., which indicate the importance of that word and therefore the importance of that sentence. The effect of such words on the importance of the sentence is determined by the following Eq. (7):

$$\mathrm{TBF}_5 = \frac{|\mathrm{Sentwords}_\mathrm{s}\ \text{is Important}|}{|\mathrm{N}_\mathrm{s}|} \tag{7}$$

#### *4.1.6 Stop Words and Insignificant Words*

Some sentences contain stop words that play no significant semantic role in the sentence. In sentences where such words are used excessively, it may indicate the lower importance of the sentence. See Eq. (8):

$$\mathrm{TBF}_6 = \frac{|\mathrm{N}_\mathrm{s}| - |\mathrm{Sentwords}_\mathrm{s}\ \text{is StopWord}|}{|\mathrm{N}_\mathrm{s}|} \tag{8}$$

#### *4.1.7 Proper Nouns*

Proper names include the names of people, places, and times. The presence of a proper noun in a sentence makes it more important. Therefore, the effect of these nouns is determined based on the following Eq. (9):

$$\mathrm{TBF}_7 = \frac{|\mathrm{Sentwords}_\mathrm{s}\ \text{is Proper nouns}|}{|\mathrm{N}_\mathrm{s}|} \tag{9}$$

#### *4.1.8 Marked Phrases*

In some sentences, certain parts are marked with quotation marks or other characters (e.g., { }, [ ] “ ”, «») to highlight the importance of that phrase [15]. The effect of these marked sections in a sentence is quantified by the following Eq. (10):

$$\mathrm{TBF}_8 = \frac{|\mathrm{Sentwords}_\mathrm{s}\ \text{is Certain Phrases}|}{|\mathrm{N}_\mathrm{s}|} \tag{10}$$

### ***4.2 Sentence-Based Features***

Besides term-based features, sentence-based features are the other key parameters of extractive summarization. The purpose of these features is to ensure the proper selection of sentences with high importance rating. Each sentence in the input document has a rank that is determined based on the weight of its features. The most important features that are considered for each sentence are described below.

#### *4.2.1 Sentence Length*

Excessively short or long sentences are usually not suitable for a summary. This is because excessively short sentences tend to convey little meaning and very long sentences make the summary longer than required. Thus, the following Eq. (11) is used to take this into consideration:

$$\mathrm{SBF}_1 = \frac{|\mathrm{Sentwords}_\mathrm{S}|}{\mathrm{Max}\ |\mathrm{Sentwords}\ |} \tag{11}$$

The denominator of this fraction is the number of words in the longest sentence of the document.

#### *4.2.2 Sentence Position*

An important feature of a sentence is its position in the paragraph. In paragraphs made of declarative sentences, usually, the beginning and end sentences are more important than others, and should therefore be given a higher weight. The importance of sentences, as indicated by their position in the paragraph, is quantified as follows Eq. (12):

$$\mathrm{SBF}_2 = \mathrm{Max}\left(\frac{1}{\mathrm{Position}_\mathrm{S}}, \frac{1}{|\mathrm{TotalS}_\mathrm{P}| - \mathrm{Position}_\mathrm{S}}\right) \tag{12}$$

In the above equation, $Positon_s$ is the position of the sentence in the paragraph, and $TotalS_p$ is the total number of sentences in the paragraph.

#### *4.2.3 Similarity to Beginning and End Sentences of the Paragraph*

Given the high importance of the beginning and end sentences of each paragraph, another feature that can benefit the selection of high-value sentences for a summary is the degree of similarity to these sentences. See Eq. (13):

$$\mathrm{SimP}_\mathrm{End} = \frac{|\mathrm{Sentwords}_\mathrm{S} \cap \mathrm{EndSentwords}|}{|\mathrm{Sentwords}_\mathrm{S}| + |\mathrm{EndSentwords}| - |\mathrm{Sentwords}_\mathrm{S} \cap \mathrm{EndSentwords}|} \tag{13}$$

#### *4.2.4 Sentence Centrality*

Sentence centrality is the degree to which the keywords of a sentence overlap with other sentences in the paragraph. The greater the overlap is, the more important and valuable that sentence will be for the text. See Eq. (14):

$$\mathrm{SBF}_4 = \frac{|\mathrm{SentKeywords}_\mathrm{S} \cap \mathrm{PKeywords}|}{|\mathrm{totalP}\ _\mathrm{Keyword}|} \tag{14}$$

In the above equation, the numerator of the fraction is the number of keywords in the sentence that are also present elsewhere in the paragraph, and the denominator is the total number of keywords in the paragraph.

### ***4.3 Paragraph-Based Features***

In addition to term-based and sentence-based features, the feature of paragraphs that consist of sequences of sentences can also benefit the selection of sentences for the summary. The parameters used for this purpose are described in the following.

#### *4.3.1 Paragraph Position*

In most texts, including news documents (depending on the writing style), the beginning and end paragraphs tend to convey more important information. Thus:

$$PBF_1 = Max\left(\frac{1}{Position_P}, \frac{1}{|TotalP_D| - Position_P + 1}\right) \tag{15}$$

In this Eq. (15), $Position_P$ is the position of the paragraph, and Total $P_D$ is the total number of paragraphs in the input document.

#### *4.3.2 Paragraph Centrality*

In a typical document, some paragraphs have more references to the main topic and carry more information about the subject. Therefore, using this feature can improve the quality of the texts to be selected for summary generation.

$$\mathrm{PBF}_2 = \frac{|\mathrm{PKeywords}_\mathrm{D} \cap \mathrm{DKeywords}|}{|\mathrm{totalD}\ _{\mathrm{Keyword}}|} \tag{16}$$

In this Eq. (16), the numerator is the number of keywords in the paragraph that are also present elsewhere in the text, and the denominator is the total number of keywords in the text.

#### *4.3.3 Similarity to Title and Keywords*

A paragraph that contains a large number of title words and keywords often conveys critical information about the document and can benefit the summarization.

$$\mathrm{PBF}_3 = \frac{|\mathrm{Pwords}_\mathrm{p} \cap \mathrm{Titlewords}|}{|\mathrm{Pwords}_\mathrm{p}| + |\mathrm{Titlewords}| - |\mathrm{Pwords}_\mathrm{p} \cap \mathrm{Titlewords}|} \tag{17}$$

The Eq. (17) represents the impact of the similarity of the content of the paragraph to keywords and title words on the summary document.

#### *4.3.4 Important Signs*

In some documents, the paragraphs that are supposed to grab the reader's attention are marked by bullet points, numbered lists, and multilevel lists. Therefore, the presence of these signs can indicate importance. See Eq. (18):

$$\mathrm{PBF}_4 = \begin{cases} 1, & \text{if Exit Important Signs} \\ 0, & \text{if NOT Exit Important Signs} \end{cases} \tag{18}$$

## 5 Summarization

### *5.1 Sentence Weighting and Initial Summary Chain Generation*

The previous section described the parameters used in summarization in the three main groups of term-based features, sentence-based features, and paragraph-based features. These features are key determinants of whether a sentence will be included in the summary text. The next step is to determine the weight of each sentence based on a combination of these parameters. This weight indicates the importance of each sentence for the text. The weight of sentence i from paragraph j is defined based on the linear combination of the described criteria as shown below Eq. (19):

$$WS_{i,j} = \sum VBF_{i,j} * \sum SBF_{i,j} * \sum PBF_j \tag{19}$$

After calculating the weight of every sentence in the input document, the sentences are sorted in descending order of their weight. Then, an initial chain of sentences is produced by taking a certain number of sentences from the top of this list depending on the desired degree of summarization (compression).

### *5.2 Final Summary Chain Generation*

An optimal summary is one that is concise and unambiguous and consists of key sentences. To reach such a summary, the initial summary produced in the previous step is refined through two processes: redundancy elimination and ambiguity elimination. These processes are described below.

#### *5.2.1 Redundancy Elimination*

To avoid including similar sentences in the summary, every two sentences of the summary text are compared with each other. Upon finding two sentences with over 75% overlap, the loner sentence will be removed from the summary text.

#### *5.2.2 Ambiguity Elimination*

The presence of pronouns with unclear antecedents can make the sentences misleading or ambiguous, and thus make the summary difficult to comprehend. Also, the antecedents of a pronoun can be in a previous sentence that is not included in the summary. To avoid this problem, the sentence before the selected ambiguous sentence in the original text is also used in the summary text.

## 6 Implementation and Results

The proposed method was implemented on 5000 texts taken from news websites, the topics of which are listed in Tab. 1. The results were then evaluated using standard criteria.

**Table 1:** Specifications of the selected news articles

| News subjects | Number |
|---|---|
| Sports | 1000 |
| Economic | 1000 |
| Political | 1000 |
| Scientific | 1000 |
| Others | 1000 |
| Total number of news | 5000 |
| Total number of sentences | 214600 |
| Total number of terms | 3565200 |

The preprocessing operation starts with normalization, which is followed by POS tagging to determine the role of words, and then tokenization to determine the boundaries of words and sentences. The other two operations of the preprocessing stage are the stop word removal and stemming. The tool used for preprocessing in this study was the ParsiPardaz tool developed by the Iranian National Cyberspace Research Institute (Telecommunication Research Center of Iran). One of the great features of this tool is its high accuracy in tagging the words of input documents.

Noun and adjective are the most important words in the input text. After the preprocessing operation, the method described in the previous section was used to build the initial summary based on term-based, sentence-based, and paragraph-based features. Finally, ambiguity and redundancy elimination was performed to turn this initial summary into the final summary text.

The summarization performance was measured by precision and recall, which are among the criteria most commonly used for this purpose. These criteria are defined as follows Eqs. (20) and (21):

$$\mathrm{Precision\,(P)} = \frac{\left|\mathrm{Sum_r}\ \cap\ \mathrm{Sum_s}\right|}{\left|\mathrm{Sum_s}\right|} \tag{20}$$

$$\mathrm{Recall\,(R)} = \frac{\left|\mathrm{Sum_r}\ \cap\ \mathrm{Sum_s}\right|}{\left|\mathrm{Sum_r}\right|} \tag{21}$$

In the above equations, $Sum_r$ is the total number of sentences in human-generated summaries and $Sum_s$ is the total number of sentences in the summaries produced by the proposed summarization system. For this evaluation, a combined recall-precision measure was also calculated as follows Eq. (22):

$$\mathrm{E_{p,r}} = \frac{2\mathrm{pr}}{\mathrm{p}+\mathrm{r}} \tag{22}$$

The results of the evaluation of the proposed method based on the above criteria are presented in Tab. 2.

**Table 2:** Comparison of the proposed method with existing methods

| Method | Compression basis = 30% | | Compression basis = 40% | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| SweSum | 0.66 | 0.68 | 0.73 | 0.77 |
| Proposed method | 0.76 | 0.79 | 0.81 | 0.81 |

As the results of the above table show, the proposed summarization method is more accurate than the SweSum method for similar data sets (See Figs. 3 and 4). Another major advantage of the proposed summary method is the elimination of redundancy by removing similar sentences from the summary and also the elimination of ambiguity by inserting the sentences on which ambiguous sentences are dependent in the summary text.

Another evaluation criterion was the average number of sentences that were present in human-generated summaries as well as those produced by the proposed method. Fig. 5 shows the results of this evaluation, which was conducted with the help of five experts.

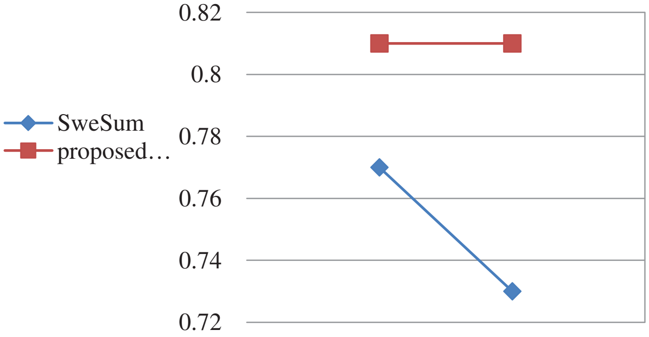

**Figure 3:** Comparison of the proposed method with existing methods for 30% compression

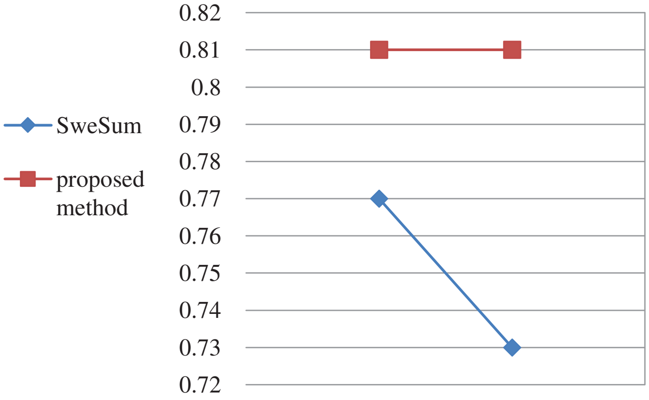

**Figure 4:** Comparison of the proposed method with existing methods for 40% compression

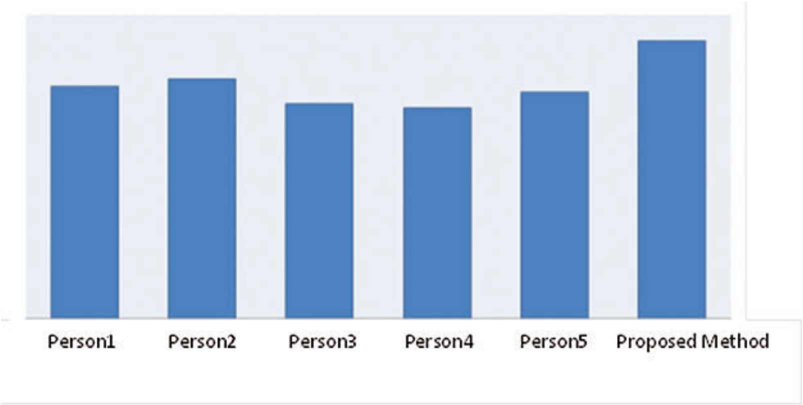

**Figure 5:** Average number of sentences shared between the summaries of the proposed method and those produced by five experts

Evaluation of the length of the produced summaries showed that the shortest document consisted of 10 sentences with 70 words and its summary contained 5 sentences with 39 words. For the longest document, which consisted of 50 sentences with 200 words, the summary contained 41 sentences with 147 words.

To evaluate the performance of the proposed method, readers' satisfaction with the summaries produced for different topics was also investigated. The results of this investigation are presented in Fig. 6.

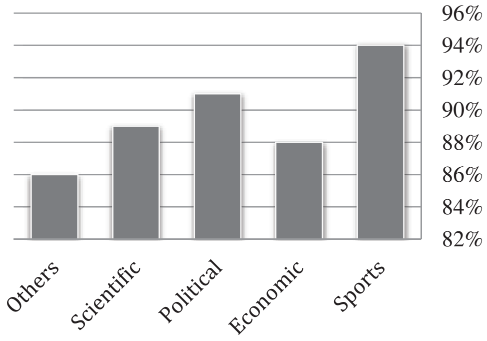

**Figure 6:** Readers' satisfaction with the summaries produced by the proposed method

This evaluation showed that on average, 89% of readers were satisfied with the summaries produced by the proposed method, which indicates the effectiveness and good accuracy of the proposed method in producing the summaries desired by users.

## 7 Conclusion

One of the main advantages of the summarization method proposed in this paper over similar methods is the combined use of statistical and linguistic methods to model the text structure and the examination of the relationship between entities of the input document in order to improve the sentence feature selection process and produce an unambiguous, concise, consistent, and coherent summary.

The proposed summarization method consists of three stages: preprocessing, feature selection, and summary generation. In the feature selection stage of this method, the main features of the input document are captured in three phases based on term-based features, sentence-based features, and paragraph-based features.

Comparing the performance of the proposed method with a similar method on a series of news texts showed that with 78.5% precision and 80%, recall, the proposed method outperformed the other method in this respect. Also, readers' satisfaction with the summaries produced by the method was 89%. The idea used in the proposed method fairly reduces issues such as incoherence, ambiguity, and redundancy in the summary text. To improve the proposed summarization method in future studies, it may be possible to use ontology techniques to create stronger lexical chains and semantic links between different entities in the input document.

**Funding Statement:** The author(s) received no specific funding for this study..

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] C. Salmet, A. Atmadja, D. Maylawati, R. Lestari, W. Darmalaksana *et al.,* "Automated text summarization for indonesian article using vector space model," *2nd Annual Applied Science and Engineering Conf.*, vol. 288, no. 1, pp. 1–6, 2017.

[2] T. Hosseinikhah, A. Ahmadi and A. Mohebi, "A new Persian text summarization approach based on natural language processing and graph similarity," *Iranian Journal of Information Processing Management*, vol. 33, no. 2, pp. 885–914, 2018.

[3] M. G. Ozsoy, F. N. Alpaslan and L. Cicekli, "Text summarization using latent semantic," *Journal of Information Science*, vol. 37, no. 4, pp. 405–417, 2011.

[4] C. Thaokar and L. Malik, "Test model for summarization Hindi text using extraction method," *IEEE Conf.*, vol. 7, no. 3, pp. 1138–1143, 2013.

[5] L. Talibali and N. Riahi, "An overview of automatic text summarization techniques," *Int. Conf. on Applied Research in Information Technology, Computer and Telecommunications*, vol. 28, pp. 75–84, 2015.

[6] S. Lagirini, M. Redjimi and N. Azizi, "Automatic Arabic text summarization approaches," *International Journal of Computer Applications*, vol. 164, no. 5, pp. 31–37, 2017.

[7] A. Pourmasoomi, M. kahani, S. A. Toosi and A. Estiri, "Ijaz: An operational system for single-document summarization of Persian news texts," *Journal Signal and Data Processing*, vol. 11, no. 1, pp. 33–48, 2014.

[8] M. Bazghandi, G. H. Tedin-Tabrizi, M. Wafai-Jahan and A. Bazghandi, "Selective synopsis of PSO clustering coherent texts," in *The First Int. Conf. on Line Processing and Farsi language*, Iran: Semnan University, 2012.

[9] F. Mohammadian, M. Nematbakhsh and A. Naghshenilchi, "Summary of Persian texts using a meaning-based method," *Second National Conf. on Soft Computing and Information Technology*, vol. 7, no. 7, pp. 9–16, 2011.

[10] H. Sotoudeh, M. Akbarzadeh-Totouchi and M. Teshnelab, "Summary of text based on selection using an anthropological approach," *18th Iranian Conf. on Electrical Engineering*, vol. 1, pp. 2266–2227, 2010.

[11] N. Mazdak and M. Hassel, "FarsiSum-A Persian text summarization," Sweden: Stockholm University, Department of linguistics, Master Thesis, 2004.

[12] H. Dalianis, "SweSum—A text summarizer for Swedish," Sweden, Technical report in interaction and Presentation Laboratory, pp. 1–15, 2000.

[13] H. Shakeri, S. Gholamrezazadeh, M. Amini-Salehi and F. Ghadmyari, "A new graph-based algorithm for Persian text summarization," *Springer Science and Business Media*, vol. 1, pp. 21–30, 2012.

[14] M. Shamsfard and Z. Karimi, "The automatic writer system of Persian texts," *12th Iranian Computer Society Conf.*, vol. 40, pp. 1–28, 2006.

[15] A. D. Chowanda, A. R. Sanyoto, D. Suhartono and C. J. Setiali, "Automatic debate text summarization in online debate forum," *Elsevier Science Direct*, vol. 116, pp. 11–19, 2017.

[16] P. M. Sabuna and D. B. Setyohadi, "Summarizing Indonesian text automatically by using sentence scoring and decision tree," *2nd Int. conf. on Information Technology, Information Systems and Electrical Engineering, Indonesia*, vol. 9, no. 3, pp. 1–6, 2018.

[17] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman and Y. J. Kumar, "Text summarization features selection method using pseudo genetic-based model," *Int. Conf. on Information Retrieval Knowledge Management*, vol. 8, pp. 193–197, 2012.

[18] C. R. Chowdary, M. Sravanthi and P. S. Kumar, "A system for query specific coherent text multi-document summarization," *International Journal on Artificial Intelligence Tools*, vol. 19, no. 5, pp. 597–626, 2010.

[19] S. Manne, P. S. Mohd and S. S. Fatima, "Extraction based automatic text summarization system with hmm tagger," *Spinger-Verlag Berlin Heidelberg*, vol. 16, no. 2, pp. 421–428, 2012.

[20] Z. Sarabi, H. Mahyar and M. Farhoodi, "ParsiPardaz: Persian language processing toolkit," in *3rd Int. Conf. on Computer and Knowledge Engineering*, Iran: Ferdowsi University of Mashhad, pp. 1–8, 2013.

[21] N. Riahi, F. Ghazali and M. Ghazali, “Improving the efficiency of the Persian abstract synthesis system using pruning algorithms in neural networks,” in *The First Int. Conf. on Line and Language Processing Persian*, Iran: Semnan University, 2012.

[22] F. Shafiei and M. Shamsifard, “The automatic dictionary of Persian texts,” *20th National Computer Society Conf.*, vol. 1, pp. 931– 936, 2014.