Farsi Conceptual Text Summarizer: A New Model in Continuous Vector Space

Mohammad Ebrahim Khademi
Faculty of Electrical and Computer Engineering, Malek Ashtar University of Technology, Iran
khademi@mut.ac.ir
Mohammad Fakhredanesh*
Faculty of Electrical and Computer Engineering, Malek Ashtar University of Technology, Iran
fakhredanesh@mut.ac.ir
Seyed Mojtaba Hoseini
Faculty of Electrical and Computer Engineering, Malek Ashtar University of Technology, Iran
mojtabahoseini@aut.ac.ir

Received: 04/Feb/2019 Revised: 24/Aug/2019 Accepted: 08/Sep/2019

Abstract

Traditional methods of summarization were very costly and time-consuming. This led to the emergence of automatic methods for text summarization. Extractive summarization is an automatic method for generating summary by identifying the most important sentences of a text. In this paper, two innovative approaches are presented for summarizing the Farsi texts. In these methods, using a combination of deep learning and statistical methods (TFIDF), we cluster the concepts of the text and, based on the importance of the concepts in each sentence, we derive the sentences that have the most conceptual burden. In these methods, we have attempted to address the weaknesses of representation in repetition-based statistical methods by exploiting the unsupervised extraction of association between vocabulary through deep learning. In the first unsupervised method, without using any hand-crafted features, we achieved state-of-the-art results on the Pasokh single-document corpus as compared to the best supervised Farsi methods. In order to have a better understanding of the results, we have evaluated the human summaries generated by the contributing authors of the Pasokh corpus as a measure of the success rate of the proposed methods. In terms of recall, these have achieved favorable results. In the second method, by giving the coefficient of title effect and its increase, the average ROUGE-2 values increased to 0.4% on the Pasokh single-document corpus compared to the first method and the average ROUGE-1 values increased to 3% on the Khabir news corpus.

Keywords: Extractive Text Summarization; Unsupervised Learning; Language Independent Summarization; Continuous Vector Space; Word Embedding.

1. Introduction

Automatic text summarization of a large corpus has been a source of concern over the years, from two areas of information retrieval and natural language processing. The primary studies in this field began in 1950s. Baxendale, Edmundson and Luhnhave done research in those years[1]–[3]. Automatic generation of summaries provides a short version of documents to help users in capturing the important contents of the original documents in a tolerable time [4]. Now humans produce summaries of documents in the best way. Today, with the growth of data, especially in the big data domain, it is not possible to generate all of these summaries manually, because it’s neither economical nor feasible.

There are two approaches to text summarization based on the chosen process of generating the summary [5]:

- Extractive summarization: This approach of summarization selects a subset of existing words, phrases, or sentences in the original text to form the summary. There are, of course, limitations on choosing these pieces. One of these limitations, which is common in summarization, is output summary length.
- Abstractive summarization: This approach builds an internal semantic representation and then uses natural language generation techniques to create a summary that is expected to be closer to what the text wants to express.

Based on the current limitations of natural language processing methods, extractive approach is the dominant approach in this field. Almost all extractive summarization methods encounter two key problems in [6]:

- assigning scores to text pieces
- choosing a subset of the scored pieces

Traditional methods of summarization were very costly and time-consuming. This led to the emergence of automatic methods for text summarization. Extractive summarization is an automatic method for generating summary by identifying the most important sentences of a text. Hitherto text summarization has traveled a very unpaved path to address this challenge unsupervisedly. In the beginning, frequency based approaches were utilized for text summarization. Then, lexical chain based

* Corresponding Author

approaches came to succeed with the blessing of using large lexical databases such as WordNet [7] and FarsNet [8], [9]. Since the most common subject in the text has an important role in summarization, and lexical chain is a better criterion than word frequency for identifying the subject of text; as a result, a more discriminating diagnosis of the subject of text was made possible which was a further improvement in summarization. However the great reliance of these methods on lexical databases such as WordNet or FarsNet is the main weakness of these methods. For the success of these methods depends on enriching and keeping up to date the vocabulary of these databases that is very costly and time consuming, removing this weakness is not feasible.

Hence, valid methods such as Latent Semantic Analysis (LSA) based approaches that do not use dedicated static sources -which requires trained human forces for producing them- became more prominent. Latent Semantic Analysis is a valid unsupervised method for an implicit representation of the meaning of the text based on the co-occurrence of words in the input document. This method is unsupervised and it is considered an advantage. But this method has many other problems:

- The dimensions of the matrix changes very often (new words are added very frequently and corpus changes in size).
- The matrix is extremely sparse since most words do not co-occur.
- The matrix is very high dimensional in general ($\approx 10^6 \times 10^6$)
- Quadratic cost to train (i.e. to perform SVD)

With the advent of machine learning methods in recent years, training more complex models on much larger datasets has become possible. Lately, the advancement in computing power of GPUs and new processors have made it possible for hardware to implement these more advanced models. One of the most successful of these cases in recent years is the use of the distributed representation of vocabularies [10].

Word Embedding model was developed by Bengio et al. more than a decade ago [11]. The word embedding model W, is a function that maps the words of a language into vectors with about 200 to 500 dimensions. To initialize W, random vectors are assigned to words. This model learns meaningful vectors for doing some tasks.

In lexical semantics, Linear Dimension Reduction methods such as Latent Semantic Analysis have been widely used [12]. Non-linear models can be used to train word embedding models [13], [14]. Word embedding models not only have a better performance, but also lacks many problems of Linear Dimension Reduction methods such as Latent Semantic Analysis.

Distributed representation of vocabularies (Word Embedding) is one of the important research topics in the field of natural language processing [10], [12]. This method, which in fact is one of the deep learning branches, has been widely used in various fields of natural language processing in recent years. Among these, we can mention the following:

- Neural language model [11], [15]
- Sequence tagging [16], [17]
- Machine translation [18], [19]
- Contrasting meaning [20]

Bengio et al. [11], Mikolov et al. [21], and Schwenk [15] have shown that Neural network based language models have produced much better results than N-gram models.

In this paper, a novel method of extractive generic document summarization based on perceiving the concepts present in sentences is proposed. Therefore after unsupervised learning of the target language word embedding, input document concepts are clustered based on the learned word feature vectors (hence the proposed method is language independent). After allocating scores to each conceptual cluster, sentences are ranked and selected based on the significance of the concepts present in each sentence. Ultimately we achieved promising results on Pasokh benchmark corpus.

The structure of the paper is as follows. Section two describes some related works. Section three presents the summary generation process. Section four outlines evaluation measures and experimental results. Section five concludes the paper and discusses the avenues for future research.

2. Related Works

Although many text summarization methods are available for languages such as English, little work is done in devising methods of summarizing Farsi texts.

In general these methods can be categorized as supervised and unsupervised, while most of the Farsi proposed methods so far have been of the former type. Supervised summarization methods presented for Farsi documents are divided into four categories of heuristic, lexical chain based, graph based, and machine learning or mathematical based methods:

- Heuristic method:
    - Hassel and Mazdak proposed FarsiSum as a heuristic method [22]. It is one of the first attempts to create an automatic text summarization system for Farsi. The system is implemented as a HTTP client/server application written in Perl. It has used modules implemented in SweSum (Dalianis 2000), a Farsi stop-list in Unicode format and a small set of heuristic rules.
- Lexical chain based methods:
    - Zamanifar et al. [23] proposed a new hybrid summarization technique that combined “term co-occurrence property” and “conceptually related feature” of Farsi language. They consider the relationship between words and use a synonym dataset to

  eliminate similar sentences. Their results show better performance in comparison with FarsiSum.
  - Shamsfard et al. [24] proposed Parsumist. They presented single-document and multi-document summarization methods using lexical chains and graphs. To rank and determine the most important sentence, they consider the highest similarity with other sentences, the title and keywords. They achieved better performance than FarsiSum.
  - Zamanifar and Kashefi [25] proposed AZOM, a summarization approach that combines statistical and conceptual text properties and in regards of document structure, extracts the summary of text. AZOM performs better than three common structured text summarizers (Fractal Yang, Flat Summary and Co-occurrence).
  - Shafiee and Shamsfard [26] proposed a single/multi-document summarizer using a novel clustering method to generate text summaries. It consists of three phases: First, a feature selection phase is employed. Then, FarsNet, a Farsi WordNet, is utilized to extract the semantic information of words. Finally, the input sentences are clustered. Their proposed method is compared with three known available text summarization systems and techniques for Farsi language. Their method obtains better results than FarsiSum, Parsumist and Ijaz.
- Graph based method:
  - Shakeri et al. [27] proposed an algorithm based on the graph theory to select the most important sentences of the document. They explain their objective as “The aim of this method is to consider the importance of sentences independently and at the same time the importance of the relationship between them. Thus, the sentences are selected to attend in the final summary contains more important subjects, and also have more contact with other sentences.” [27] Evaluation results indicate that the output of proposed method improves precision, recall and ROUGE-1 metrics in comparison with FarsiSum.
  - Hosseinikhah et al.[28]proposed an extractive method by combining natural language processing and text mining techniques. Part of speech tagging is used for calculating coefficient of words’ importance and graph similarity’s methods are used to select sentences without redundancy problem.
- Machine learning and mathematical based methods:
  - Kiyomarsi and Rahimi [29] proposed a new method for summarizing Farsi texts based on features available in Farsi language and the use of fuzzy logic. Their method obtains better results as compared with four previous methods.
  - Tofighy et al. [30] proposed a new method for Farsi text summarization based on fractal theory whose main goal is using hierarchical structure of document to improve the summarization quality of Farsi texts. Their method achieved a better performance than FarsiSum, but weaker than AZOM.
  - Bazghandi et al. [31] proposed a textual summarization system based on sentence clustering. Collective intelligence algorithms are used for optimizing the methods. These methods rely on semantic aspect of words based on their relations in the text. Their results is comparable to traditional clustering approaches.
  - Tofighi et al. [32] proposed an Analytical Hierarchy Process (AHP) technique for Farsi text summarization. The proposed model uses the analytical hierarchy as a base factor for an evaluation algorithm. Their results show better performance in comparison with FarsiSum.
  - Pourmasoumi et al. [33] proposed a Farsi single-document summarization system called Ijaz. It is based on weighted least squares method [34]. Their results proved a better performance as compared with FarsiSum. They also proposed Pasokh [35], a popular corpus for evaluation of Farsi text summarizers.
  - Farzi and Kianian [36] proposed a Farsi summarizer based on a semi-supervised summarization approach, which is a combination of co-training and self-training algorithms. Co-training is a machine learning algorithm used when there are only small amounts of labeled data and large amounts of unlabeled data. They took a semi-supervised approach to overcome the absence of sufficient labeled data.

As an unsupervised method, Honarpisheh et al. [37] proposed a new multi-document multi-lingual text summarization method, based on singular value decomposition (SVD) and hierarchical clustering.

Success of Lexical chain based methods and supervised machine learning methods depends on enriching and keeping up to date lexical databases and training labeled datasets respectively, that is very costly and time consuming. These methods often use language-dependent features and cannot be generalized to other languages. On the other hand unsupervised methods such as SVD based methods have many problems that are mentioned in the previous section.

The proposed generic extractive method is a novel method that not only is unsupervised, but also does not have many problems of SVD-based methods and without using any hand-crafted features, achieves much better performance compared to supervised methods.

Although many text summarization methods are available for languages such as English

As to the related works done for English language, two of the latest research accomplished in the realm of deep learning are:

- Joshi et al. [38]proposed SummCoder, a methodology for genericextractive text summarization of single documents.The approach generates a summary according to three sentence selection metrics:
  - The sentence content relevance is measured using a deep auto-encoder network,
  - The novelty metric is derived by exploiting the similarity among sentences represented as embeddings.
  - The sentence position relevance metric is a hand-designed feature, which assigns more weight to the first few sentences through a dynamic weight calculation function.

  Finally Sentence Ranking & Selection module fuses the three scores and computes the final ranks for each sentence to select high-ranked sentences for the final document summary.
- Nallapati et al. [39] proposed a Recurrent Neural Network (RNN) based sequence model for extractive summarization of documents and show that it achieves performance better than or comparable to state-of-the-art.

3. Proposed Method

In this section we propose a novel method of extractive generic document summarization based on perceiving the concepts present in sentences. It is an unsupervised and language independent method that does not have many problems of SVD-based methods. For this purpose, firstly, the necessary preprocesses are performed on the Hamshahri2 [40]corpus texts. Subsequently, the Farsi word embedding is created by unsupervised learning of Hamshahri2 corpus. Then the input document keywords are extracted. Afterward the input document concepts are clustered based on the learned word feature vectors (hence the proposed method can be generalized to other languages), and the score of each of these conceptual clusters are calculated. Finally, the sentences are ranked and selected based on the significance of the concepts present in each sentence. The chart of this method is presented in **Error! Reference source not found.**. The following sections will be described based on this chart.

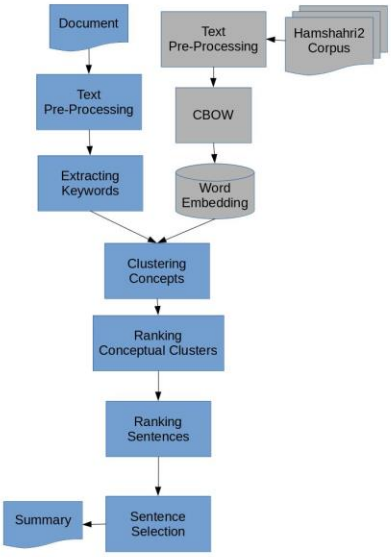


Fig. 1 Conceptual text summarizer

3.1 Text Pre-Processing

To learn a Farsi language model, we use Hamshahri2 [40] corpus. We need to produce a dictionary of vocabularies of Hamshahri2 corpus. To do this, we tokenize the words of each text file of the corpus using Hazm word_tokenize function[41] library. Hazm is an applicable open source natural language processing library in Farsi. Then we compose a dictionary out of these words by counting the frequency of each word throughout the corpus. This dictionary will be used in succeeding steps.

We constitute a complete list of Farsi stopwords out of frequent words in the prepared dictionary along with stopword lists in other open source projects.

3.2 Unsupervised Learning of Farsi Word Embedding

The Hamshahri2 [40] corpus has 3206 text files in unlabeled text sections. Each of these files is a concatenation of hundreds of news and articles. These news and articles are from different fields of cultural, political, social, etc.

To construct a suitable Farsi word embedding set, we use CBOW model [42]. This model is a neural network with one hidden layer. To learn the model a small window moves across the corpus texts and the network tries to predict *the central word of the window* using the words around it.

We assume a window with nine words length and it goes across the unlabeled texts of Hamshahri2 corpus to learn the weights of the network as Farsi word embedding vectors. The first and the last four words of each window is assumed to be the input of the network. The central word of the window is assumed to be the label of the output. Thus we have a rich labeled dataset.

Completing the learning process of network weights on all windows of Hamshahri2 corpus, we will have a suitable Farsi word embedding set, whose words' dimension is equal to the size of the hidden layer of the network. The hidden layer size is assumed to be 200 in this work.

The Farsi word embedding generated at this stage, maps every words of the Hamshahri2 corpus to a vector in a 200 dimensional vector space. The generated Farsi word embedding set contains 300,000 words.

The t-SNE method for visualization can be used to better understand the word embedding environment.

In the mapping of the words in **Error! Reference source not found.**, similar words are closer to each other. This issue can also be examined from other dimensions, as another example in **Error! Reference source not found.**, the closest vocabularies to the header terms is given (using the proposed Farsi word embedding generated in this work).

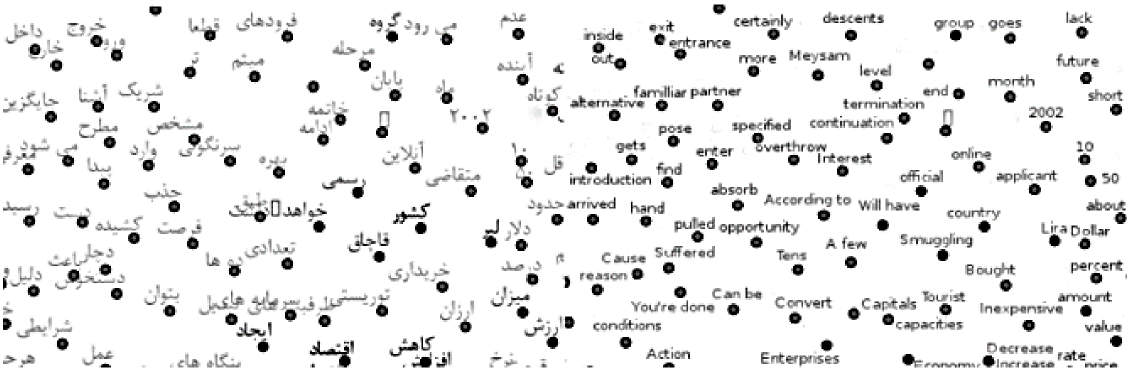
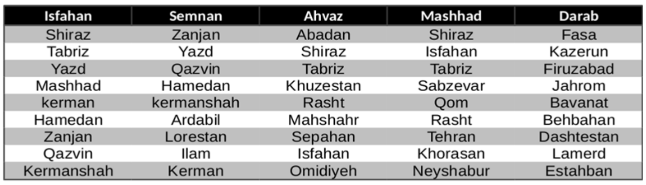


Fig. 2 A Persian word embedding visualization using t-SNE method. Part of the words of one of the texts of the Pasokh corpus visualized in this figure

In the proposed method, using the relationship between words, the concepts of the input document are represented. In this method, the importance of sentences is determined using semantic and syntactic similarities between words. And Instead of using single words to express concepts, multiple similar words are used. For example, the occurrence of words: computer, keyboard, display, mouse and printer, even though they are not frequently repeated singly in the input document, express a certain concept.

As stated in the introduction, the great reliance of lexical chain based methods on lexical databases is the main weakness of these methods. At this stage, to remove this weakness, an appropriate word embedding for summarization is created that encompasses the semantic and syntactic communication of the words in a broader and more up to date lexical range than databases that of lexical.

The word embedding presented in this work is able to discover relationships present in the outside world that do not exist in common vocabulary databases. For example, this word embedding can detect the relation between the words of Mashhad, Neyshabur and Khorasan (**Error! Reference source not found.**). Mashhad is the capital of Khorasan province and Neyshabur is one of the cities of this province. (The common vocabulary databases that cannot discover such relationships, are comprehensive lexical databases that carry different meanings for each word along with relationships between them such as: synonyms, antonyms, part of / containing, or more general / more specific relationships. But their construction is manual, costly and time-consuming.)



Isfahan	Semnan	Ahvaz	Mashhad	Darab
Shiraz	Zanjan	Abadan	Shiraz	Fasa
Tabriz	Yazd	Shiraz	Isfahan	Kazerun
Yazd	Qazvin	Tabriz	Tabriz	Firuzabad
Mashhad	Hamedan	Khuzestan	Sabzevar	Jahrom
kerman	kermanshah	Rasht	Qom	Bavanat
Hamedan	Ardabil	Mahshahr	Rasht	Behbahan
Zanjan	Lorestan	Sepahan	Tehran	Dashtestan
Qazvin	Ilam	Isfahan	Khorasan	Lamerd
Kermanshah	Kerman	Omidiyeh	Neyshabur	Estahban

Fig. 3 The closest vocabulary to the header terms is given (using the proposed Persian word embedding generated in this work)

3.3 Extracting the Keywords of the Document

For extracting the keywords of the input document, we first tokenized the words of the document using Hazm tokenizer [41]. Then we excluded stopwords from input document tokens. The score of each word of the input document calculated using equation (1) [43]:

$$point(w) = TF_{ij} \times IDF_i \qquad (1)$$

where w is the intended word, TF calculated from equation (2):

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \qquad (2)$$

where f_{ij} is frequency of the i-th word in the j-th document and $max_k f_{kj}$ is maximum frequency of the words in the input document. The TF is normalized using this division.

Finally, IDF in equation (1) was calculated from equation (3):

$$IDF_i = log_2(N/n_i) \qquad (3)$$

where N is the number of documents of the Hamshahri2 corpus and n_i is the number of documents in the corpus that the i-th word has been observed there.

If a word is not in the Hamshahri2 corpus, there will not be a score for it. Also due to the absence of a vector in the continuous vector space for this word, it is deleted from the decision making cycle. Therefore, learning word embedding on a richer Farsi corpus will cause to increase the accuracy of the method.

3.4 Clustering Concepts

In this phase, the concepts present in the input document are constructed using the Farsi word embedding obtained in section 3.2. For this purpose:

1. First we sort the keywords of the previous phase according to their calculated scores.
2. Then we map all input document terms into a 300-dimensional space using the prepared Farsi word embedding
3. We cluster the concepts of this document into ten different clusters using K-means algorithm:
    - To select the initial centroids, we used the top keywords selected in Section 3.3, starting with the top keyword and selecting it as the first centroid. Then, using the cosine distance criterion, we extract 150 of the words most similar to this keyword from the word embedding found in Section 3.2. To select a second centroid, we go to the next keyword selected in Section 3.3 and check to see if it is among the most similar words to the previously selected centroids. If available, we will skip this keyword and move on to the next keyword. Otherwise we choose this keyword as the next centroid. We continue these steps until we have extracted the first 10 centroids suitable for clustering. Excluding keywords that are among the words similar to the preceding cenroids makes the selected centroids of the most important keywords of the text less semantically similar. The selected initial centroids thus help to differentiate created clusters conceptually.
    - Then we cluster the entire words of the input document using the obtained initial centroids.
    - Each obtained cluster can be considered as a concept. Thus ten key concepts of the document are constructed.
    - Finally, we consider the nearest word to each cluster center as the criterion word for that cluster or concept.
    - The total score of each concept is calculated using the equation (4):

$$point(C) = \sum_{w \in C} \left(point(w) \times nearness(w) \right) \quad (4)$$

where w is the word, C is the concept and point(w) is the total score of each word that was calculated based on equation (1).

The nearness(w) indicates the closeness of each word in the intended concept to the concept's criterion word. Therefore the words nearer to the concept's criterion word will have larger linear coefficients and the words farther to that criterion word will have smaller linear coefficients. Thus the nearness of each word to its concept's criterion word affects the final score of the concept. Hence, repetition of more closely situated words in the input document will result in a higher score than repetition of farther words.

3.5 Sentence Ranking

For ranking sentences, the following steps are taken:

- First, the input document is read line by line and the sentences of each line are separated using Hazm sentence tokenizer.
- For scoring extracted sentences, equation (5) is used:

$$score(S) = \frac{\sum_{w \in S} point(C)}{N} \quad (5)$$

where S is a sentence, N is its number of words and point(C) is the score of the intended word's concept.

- By dividing the sentence score into its number of words, we normalized the obtained score, so that shorter and longer sentences would have equal chance of selection.
- Sentences are sorted according to their normalized scores.
- According to the desired summary length, some sentences with the highest score are selected, and are displayed in the order they appear in the document.

3.6 Taking advantage of titles

As discussed in the previous section, the first method presented in this article does not exploit the benefits of the title of the text in the summarization process. The text title usually contains the most important text message. So the concepts mentioned in the title can earn more points and the title explanatory sentences in the text thus gain more prominence in the summary. This section examines the importance of the title in the second proposed method. For this purpose, in the first step of the summarization process, the coefficient of title effect has been added to the calculation of the score of each of the words of the input document (using equation 1). This change highlights the effect of the words in the title when calculating TF_{ij} (using equation 2):

$$TF_{ij} = \frac{f_{ij} + CTE_i}{max_k f_{kj}} \quad (6)$$

In the above equation, CTEi is the coefficient of title effect. This coefficient in the case of presence of thei-th word in the title, is equal to the positive constant value, and in the absence of it, is equal to zero. By increasing the coefficient of title effect, the words that appear in the title are scored more and, in the next step (clustering) are more likely to be considered as the primary cluster centers. Given that in the next steps in order to score the words of each sentence, the score of the cluster containing the word is considered as the score of each word, the score of the words in the title affects the score of the cluster and in fact the clusters containing the words that are In the title will score more points and the sentences containing their words will also earn more points.

The results of the second proposed method are reported in Table 1,

Table 2 and Table 3.

4. Experimental Results

In this section using ROUGE criterion, our system generated summaries on single-document Pasokh corpus is evaluated and the obtained results are compared with other available Farsi summarizers.

4.1 Evaluation Measures

ROUGE-N is a measure for evaluation of summarizations [44]. This recall based measure is very close to human evaluation of summaries. This measure calculates the number of common n-grams between the system generated summaries and the reference human made summaries. It’s therefore a suitable measure for automatically evaluating summaries produced in all languages. For this work, two public ROUGE evaluation tools are studied:

1. ROUGE: Is a Perl implementation of ROUGE measure that was developed by Mr. C. Lin et al. at the University of Southern California [44]. This implementation does not support unicode and it generates unrealistic results for the Farsi summary evaluation. After obtaining the exaggerated results of this tool for Farsi summaries, we realized this great weakness.
2. ROUGE 2: Is a Java implementation of ROUGE-N measure developed by Rxnlp team and is publicly accessible [45]. This tool supports unicode and the obtained results are accurate, but it has only implemented ROUGE-N and not any other variations of ROUGE measure.

In this work, a python implementation of ROUGE-N was developed based on Mr. C. Lin’s paper [44]. This tool supports unicode and verifies the results of the ROUGE-2 implementation [45]. According to the above descriptions, the ROUGE-2 is used for summary evaluation in this study.

4.2 Pasokh Corpus

Pasokh [35] is a popular corpus for the evaluation of Farsi text summarizers. This dataset consists of a large number of Farsi news documents on various topics. It contains human-written summaries of the documents in the forms of single-document, multi-document, extractive and abstractive summaries.

The single-document dataset of Pasokh contains 100 Farsi news texts that five extractive and five abstractive summaries for each of these news are generated by different human agents.

One hundred news texts of the single-document Pasokh dataset were summarized using the proposed algorithm in this work. The compression ratio of our system summaries was 25 percent. Then we needed to calculate ROUGE-N between each of our system generated summaries and the related 5 Pasokh extractive reference summaries (human-made summaries). For this purpose, ROUGE 2.0 (Java implementation) tool was used, which is mentioned in Evaluation tool section earlier. The average of the 5 ROUGE-N is considered as the evaluation of each of our system summaries. Finally, the average of 100 system summary evaluations was calculated as the final evaluation result.

It should be noted that the news headlines of Pasokh corpus has not been used in summarization process and the results are obtained without taking advantage of headlines.

Pourmasoumi et al. [33] presented Ijaz as an extractive single-document summarizer of Farsi news in 2014 which is available online. In this experiment one hundred news texts of the Pasokh corpus were summarized using Ijaz summarizer. The compression ratio was 25 percent, and the results were obtained without using headlines.

The results are reported in Table 1,

Table 2 and Table 3.

Table 1 ROUGE-1 scores (percent) on Pasokh single-document dataset

Systems	***ROUGE-1***		
	Avg_Recall	***Avg_Precision***	***Avg_F-Score***
Shafiee and Shamsfard method [26]	38.8	42.5	39.1
Ijaz [33]	39.3	44.8	40.5
Our First Proposed Method (without using titles)	45.4	52.4	46.8
Our Second Proposed Method (Coefficient of Title Effect: 100)	**45.6**	**53.1**	**47.2**
Pasokh Authors	53.9	53.9	49.7

Table 2ROUGE-2 scores (percent) on Pasokh single-document dataset

Systems	***ROUGE-2***		
	Avg_Recall	***Avg_Precision***	***Avg_F-Score***
Shafiee and Shamsfard method [26]	21.6	24.7	22.1
Ijaz [33]	22.6	27.6	15.4
Our First Proposed Method (without using titles)	30.1	37.3	31.9
Our Second Proposed Method (Coefficient of Title Effect: 100)	**30.5**	**38.2**	**32.6**
Pasokh Authors	39.7	40.4	36.5

Table 3ROUGE-3 scores (percent) on Pasokh single-document dataset

Systems	***ROUGE-3***		
	Avg_Recall	***Avg_Precision***	***Avg_F-Score***
Shafiee and Shamsfard method [26]	16.7	19.3	17.1
Ijaz [33]	18.0	22.4	19.3
Our First Proposed Method (without using titles)	26.7	34.0	28.5
Our Second Proposed Method (Coefficient of Title Effect: 100)	**27.1**	**35.0**	**29.2**
Pasokh Authors	35.1	12.11	17.59

Thus our proposed method in this work has the following advantages over Pourmasoumi et al. method:

- Our proposed method achieves much better results than the proposed method of Pourmasoumi et al. in all ROUGE-1, ROUGE-2 and ROUGE-3 measures.
- The method proposed by Pourmasoumi et al. [33] has taken a supervised learning approach, while our learning approach is unsupervised. As defined by authorities supervised learning requires that the algorithm’s possible outputs are already known and that the data used to train the algorithm is already labeled with correct answers. While, unsupervised machine learning is more closely aligned with what some call true artificial intelligence, the idea that a computer can learn to identify complex processes and patterns without a human to provide guidance along the way. Although unsupervised learning is prohibitively complex for some simpler enterprise use cases, it opens the doors to solving problems that humans normally would not tackle.
- Their proposed method is a Farsi specific method, while our proposed method can be generalized to other languages.

Shafiee and Shamsfard [26] proposed an approach in extractive single-document Farsi summarization in 2017.

Unfortunately, neither their summarizer nor summaries generated by their proposed algorithm are available for comparison, therefore, the algorithm has been implemented.

In this experiment one hundred news texts of the Pasokh corpus were summarized using developed summarizer. The compression ratio was 25 percent, and the results were obtained using headlines. The results are reported in Table 1, Table 2 and Table 3.

Our approach has the following advantages over Shafiee and Shamsfard's approach:

- Our proposed method achieves much better results than the "number of similar and related sentences" method of Shafiee and Shamsfard in all ROUGE-1, ROUGE-2 and ROUGE-3 measures.
- Shafiee and Shamsfard's method is supervised, while ours is unsupervised. In order to calculate a feature's weight, they utilize one-third of the Pasokh single-document corpus. To compute a feature's weight, the mean of F-measure scores is calculated to be considered as the final weight of the selected feature for single-document summarization.
- Their proposed method depends on enriching and keeping up to date the FarsNet lexical database, that is very costly and time consuming, while our method depends on unsupervised learning of the target language word embedding.
- Their proposed method is a Farsi specific method, while our proposed method can be generalized to other languages.
- Their method has used the news headlines in the summarization process, while our method has obtained the results without using headlines.

In order to have a better understanding of the results, we have evaluated the human summaries generated by the contributing authors of the Pasokh corpus as a measure of the success rate of the proposed method. Assuming that the best summaries are produced by human factors (the various authors of the Pasokh corpus), these summaries should be the most ideal in the evaluation. For this purpose, we compared the summaries generated by each of the authors of the Pasokh corpus with summaries from other authors of this corpus (with the ROUGE-N criterion). Given the varying number of texts summarizing by different authors, we consider the weighted average of evaluations made by all authors as the final evaluation of the corpus authors. The results obtained in Table 1,

Table 2 and Table 3 show that:

- In terms of recall, our proposed method has achieved favorable results.
- In terms of accuracy, based on ROUGE-1 and ROUGE-2 our proposed method yielded results close to the results of the authors of the corpus and in the ROUGE-3 criterion, it reached a much higher accuracy than the authors of this corpus (2.8 times).
- In terms of F-Score, ROUGE-1 and ROUGE-2 our proposed method also yielded results close to the results of the authors of the Pasokh corpus and in the ROUGE-3 criterion, it achieved a much higher score than the authors of this corpus (1.6 times).

The comparison of the results of our proposed methods and the authors of the Pasokh corpus shows that the results obtained in this paper are close to the ideal results and, in some cases, outweigh the human abstracts.

Hassel and Mazdak created FarsiSum [22] in 2004 as one of the first Farsi text summarizers reported in related literature. The available version of FarsiSum summarizer in their website has a number of bugs. For example, the length of the summary FarsiSum produces has a significant difference with the requested compression ratio percentage. According to previous studies [26], [33], the results of our proposed method on Pasokh corpus are much higher than the results obtained by FarsiSum summarizer.

4.3 Taking advantage of titles

Now, we examine the effect of the coefficient of title effect on the summaries generated in two corpuses of the Pasokh and the Khabir news corpus: As shown in **Error! Reference source not found.**, by applying different values of the coefficient of title effect and summarizing all the records of the Pasokh corpus by this coefficient, it can be seen that the overall observed behavior is that by increasing the coefficient of title effect, the mean of the F-Score values is Gradually increased. Noteworthy, the exception is the overall behavior in the range of 0 to 12, and the observation of the gradual decrease of the F-Score values. In fact, the sentences selected in this method are sentences that relate more to clusters containing title words. By factoring in the words of the title, we actually increase the weight of clusters associated with them.

For example, if the coefficient of title effect in the Pasokh corpus is considered 25, the mean value of the F-Score for all the summarized texts of the corpus is approximately equal to that of the zero coefficient. In this case, 54 summarized texts are exactly the same as

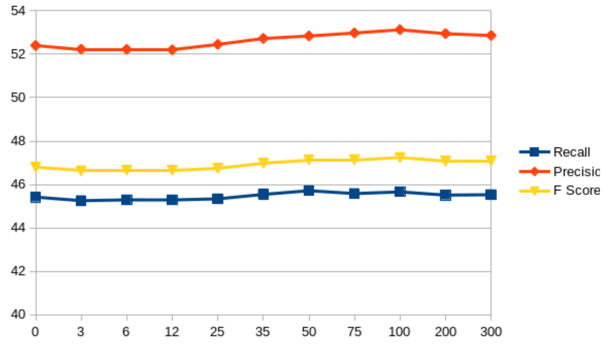


Fig. 4 Average score of all summarized texts of the Pasokh corpus for various coefficients of title effect

generated abstracts with a zero coefficient and 46 other texts are often created with a little difference. By analyzing the differences, it can be said that most of the sentences with more literal similarity to the title have been selected. In many cases, this has caused a weaker statement. In some cases, with a marginal content is selected as the title, the selected sentences are closer to the content of the headline and have been removed from the original content of the text.

Since the small number of texts of the Pasokh corpus and the observation of the results made it difficult to make the final conclusion, we repeated this experiment on the Khabir corpus. This corpus is made up of 80 thousand Farsi news and the lead of them is considered as a summary of the news. The abundance of abstracted texts in the Khabir corpus eliminates the influence of the rare factors in the final results. In this experiment, summaries with a length of 10% were produced in separate experiments. As shown in **Error! Reference source not found.**, this test was performed for various coefficients of title effect and the average score of all generated summaries in each test was calculated. By increasing the value of the coefficient of the title, the average of recall based ROUGE-1 measure initially grows linearly. After the growth of the coefficient of title effect up to 400, this linear growth is significantly reduced and the results tend to be a constant number. It is therefore clear that the increase in the coefficient of title effect in the Khabir corpus is a 3% improvement in the recall based ROUGE-1 measure of the generated summaries.

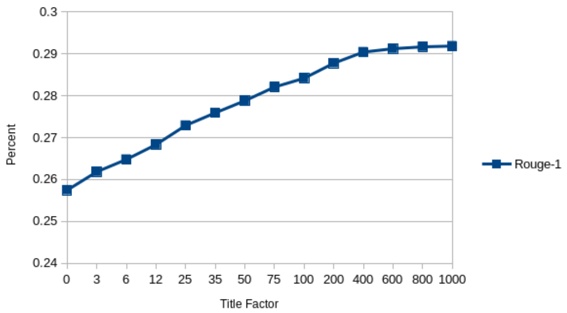


Fig. 5 Average score of all Khabir corpus texts for various coefficients of title effect

5. Conclusion

In this paper, two novel methods of extractive generic document summarization based on perceiving the concepts present in sentences are proposed. Therefore after unsupervised learning of the target language word embedding, input document concepts are clustered based on the learned word feature vectors (hence the proposed methods can be generalized to other languages). After allocating scores to each conceptual cluster, sentences are ranked and selected based on the significance of the concepts present in each sentence.

One of the most important challenges in recent researches in the field of summarizing Farsi texts is the lack of a rich lexical database in Farsi language that can be used to measure semantic similarities. In this research, by constructing a Farsi word embedding using Hamshahri2 corpus, we were able to correctly answer this shortage and provide two new methods for summarizing the texts according to the semantic and syntactic relations learned.

Using the relationship between words, the concepts discussed in the input document are represented. In these methods, the importance of sentences is determined using semantic and syntactic similarities between words. Instead of using single words to express concepts, different related words are used. We evaluated the proposed methods on Pasokh single-document dataset using the ROUGE evaluation measure. Without using any hand-crafted features, our proposed methods achieved state-of-the-art results. For system summaries generated with 25 percent compression ratio on Pasokh single-document corpus using our first method, ROUGE-1, ROUGE-2 and ROUGE-3 recall scores were 45, 30 and 27 percent, respectively.

In the second proposed method in this paper, by applying various coefficients of title effect and summarizing all the records of the Pasokh corpus by this coefficient, the overall observed behavior is the gradual increase of the mean value of the F-Score by increasing the coefficient of title effect. The comparison of the results of our proposed methods and the authors of the Pasokh corpus also demonstrates that the results obtained in this paper are close to the ideal results, and even in some cases, outweigh the human abstracts.

Evaluation of our proposed methods for summarization of other languages is suggested for future works. Learning word embedding on richer Farsi corpuses may be effective in increasing the accuracy of our methods. Using PageRank [46] algorithm to produce the concept similarity graph and to find more significant concepts may also increase the accuracy of our concept selection algorithm. Using exploited MMR (Maximum Marginal Relevance) [47] greedy algorithm in sentence selection process may decrease the redundancy of the selected sentences in our proposed methods.

References

[1] P. B. Baxendale, “Machine-made index for technical literature—an experiment,” *IBM J. Res. Dev.*, vol. 2, no. 4, pp. 354–361, 1958.

[2] H. P. Edmundson, “New methods in automatic extracting,” *J. ACM JACM*, vol. 16, no. 2, pp. 264–285, 1969.

[3] H. P. Luhn, “The automatic creation of literature abstracts,” *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, 1958.

[4] H. Khanpour, “Sentence extraction for summarization and notetaking,” University of Malaya, 2009.

[5] W. Song, L. C. Choi, S. C. Park, and X. F. Ding, “Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization,” *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9112–9121, 2011.

[6] F. Jin, M. Huang, and X. Zhu, “A comparative study on ranking and selection strategies for multi-document summarization,” in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 525–533.

[7] G. A. Miller, “WordNet: a lexical database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[8] M. Shamsfard, “Developing FarsNet: A lexical ontology for Persian,” in *4th Global WordNet Conference, Szeged, Hungary*, 2008.

[9] M. Shamsfard *et al.*, “Semi automatic development of farsnet; the persian wordnet,” in *Proceedings of 5th global WordNet conference, Mumbai, India*, 2010, vol. 29.

[10] G. E. Hinton, J. L. Mcclelland, and D. E. Rumelhart, *Distributed representations, Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*. MIT Press, Cambridge, MA, 1986.

[11] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” *J. Mach. Learn. Res.*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[12] P. D. Turney and P. Pantel, “From frequency to meaning: Vector space models of semantics,” *J. Artif. Intell. Res.*, vol. 37, pp. 141–188, 2010.

[13] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[14] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[15] H. Schwenk, “Continuous space language models,” *Comput. Speech Lang.*, vol. 21, no. 3, pp. 492–518, 2007.

[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *J. Mach. Learn. Res.*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[17] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160–167.

[18] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. M. Schwartz, and J. Makhoul, “Fast and Robust Neural Network Joint Models for Statistical Machine Translation.,” in *ACL (1)*, 2014, pp. 1370–1380.

[19] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[20] Z. Chen *et al.*, “Revisiting Word Embedding for Contrasting Meaning.,” in *ACL (1)*, 2015, pp. 106–115.

[21] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Černockỳ, “Empirical evaluation and combination of advanced language modeling techniques,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[22] M. Hassel and N. Mazdak, “FarsiSum: a Persian text summarizer,” in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, 2004, pp. 82–84.

[23] A. Zamanifar, B. Minaei-Bidgoli, and M. Sharifi, “A new hybrid farsi text summarization technique based on term co-occurrence and conceptual property of the text,” in *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD’08. Ninth ACIS International Conference on*, 2008, pp. 635–639.

[24] M. Shamsfard, T. Akhavan, and M. E. Joorabchi, “Persian document summarization by PARSUMIST,” *World Appl. Sci. J.*, vol. 7, pp. 199–205, 2009.

[25] A. Zamanifar and O. Kashefi, “AZOM: a Persian structured text summarizer,” *Nat. Lang. Process. Inf. Syst.*, pp. 234–237, 2011.

[26] F. Shafiee and M. Shamsfard, “Similarity versus relatedness: A novel approach in extractive Persian document summarisation,” *J. Inf. Sci.*, p. 0165551517693537, 2017.

[27] H. Shakeri, S. Gholamrezazadeh, M. A. Salehi, and F. Ghadamyari, “A new graph-based algorithm for Persian text summarization,” in *Computer science and convergence*, Springer, 2012, pp. 21–30.

[28] T. Hosseinikhah, A. Ahmadi, and A. Mohebi, “A new Persian Text Summarization Approach based on Natural Language Processing and Graph Similarity,” *Iran. J. Inf. Process. Manag.*, vol. 33, no. 2, pp. 885–914, 2018.

[29] F. Kiyomarsi and F. R. Esfahani, “Optimizing persian text summarization based on fuzzy logic approach,” in *2011 International Conference on Intelligent Building and Management*, 2011.

[30] M. Tofighy, O. Kashefi, A. Zamanifar, and H. H. S. Javadi, “Persian text summarization using fractal theory,” in *International Conference on Informatics Engineering and Information Science*, 2011, pp. 651–662.

[31] M. Bazghandi, G. T. Tabrizi, M. V. Jahan, and I. Mashahd, “Extractive Summarization Of Farsi Documents Based On PSO Clustering,” *jiA*, vol. 1, p. 1, 2012.

[32] S. M. Tofighy, R. G. Raj, and H. H. S. Javad, “AHP techniques for Persian text summarization,” *Malays. J. Comput. Sci.*, vol. 26, no. 1, pp. 1–8, 2013.

[33] P. Asef, K. Mohsen, T. S. Ahmad, E. Ahmad, and Q. Hadi, “IJAZ: AN OPERATIONAL SYSTEM FOR SINGLE-DOCUMENT SUMMARIZATION OF PERSIAN NEWS TEXTS,” vol. 0, no. 121, pp. 33–48, Jan. 2014.

[34] T. Strutz, *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*. Vieweg and Teubner, 2010.

[35] B. B. Moghaddas, M. Kahani, S. A. Toosi, A. Pourmasoumi, and A. Estiri, “Pasokh: A standard corpus for the evaluation of Persian text summarizers,” in *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on*, 2013, pp. 471–475.

[36] S. Farzi and S. Kianian, “Katibeh: A Persian news summarizer using the novel semi-supervised approach,” *Digit. Scholarsh. Humanit.*, vol. 34, no. 2, pp. 277–289, 2018.

[37] M. A. Honarpisheh, G. Ghassem-Sani, and S. A. Mirroshandel, “A Multi-Document Multi-Lingual Automatic Summarization System.,” in *IJCNLP*, 2008, pp. 733–738.

[38] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, “SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders,” *Expert Syst. Appl.*, vol. 129, pp. 200–215, 2019.

[39] R. Nallapati, F. Zhai, and B. Zhou, “Summarunner: A recurrent neural network based sequence model for extractive summarization of documents,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[40] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, “Hamshahri: A standard Persian text collection,” *Knowl.-Based Syst.*, vol. 22, no. 5, pp. 382–387, 2009.

[41] *hazm: Python library for digesting Persian text*. Sobhe, 2017.

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. 2013, pp. 3111–3119.

[43] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge university press, 2014.

[44] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out: Proceedings of the ACL-04 workshop*, 2004, vol. 8.

[45] *ROUGE-2.0: Java implementation of ROUGE for evaluation of summarization tasks. Stemming, stopwords and unicode support*. 2017.

[46] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the web.,” Stanford InfoLab, 1999.

[47] J. Carbonell and J. Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 335–336.

Mohammad Ebrahim Khademi received his M.S. degree in computer engineering from the Malek Ashtar University of Technology, Iran, in 2013. He is currently a PhD candidate in computer engineering there. His research interests include machine learning (deep learning) and natural language processing.

Mohammad Fakhredanesh received his B.S., M.S. and PhD degree in computer science and Engineering from the Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2005, 2007, and 2014 respectively. He is currently an assistant professor at the Malek Ashtar University of Technology. His research interests are the fields of artificial intelligence, pattern recognition, and text summarization**.**

Seyed Mojtaba Hoseini received his B.S. degree in Electronic Engineering from Malek Ashtar University of Technology in 1991. He also received his M.S. and PhD degrees in Computer Architecture Engineering from Amirkabir University of Technology in 1995 and 2011 respectively. His research interests include Wireless sensor Networks, with an emphasis on target coverage and tracking applications, image and signal processing, and evolutionary computing.