Context-Based Persian Multi-Document Summarization (*global view*)

Asef poormasoomi
Computer Engineering
Dept. Ferdowsi University
of Mashhad, Iran
as.poormasoomi@stu-mail.um.ac.ir

Mohsen Kahani
Computer Engineering
Dept. Ferdowsi University
of Mashhad, Iran
kahani@um.ac.ir
www.um.ac.ir/~kahani

Saeed Varasteh Yazdi
Computer Engineering
Dept. Amirkabir
University of Technology,
Tehran, Iran,
saeed.varasteh@aut.ac.ir

Hossein Kamyar
Computer Engineering
Dept. Ferdowsi University
of Mashhad, Iran
hossein.kamyar@stu-mail.um.ac.ir

Abstract— **Multi-document summarization is the automatic extraction of information from multiple documents of the same topic. This paper proposes a new method, using LSA, for extracting the global context of a topic and removes sentence redundancy using SRL and WordNet semantic similarity for Persian language. In the previous approaches, the focus was on the sentence features (local view) as the main and basic unit of text. In this paper, the sentences are selected based on the main context hidden in the all documents of a topic. The experimental results show that our proposed method outperforms other Persian multi-document systems.**

Keywords- Multi-document summarization, LSA, Semantic Similarity, Semantic Role labeling

I. INTRODUCTION

With the impressive growth of available data on the web, continuous increase of research resources and existence of numerous amount of news website, document summarization methods have been the subject of many researches in the area of information retrieval. There are two general types of summarization [1]: single document and multi document summarization.

In multi-document task, there are several documents per topic, each of which talks about the topic from a different perspective. For example, consider the topic "*world-wide water shortage problem*". There could be several documents; One talks about "*water shortage in Iran*", another about "*water shortage in China*".

There are several important challenges in multi-document summarization. The major issues are [2]:

a) Since there are several documents per topic, each of which has a distinct perspective of the topic, it is difficult to create a readable and coherent summary.

b) Information redundancy is an important matter that must be considered in multi-document summarization. As there are several documents in each topic, overlapping would occur. Therefore, it is essential to use effective methods for recognizing and removing redundancy.

c) Another challenge is to determine the differences among documents and cover all important information.

In this work, we focus on multi-document summarization of Persian documents. LSA has been used for extracting the global context of topic and WordNet based semantic similarity for recognizing similar sentences and removing the redundancy. Unlike previous works that have used term-sentence matrix, here, term-document matrix has been utilized. The most related concept of each topic has been extracted using cosine similarity. Then, we rank sentences based on their similarity to the assigned concept of the related topic. After that, using a semantic role labeling (SRL) method the roles of sentence units are obtained and the semantic similarities (using WordNet) of the top ranked sentences in the determined roles are calculated.

The remainder of the paper is organized as follows: Section 2 and 3 discuss related works. In Section 4 and 5, we describe the proposed method in details. The experimental results are presented in Section 6, and finally a conclusion is drawn and future works are discussed.

II. RELATED WORKS

Generally, automatic text summarization methods can be divided into two global categories: supervised and unsupervised approaches [1]. In supervised methods, there are large amount of document-summary pairs that have been created by human and can be used as training data. These approaches are model based, meaning that they would work properly when the documents are related to the human summaries. So, if documents were not similar to the model, it might not work properly [3][4]. Despite supervised methods, unsupervised approaches do not need large amount of human-made summaries for the training phase. Many methods belong to this category [5].

Latent Semantic Analysis was first proposed to address the problems of synonymy and polysemy in information retrieval [6]. Since then, LSA has become more and more attractive and a lot of researchers in Natural Language Processing have analyzed it theoretically [7]. Papadimitrious et al. in [7] performed a probabilistic analysis on LSA and proved that under certain conditions, LSA succeeds in capturing the underlying semantics of the corpus and improves the retrieval performance, while addressing the synonymy and polysemy problems. This method has also been used in text summarization.

Gong and Liu [5] proposed a method for selecting most meaningful sentences by using LSA. They have used term-sentence matrix $A = [A_1, A_2, \ldots, A_n]$, where each column A_i represents the weighted term-frequency vector of the sentence i in the document and then have applied singular decomposition value (SVD) to exploit significant sentences.

Steinberger et al. [8] improved Gong`s approach by incorporating LSA with anaphora resolution. They show that adding anaphoric information as an input to SVD significantly improves the performance of the previously proposed methods that only have used lexical terms.

Yeh et al [4] proposed a document summarization method using LSA and text relationship map (TRM). They used LSA to derive the semantic matrix of a document and used semantic representation of a sentence to construct a semantic text relationship map. All of these LSA-related

978-0-7695-4554-7/11 $26.00 © 2011 IEEE
DOI 10.1109/IALP.2011.53

methods concentrate on semantic features of sentences. However, some questions have been raised. The first question is that whether only focusing on the semantic property of sentences can lead to deep understanding in multi-document text summarization. Also, if using sentence based term weighting approaches like TF-ISF as the input of LSA can represent the different perspective hidden in several documents of a topic.

To answer these questions, careful studies of the basic assumptions are required. Term Frequency-Inverse Sentence Frequency weighting schema is calculated as:

$$(TF-ISF)_{i,j} = \left(\frac{tf_{i,j}}{\sum_k tf_{k,j}}\right) Log_2\left(\frac{|S|}{|\{s:t_i \in s\}|}\right) \quad (1)$$

where $tf_{i,j}$ is the number of occurrence of term i in sentence j and $|S|$ is the total number of sentences in the corpus. It means that $TF-ISF$ just represents the behavior of terms in each sentence not in the whole document or corpus. So given the term-sentence matrix, LSA cannot completely extract the *'hidden pattern*' within the documents, because the input matrix is based on the sentences, not on the whole documents. In the other words, comparing sentences without paying attention to the context of the documents or corpus cannot lead to accurate results.

III. PERSIAN SUMMARIZATION

Unlike English text summarization methods, summarization of single and multiple documents written in Persian language is a relatively new field of research.

The first work on Persian Language is FarsiSum in 2004[9]. It is a Web based application programmed in Perl and based on SweSum [10]. FarsiSum select sentences from documents with the main body of language independent modules implemented in SweSum. It has added the Persian stop-list in Unicode format and adapted the interface modules to accept Persian texts.

The next work is done by Karimi and Shamsfard [11]. It is a Persian single document summarization method based on lexical chains and graph based methods. It uses five measures to score a sentence. These measures are: similarity to other sentences, similarity to the user's query, similarity to the title, number of common words and cue words.

Zamanifar in [12] proposed an integrated method for Farsi text summarization, which combines the term co-occurrence property and conceptually related feature of Persian language. They consider the relationship between words and use a synonym dataset to eliminate the similar sentences.

IV. PROPOSED METHOD

This paper proposes a new method using main property of multi-document summarization and the proper use of LSA. We try to consider the effect of all documents of a topic (global view) in the process of summary generation. The overall process is shown in Figure 1. The process consists of the following major phases:

A. Pre-processing phase : Text preprocessing step plays an important role for improving the accuracy. Persian language differs from English language both morphologically and semantically[12]. These differences are indicated when explaining the process.

- The first step in pre-processing phase is tokenization. A locally developed Persian Tokenizer add-on for GATE tool was used for this purpose.
- In the second step, unlike English pre-processing, we first performed stemming and then removed the stop words. This is because, there is a possibility that a preverbal element is considered as a stop word. For stemming, we also used our Persian stemmer add-on for GATE tool.
- After processing the documents, the term-document matrix is constructed. Different weighting scheme can be used at this stage. In this paper, we have used $tfidf$ weights, which is defined as below:

$$(TF-IDF)_{i,j} = \left(\frac{tf_{i,j}}{\sum_k tf_{k,j}}\right) Log_2\left(\frac{|D|}{|\{d: t_i \in d\}|}\right) \quad (2)$$

where $tf_{i,j}$ denotes the frequency that term i occurs in document j and $|D|$ is the number of documents in the corpus.

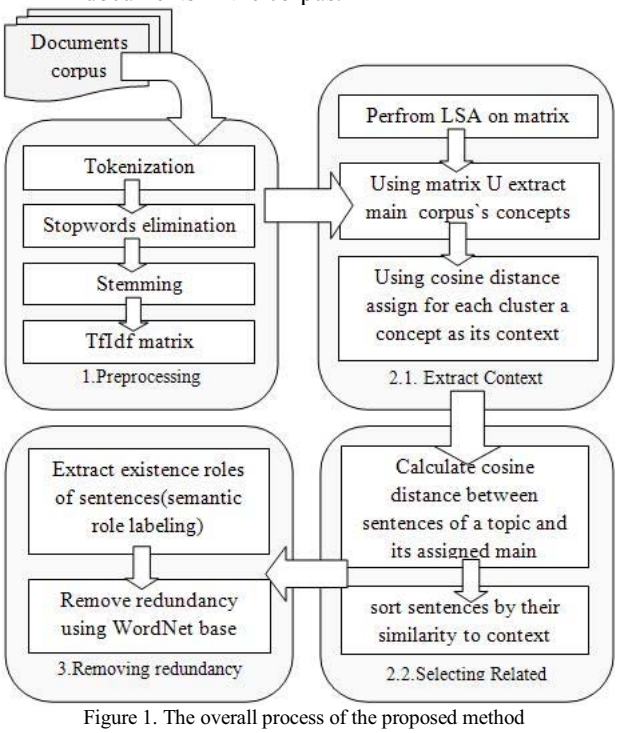

Figure 1. The overall process of the proposed method

B. Sentence Selection

1) Extracting the main existent concepts of documents: In this phase, LSA has been used for extracting the main concepts of the corpus. Then, the cosine distance of the concept vector and the document vector is calculated. This value represents the amount of similarity of each concept with a topic in the corpus. In the other words, we extract the main context of each topic.

2) Selecting related sentences: In this step, we calculate the similarity of the sentence frequency vector

and the concept vector of the related topic and sort the sentences in the descending order according to their relevance similarity with the main concept of a topic.

C. *Calculating the semantic similarity of sentences.* In this step, the Semantic Role Labeling algorithm has been used to extract the role of semantic units of the sentences. Then, with calculating the similarity of words in the same semantic role using WordNet, the pairwise sentence similarity is calculated. Finally, the similar sentences are specified and by considering the size of compression, the semantically repeated sentences are removed in order to cover all different views in the documents.

A detailed description of the proposed approach is presented in the next sections.

V. DETAILS OF THE PROPOSED METHOD.

A. Extracting main concepts of documents: LSA is based on Singular Value Decomposition (SVD) method which is akin to factor analysis. Given an $m \times n$ matrix **A**, where without loss of generality, $m \geq n$, the SVD of **A** is defined as:

$$A = USV^T \quad (3)$$

Where $\boldsymbol{U} = [u_{ij}]$ is an $m \times n$ column-orthonormal matrix whose columns are called left singular vectors; $\boldsymbol{S} = diagonal(\sigma_1, \sigma_2, \dots, \sigma_n)$ is an $n \times n$ diagonal matrix, whose diagonal elements are non-negative singular value sorted in descending order, and $\boldsymbol{V} = [v_{ij}]$ is an $n \times n$ matrix, whose columns are called right singular vectors. If $rank(\mathbf{A}) = r$, the $\boldsymbol{S}$ satisfies:

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq \sigma_{r+1} = \dots = \sigma_n = 0 \quad (4)$$

From natural language processing or semantic point of view, what SVD does is extracting the latent semantic structure of the documents represented by matrix **A**. This operation reflects a breakdown of original document into r linearly-independent base vectors or concepts. From mathematical perspective, SVD derives a mapping between the *m*-dimensional weighted vectors space and *r*-dimensional singular vector space and it is used for space reduction of sparse matrix in many applications.

In this paper, we proposed a new method that uses LSA globally and effectively. In previous approaches, the focus is on *sentence features* (local view) as the main and basic unit of text. In these approaches, with constructing sentence frequency vectors through computing weighting schema on *sentence unit* like $tfisf$, term-sentence matrix is constructed and using SVD the right singular vector matrix V^T is extracted. Then, the sentence with the largest index value in the k[th] right singular vector is selected and included in the summary.

This approach has some problems. It selects one sentence per each extracted singular vector (concept) and does not pay attention to the significance of these vectors. It has some other pitfalls that previously is expressed in other papers [4][8]. However, using LSA in this sense for summarization is the lack of attention to the main concepts hiding in the whole documents. A good summary in multi-document summarization must indicate different views hidden in the context of all documents of a topic.

In this paper, the sentences are selected based on the main context hidden in all documents of a topic. So, at first, instead of using $tfisf$, we compute $tfidf$ weighting criteria in *document unit* and then, construct the term-document matrix. In the next stage, SVD is run on this matrix, singular vector matrix is extracted. In column-orthonormal matrix **U**, column vectors are linearly independent. From NLP point of view, these columns are the independent hidden concepts of the text. Therefore, from the semantic perspective, the matrix **U** is a term-concept matrix. Calculating cosine distance between column vectors of matrix **U** (vectors of concepts) $C_i = [c_{i1}, c_{i2}, \dots, c_{im}]$ and documents vector $D_j = [d_{j1}, d_{j2}, \dots, d_{jm}]$, we assign a concept for each topic in corpus. The cosine distance between these vectors is calculated as:

$$cos(C_i, D_j) = \frac{\sum_k c_{ik} d_{jk}}{\sqrt{\sum_k (c_{ik})^2} \times \sqrt{\sum_k (d_{jk})^2}} \quad (5)$$

After assigning a concept vector to each topic as its context, the distance between the sentences of documents of a topic is calculated and the concept of that topic is assigned. For this propose, the weighted vector of sentences is created and then cosine distance between sentences vector $S_j = [s_{j1}, s_{j2}, \dots, s_{jm}]$ and concept vector $C_i = [c_{i1}, c_{i2}, \dots, c_{im}]$ is calculated and sentences are reordered based on their similarity to the topic`s concept /context.

B. Calculation of semantic similarity: In the previous phase, sentences were arranged based on their cosine similarity to the text concept/context. The nature of multi-document indicates that many sentences are semantically similar to each other. Thus, this redundancy must be removed from the final generated summary. At this phase, unlike the previous phase, using cosine distance can not specify the similarity between sentences. For more explanation, consider the following sentences as example:

S1 = United States Army successfully tested an anti-missile defense system.

S2 = U.S. military projectile interceptor streaked into space and hit the target.

S3 = Iran's weekend test of a long-range missile underscored the need for a U.S. national missile defense system.

S1 and S2 almost are expressing the same information, however as they don't have common terms; the similarity becomes zero (from cosine distance view they are not similar to each other). In the other hand, from cosine similarity perspective, S1 and S3 are similar to each other because of their common terms; though they denote very different information.

This means that cosine distance cannot compute semantic similarity of sentences properly, as it only considers the lexical aspect of terms in the sentences and neglects the position and role of terms in the sentences. This problem did not occur for the previous phase, as the aim was to extract the sentences relating to the context of documents.

A semantic role is "a description of the relationship that a constituent plays with respect to the verb in the

sentences" [13]. The roles order in Persian language is different from that of the English language. The order of the main roles in Persian are [Subject][Object][Verb], which is different from the order in English, ie [Subject][Verb][Object].

In this phase, firstly, we decomposed the sentences into their semantic units by means of SRL. Then, used Persian WordNet for measuring the relation between terms in the same semantic roles. If two words in the same semantic role are identical or have the semantic relations such as synonym, hyponym and hypernym, the terms are semantically related. If P_{ak} and P_{bl} are two propositional unit of S_a and S_b sentences respectively, and $R = \{r_1, r_2, \dots, r_{roleNum}\}$ is the set of existing roles and $Term_k(r_i)$ is the terms set of P_k in semantic role r_i, then semantic similarity between two sentences is calculated as:

$$sim(S_a, S_b) = \frac{\sum_{k=1}^{m}\sum_{l=1}^{n} Semantic_{Sim}(P_{ak}, P_{bl})}{m+n} \quad (6)$$

$$Semantic_{Sim}(P_k, P_l) = \frac{\sum_{i=1}^{roleNum} RelatedT(Term_k(r_i), Term_l(r_i))}{|Term_k(r_i)| + |Term_l(r_i)|} \quad (7)$$

in which $RelatedT(Term_k(r_i), Term_l(r_i))$ function specifies the number of words in P_k and P_l that are related. The similarity scores are between 0 and 1. In final step, for generating summary, redundancy must be removed. So, we firstly add 1th sentence of descending sorted list of sentences (sorted in previous phase) to summary. Then with respect to compression size, add the other sentences that their semantic similarity with summary`s sentences is less than 0.65.

VI. EVALUATION

In this section we describe the data set used for the evaluation, implementation issues and the experimental results.

1) Dataset: Unfortunately, there is no standrad dataset such as DUC[1] dataset for text summarization in Persian language . So, we gathered more than 180 documents for 6 topics from 5 famous news web site[2], about 30 documents per each topic. Table 1 gives a brief description of this data sets. Then, we asked Four expert persons to generate an extractive summary for each topic. The implemented systems is then compared with these human created summaries.

Number of topics	6
Number of documents per topics	30
Number of Sentences	3804
Number of terms without stopwords	17562
Compression size	250 w

Table1. **Description of the data sets**

2) Evaluation Tool: As there is no standard tools to do comparison for Persian, we desided to evaluate our method in two ways. First we implement a ROUGE[3] [14] evaluation tools for Persian language. ROUGE is the most commonly used tool for text summarization comparison in English. Each method estimates recall, precision and F-measure between human written reference summaries and the candidate summaries of the proposed system. . For example, in ROUGE-N n-gram recall is computed as follows:

$$ROUGE-N = \frac{\sum_{S\epsilon\{Reference\ Summaries\}}\sum_{gram_n} Count_{match}(gram_n)}{\sum_{S\epsilon\{Reference\ Summaries\}}\sum_{gram_n} Count(gram_n)} \quad (8)$$

Where *n* stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. In the second method, we asked some MS students in linguistics to evaluate the summaries by scoring on a Likert scale; 1(very poor) and 5(very good). The comparison results are shown in Figure 4.

Implemented Summarization Systems: : In order to evaluate our proposed method properly, we compared our method by the only publicly available system in Persian, FarsiSum. In addition, we implemented the following approaches

- LSA base summarizer proposed by Gong [5].
- FarsiSum online summarizer
- Random base summarizer.
- Our proposed method with only phases 1 and 2 (our1).
- Our complete proposed method including all phases (our2).
- Human created summaries.

3) Experimental Results: : For evaluating, at first we performed pre-processing operations on the documents of the corpus, constructed term-document matrix, performed LSA and extracted main concepts of the corpus. Then, we assigned a concept to *each topic as its main context* and followed other operations. ROUGE-*1* and ROUGE-*2* comparison results are shown in Figure 2 and 3.

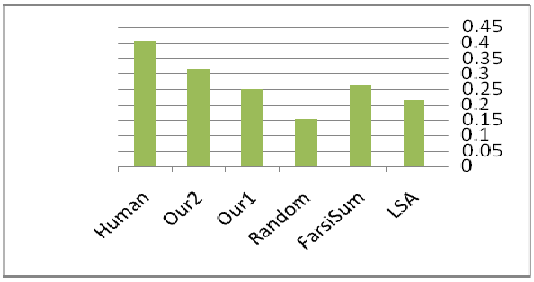

Figure2. Overall performance (recall) comparison using ROUGE-1.

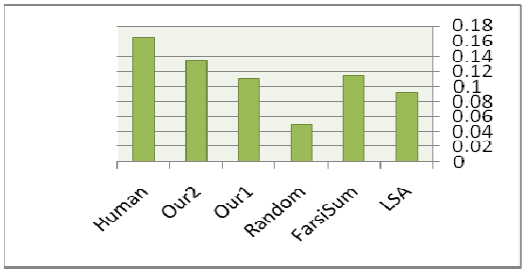

Figure3. Overall performance (recall) comparison using ROUGE-2.

[1] - Document Understanding Conference(duc.nist.gov/pubs.html)
[2] - www.tabnak.ir,www.iribb.ir,www.isna.ir, www.irna.ir,www.presstv.ir
[3] - http://berouge.com/default.aspx

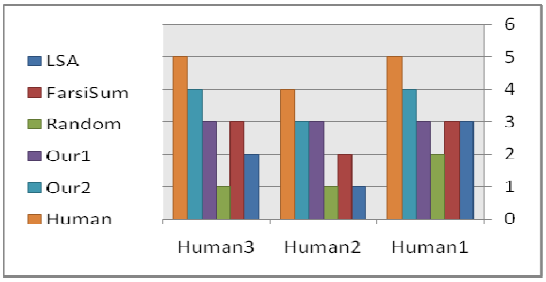

Figure4.The results of human evaluation.

CONCLUSION AND FUTURE WORKS

This paper presents a new multi-document summarization method using LSA and WordNet based semantic similarity for Persian. In this work, unlike previous methods, the sentences are selected based on the main context hidden in all documents of a topic. In previous approaches, the focus is on sentence features (local view) as the main and basic unit of text. Main context is extracted using LSA. We have also used SRL and WordNet to calculate semantic similarity between ranked sentences and redundancy elimination. The experimental results show that a global view in multi-document summarization can improve precision significantly.

ACKNOWLEDGMENT

This work was supported by the Web Technology Laboratory of Ferdowsi University of Mashhad. We would like to thank linguistics and WTLAB group members who helped us in this work.

REFERENCES

[1] I. Mani. *Automatic summarization*. John Benjamins Publishing Company, 2001.

[2] X. Wan, J. Yang, and J. Xiao. Manifold-ranking based topic-focused multi document summarization. In *Proceedings of IJCAI,* 2007.

[3] Amini, M. R., & Gallinari, P. The use of unlabeled data to improve supervised learning for text summarization. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* , SIGIR`02, 2002.

[4] Yeh, J. Y., Ke, H. R., Yang, W. P., & Meng, I. H. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing and Management, 41, 75-95,* 2005.

[5] Gong, Y., & Liu, X. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval , SIGIR`01,* New Orleans, 2001.

[6] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. *Indexing by latent semantic analysis*. Journal of the American Society for Information Science and Technology(JASIS), 41(6):391-470, 1990.

[7] C. H. Papadimitrious, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing:A probabilistic analysis. *J. Comput. Syst. Sci., 61(2):217-235, 2000.*

[8] Steinberger, J., & Kabadjov, M.A. & Poesio, M., & Sanchez-Graillet,O . Improving LSA-based summarization with anaphora resolution. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 2005.

[9] N. Mazdak, "FarsiSum-a Persian text summarizer", *Master thesis, Department of linguistics, Stockholm University*, 2004.

[10] H. Dalianis, "SweSum A Text Summarizer for Swedish", *Technical report", TRITANA-P0015, IPLab-174*, 2000

[11] Z. Karimi, M. Shamsfard, "Summarization of Persian texts" *In Proceedings of 11th International CSI computer Conference , Tehran, Iran,* 2006

[12] A. Zamanifar, B. Minaei, M. Sharifi, "A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of the Text*", Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, DOI 10.1109/SNPD.*2008.57

[13] D. Arnold, L. Balkan, S. Meijer, R. Humphreys, and L. Sadler. Machine Translation: an Introductory Guide. *Blackwells-NCC,* 1994.

[14] C. –Y. Lin and Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. *In Proccedings of NLT-NAACL, 2003.*