Extractive Summarization Of Farsi Documents Based On PSO Clustering

Mehdi Bazghandi[1], Ghamarnaz Tadayon Tabrizi[2] and Majid Vafaei Jahan[3]
[1] Islamic Azad University, Mashhad Branch
Mashahd, Iran

[2] Islumic Azad University, Mashhad Branch
Mashahd, Iran

[3] Islumic Azad University, Mashhad Branch
Mashahd, Iran

Abstract

As there is an ever-increasing number of textual resources, users nowadays enjoy access to a wider range of data; hence, accessing accurate and reliable ones has become a problematic issue. Automated summarization systems can play a principal role in covering the main ideas of the texts and removing time limitations. The present study presents a textual summarization system based on sentence clustering. There are some methods proposed to solve clustering problems so that for reaching a desirable 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. Ultimately, appropriate sentences are selected from each cluster after clustering the sentences on the basis of the aforementioned criteria. A collection of Persian sports news articles are selected for the assessment. The findings reveal that the presented method yields more accurate results than others.

Keywords: *Summarization, clustering, semantic similarity, efficiency, PSO algorithm*

1. Introduction

With an increase in the amount of data available on the web, cumulative rise of news websites, publication of various electronic books, and a significant growth in the number of published articles in different fields of study, one main problem of researchers in the 21st century has been that of accessing accurate and reliable data. The vast body of data bases on the one hand and time limitation on the other have directed the researchers to the interesting area of summarizing texts.

Automated document summarizers produce a summarized version of the main document by a computer program while keeping its main features and points [1]. The ultimate goal of summarizing is producing summaries which could compete human summarization. Yet, there are numerous challenges the most important of which at the first step is selecting the key sentences of the main text so that it covers the main ideas of the whole text and the same time does not repeat the same or similar sentences. This is the main focus of summarization systems. Text summarizing methods are categorized from different aspects [2]. Summarization methods differ with respect to different input, the object of summarization, and the type of the output required. Summarizers are divided into single-document and multi-document categories regarding the type of the input. In single-document summarization models, the input of the summarization system is just one document [1]. Single-document summarization is much less complicated than the multi-document one since it only deals with one single document which continuously discusses one topic and does not entail paradoxical sub-topics. Many methods have been presented for single-document summarization some which are listed in [3].

In multi-document summarization, the input encompasses multiple documents; it is closely tied to answering systems and search-based summarization [4]. Regarding the output, summaries are divided into two major categories of abstractive and extractive. In extractive summarization, which also forms the basis of most of abstractive summarization systems, some parts of the text are often selected (at sentence level) and then arranged as the summary. Most methods follow the same principle. In abstractive summarization, the structure of the sentence could be altered besides selection thereof. In this method, sentences could be omitted or changed or even new sentences could be generated. It should be noted that this method is very complicated and even more complicated than 'machine translation'. Text summarization systems emerged in 1950's which focused on the form of the text such as the position of the sentences in the text due to lack of powerful computers and the problems in natural language processing [NLP]. Artificial intelligence appeared in 1970's [5]. The idea behind AI was extraction of knowledge such as frames or patterns for identifying implied entities of the text and extraction of the relation among the entities using conceptual mechanisms; the main problem was that the frame or the patterns have some limitations and lead to incomplete analyses of the conceptual entities. Since 1990's, information retrieval (IR) has been used.

2. Semantic Relations Between Words

To begin with, the corpus needs to be analyzed and its candidates be extracted. In English such tools as TNT-Tager and SENTA could be used to extract compound and non-compound words. Since Persian lacks such language tools and resources, in the present study the candidates were manually stored in a file. This is done to determine the relations and the extent of similarity among the terms and store thereof in the similarity matrix. An N-dimensional vector with maximum relations with the term is envisaged for each of the candidates. Therefore, the highest co-occurrences of this term with the others should be found. Symmetric Conditional Probability offers a method which enables us to estimate the relations and co-occurrence of the terms [6] through the following equation.

$$SCP(w_1, w_2) = \frac{p(w_1, w_1)}{p(w_1)\,p(w_2)} \quad (1)$$

Where $p(w_1, w_1)$ is the possibility of co-occurrence of two words in a particular context. $p(w_1)$ and $p(w_2)$ respectively represent the possibility of words and alone. In the application stage, a text window of 20 words is given.

The words that appear in the context vector of a specific word have various relations with it. After extracting the context-vectors of some of the candidates, the extent of the similarity of the two words in the space vector is established, and if the two vectors are close (by comparing the two vectors based on cosine criterion), it means that the two words are similar and are used in similar texts and that their relations are also similar whereas the words that appear in the context vector of a specific word have various relations with it [7]. The basic idea of our informative similarity measure is to integrate into the cosine measure the word co-occurrence factor inferred from a collection of documents with the Equivalence Index association measure. This can be done straightforwardly as defined in equation 3 where $Coh(w_{i,k}, w_{j,l})$ is the Equivalence Index value between $w_{i,k}$, the word that indexes the vector of the document i at position k, and $w_{j,l}$ the word that indexes the vector of the document j at position l. In fact, the informative similarity measure can simply be explained as follows. Let's take the focus sentence x_i and a block of sentences x_j., the similarity of the two sentences is estimated using the equations (2,3).

$$S_{i,j} = f(x_i, x_j) \quad (2)$$

$$InfosimBA(x_i, x_j) = \frac{A_{I,J}}{B_i * B_j + A_{i,j}} \quad (3)$$

Where

$$A_{i,j} = \sum_{k=1}^{p}\sum_{l=1}^{p} X_{i,k} * X_{j,l} * coh(w_{i,k}, w_{j,l})$$

$$\forall i, B_i = \sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{i,k} * X_{j,l} * SCP(w_{i,k}, w_{j,l})}$$

In this equation, $X_{I,J}$ is related to $W_{I,J}$. $Coh(w_{i,k}, w_{j,l})$ shows the extent of the similarity of the words based on co-occurrence. When the criterion for closeness of two context-vectors of two words is to be used, $SCP(w_{i,k}, w_{j,l})$ will be applied (the extent of the similarity of the related vectors will be stored in Coherence matrix). Due to the process equation estimated using (4)

$$Tf - Isf(w, s) = \frac{stf(w, s)}{|s|} * \log_2 \frac{N_S}{sf(w)} \quad (4)$$

where $stf(w, s)$ represents the frequency of the word in the sentence, |s| shows the sentence length, and N_S shows the total number of the sentences in the text, and $sf(w)$ represents the number of the sentences in which w occurs.

3. Clustering

Automated clustering algorithms assign a set of items in a cluster so that the best number for clustering is determined by the algorithm. The items within a cluster must bear the highest possible similarity while also be different from the other items, in other words, the two following points should be observed [8]: The sentences should be clustered in a way that the most similar sentences fall within the same cluster (intra-cluster function) , The sentences within a cluster should be distant from other sentences in other clusters (inter-cluster function). We use the different methods that are defined in (5)–(9). These methods optimize various aspects of intra-cluster similarity (5), inter-cluster dissimilarity (6) and heir combinations (7)–(9). These methods are defined as follows:

$$IntraSim = \left(\sum_{p=1}^{k} c_p \sum_{s_i, s_j} sim(s_i, s_j)\right) \rightarrow MAX \quad (5)$$

$$InterSim = \left(\sum_{p=1}^{k} c_p \sum_{s_i, s_j} sim(s_i, s_j)\right) \rightarrow MAX \quad (6)$$

$$M = IntraSim * 1/InterSim \rightarrow MAX \quad (7)$$

$$CL = w_2 * IntraSim + (1 - w_2) * {}^{1}\!/_{InterSim} \rightarrow MAX \quad (8)$$

$$H = \frac{2}{\frac{1}{P} + R} = \frac{2P}{1 + PR} \rightarrow MAX \quad (9)$$

The combinational function M is the multiplication of $IntraSim$ and $IntrerSim$, function CL is the combined weight of the two functions, and function H is the average harmonic of the two functions. Also function CL shows the extent of the collaboration of functions $IntraSim$ and $InterSim$; then if w_2 is equal to zero, function CL^{-1} will represent $InterSim$ function; if it is equal to 1, it represents $IntraSim$; if it is equal to 0.5, then both of the functions ($InterSim$ and $IntraSim$) will have equal shares. In these functions, K shows the number of the clusters (regarding that the rate of summarization is determined by the user and only one sentence is chosen from each cluster, the number of the clusters will equal the number of the summarized sentences). C_p is the number of the sentences in the cluster. Finally, $sim(s_i, s_j)$ is the extent of the similarity stored in the similarity matrix (Similarity of sentence pairs in the texts according to " $InfosimB$" of the candidates terms).

3.1. The Poposed Clustering Based On PSO

Particle swarm optimation algorithm is a complementary processing method to optimize non-linear functions modeled from the social behavior of birds.

In this defining structure, the length of each particle is defined as the number of the sentences of the text so that a number must be chosen from {1, 2, 3, …K} for each sentence. In this structure all the K numbers must be used. Each sentence has a unique number. The numbers are not unique along the particle length, though. Therefore, the numbers in each particle (with regards to the number of clusters) must be chosen in a way that the number which represents the number of the cluster is repeated, but all K numbers must be used [8]. In this study, the structure of each particle is defined as in Figure 1.

1	1	3	2	2	1	3	4	4

Fig. 1 The structure of a particle

This particle shows that sentences 1, 2, and 6 are in the first cluster; sentences 3 and 7 are in the third cluster; sentences 4 and 5 are in the second cluster; sentences 8 and 9 are in the fourth cluster. The fitness function used in the algorithm includes one intra cluster, one inter cluster and one combinational function from section 3. One point to be noted in this algorithm is that the particles produced should not yield values beyond the aforementioned conditions. Yet this is inevitable in PSO algorithm which arises owing to velocity vectors of the particle, in other words, the values might go beyond the limitations (or even appear in decimal forms). To overcome such problems, genetic algorithm and a vector called mutation vector could be used as follows.

$$m_{i,j}(t+1) = \begin{cases} 1, if Rand_j \le sigm(v_{i,j}(t+1)) \\ 0, otherwise \end{cases} \quad (10)$$

Influencing on $x_{i,j}(t)$,the $m_{i,j}(t+1)$ changes this vector to $x_{i,t}(t+1)$. If the Jth element of vector is equal to 1, the Jth element of $x_{i,j}(t)$ is transferred to the Jth element of $x_{i,t}(t+1)$ without any change; and if the Jth element of $m_{i,j}(t+1)$ is equal to zero, the Jth element of $m_{i,j}(t+1)$ will have a mutation which is the effect of the reverse function on it. Figure 2 shows the formation of $x_{i,t}(t+1)$ using $m_{i,j}(t+1)$.

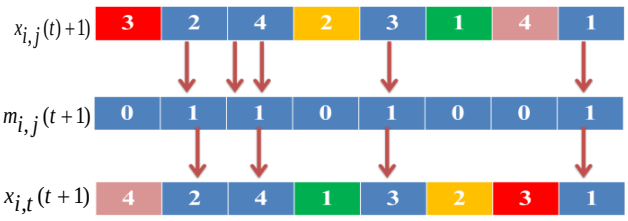

Fig.2 mutation vector

4. Experimental Results

A set of sports news from ISNA (Iranian Students News Agency) is chosen as the corpus of the study. All the assessments and tests have been administered on the 8 subsets. This set entails 2-4 sports news. An attempt has been made in this study to use [9] assessment method which uses examiners instead of recalling criteria. In this method some examiners are asked to score the sentences of a text in terms of their importance (with 1 being the least important and 10 being the highest score representing the most important sentence). The initial population of the particles is 40 (N=40). The number of repetitions is 250 (t=250), and the cognitive and social parameters are $c_1 = 2.5, c_2 = 4 - c_1$ and inertia is w_2 =0.5. All the answers are recorded, and the average number of the

repetition of each method is regarded as 10, and in all these methods, summarization is done with the rate of 50%. The proposed method will be compared with the following methods. Clustering based on PSO , Clustering based on Kmeans, PSO Three examiners are asked to score the sentences of a Set of news and the summary of the text is generated using each of the methods.

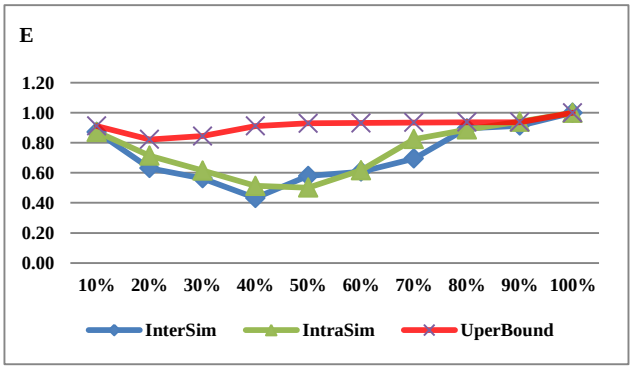

Fig.3 The efficiency of IntraSim and InterSim methods are compared
According to the graph in Figure 3, neither of the methods has any priority over the other (Considering the rate).

Meth	Rough	Intra	Inter	M	CL	H	PSO	PSO C
Intra	0.201	X	+0.99	+1.4	+0.99	+2.9	-0.49	+1.43
Inter	0.203	-0.98	X	+0.49	0	+1.97	-1.47	+0.49
M	0.204	-1.47	-0.49	X	-0.49	+1.47	-1.96	0
CL	0.203	-0.98	0	+1.4	X	+1.97	-1.47	+0.49
H	0.207	-2.89	-1.93	-1.44	-1.93	X	-3.38	-2.89
PSO	0.200	+0.50	+1.50	+1.96	+1.50	+3.50	X	+1.96
PSO-C	0.204	-1.47	-0.49	0	-0.49	+1.47	-1.96	X

Table 1:Evaluatoin of maethods , among our methods the best result obtained by the hybrid function H. In spite of the fact that, among our methodsthe worst result is obtained by the method *InterSim.*

In Tables 1 "+" means the result outperforms and " – " means the opposite.The efficiency of method CL which is chose as 0.35 here. The average of each of the methods on the all of news set is.

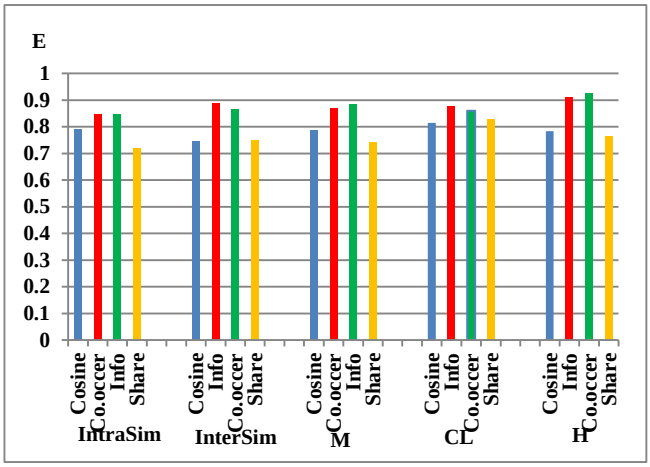

Fig. 4 The comparison of criteria

According to Figure 4, the Co-Occerance and InfoSIm criteria in all ways a better performance than the cosine measure.

5. Conclusions

An important issue in summarization systems is the set of the documents used as input including the information relevant or irrelevant to the main topic of the text. A cluster-based method was proposed in this study for text summarization whose advantages include: an alternative method was used instead of cosine distance to identify the similarity between the sentences. In addition to being costly, cosine distance takes only the form of the words into account (K-Means, PSO Clustering methods). However, in the proposed method, the distance of the words is estimated using context-vector similarity (InterSim, IntraSim, M, CL, H). Moreover, method observes the physical distance. most methods, initially focus on mere clustering of the documents using clustering methods, and then extract the main sentences using ranking (in a cluster of related documents). Yet, in the preset study sentence clustering is carried out only after documents clustering (PSO). an important consideration in this study is optimization of the presented functions. An attempt was made to use PSO algorithm in optimizing the functions.

References

[1] Mani, I. & Maybury, M. T. (Eds.). "Advances in automated text summarization". Cambridge, MA: The MIT Press, 1999.

[2] Hovy, E. & Lin, C. Y. "Automatic text summarization in SUMMARIST". In Proceedings of the ACL˝97/EACL˝97 workshop onintelligent scalable text summarization (pp. 18–24), Madrid, Spain, 1997.

[3] Mihalcea, R. and Tarau, P. "An Algorithm for Language Independent Single and Multiple Document Summarization". In Proceedings of the International Joint Conference on Natural Language Processing, Korea, 2005.

[4] Hirao, T., Sasaki, Y., Isozaki, H. "An extrinsic evaluation for question-biased text summarization on qa tasks". In Proceedings of NAACL workshop on Automatic summarization, 2001.

[5] DeJong, G. F. "Skimming stories in real time: an experiment in integrated understanding". Doctoral Dissertation. Computer Science Department, Yale University, 1979.

[6] Silva, G. Dias, S. Guillor´e and J.G.P. Lopes. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units.In Proceedings of 9th Portuguese Conference in Arti- cial Intelligence. Springer-Verlag. 1999.

[7] Moshki.Analoei," Extractive Multi Document Farsi Text Summarization" irst National Conference on Software Engineering, University Roodehen, 1388".

[8] Ramiz Aliguliev." Clustering Techniques And Discrete Particle Swarm Optimization Algorithm".Computational Intelligence, Volume 26, Number 4, 2010.

[9] D. R. Radev, H. Jing, M.¬Stys, and D. Tam (2004), Centroid-based summarization of multiple documents. Information Processing and Management, 2004