International Journal of Computer Networks and Communications Security

VOL. 3, NO. 11, NOVEMBER 2015, 421–423
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

A New Method for Text Segmentation in Persian Based on Lexical Cohesion

Selma Mokhtar Zadeh shahraki[1] and Mashallah. Abbasi Dezfouli[2]

[1, 2] Department of Computer, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran.

E-mail: [1]mok.s.computer@gmail.com, [2]abbasi_masha@yahoo.com

ABSTRACT

In this paper, we present a new segmentation algorithm based on Latent Semantic Analysis for segmentation of Persian texts. The presented algorithm is fully automatic, without training and based on lexical cohesion and it performs segmentation using semantic relationship between the blocks. Evaluation of results shows that our algorithm acts better than unsupervised Persiantiling segmentation method and F_measure with 70.97% had a significant improvement.

Keywords: *Latent Semantic Analysis, Text Segmentation In Persian, Unsupervised Algorithm, Untrained Segmentation, Persiantiling Algorithm, Evaluation Criteria.*

1 AN OVERVIEW OF B&A SPY AGENCY

A text document contains a series of words that are meaningful by placing together and form and express a more specific structure and content. Today, with increasing rate of document, it is difficult to find a specific content from a document, so the methods and tools for easier access to these documents are necessary. Using text segmentation tools, we can provide a favorable conditions for easier access to a certain content in a text. So one of the applications of such tools and techniques is in information retrieval and summarization systems.

A variety of methods are presented for text segmentation in other languages, including English, mostly based on word, [1], [2], [3], [4], [5], [6], [10] are examples of such systems. In the Persian language, there are some methods that could divide a text document to the integrated and discrete parts, [7], [8], [9], but unfortunately no activity had done in this area, now we provide a new way to improve the performance of Persian text segmentation.

2 PROPOSED ALGORITHM

Our proposed algorithm called TSUL, is a vector space model and uses the semantic relationship between the blocks to identify the boundary between the subjects. In the Figure (1) we show the general framework of TSUL algorithm.

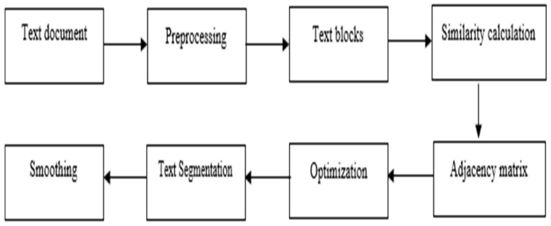

Fig. 1. General framework of TSUL algorithm

In most cases, the text contains numbers, symbols and words which presence in the text which only use memory and prolong the system output time. As a result, before evaluating the TSUL proposed algorithm and Persiantiling, we need to normalize the input samples. To normalize the text, we convert the text document that we want to segment it to building tokens, we then split the text into the building blocks of the same size. Each building block can be sentence, paragraph or any particular size of a text that explains a specific issue. Since a paragraph can contain several different topics, we here consider the size of each building block as a sentence. After blocking out the text, we calculate the similarity of blocks using equation (1).

$$\cos\left(D_i, D_j\right) = \frac{\sum_{k=1}^{n} W_{ik} W_{jk}}{\sqrt{\sum_{k=1}^{n} W_{ik}^2 \sum_{k=1}^{n} W_{jk}^2}} \qquad (1)$$

Where, D_i, D_j are two blocks of text and $W_{ik}$is the weight of k word in i-th block. n here is the number of unique words in the text and since in the obtained matrix, many blocks are not similar, we calculate the SVD to reduce and optimize the semantic space. Now, using the optimized matrix called Λ_k, we calculate the characteristic vectors according to equation (2).

$$\lambda_i = \sum_j S_{ij} * \Lambda_k(j) \qquad (2)$$

In the above equation, S_{ij} is the similarity of two vectors.Finally, to identify the final piece, we use the cosine standard equation, and using equation (3), we smooth the output graph. The smoothing method works so that we consider a fixed paired width of (r) for any g gap and for every gap, we calculate the mean values using the smoothing equation and considering $\frac{r}{2}$ on the left and the right, and then we attribute the obtained value to the gap [6].

R value is selected according to the database and the length of documents, that for big database, a great r would be appropriate and for small databases and documents such as newspaper and articles, small r value would be appropriate [4]. Considering our database and evaluation of the results obtained from the proposed method TSUL, we saw that the best results obtained when we consider that r equals 2.

Di = (Si-1 − Si) + (Si+1 − Si) = Si + Si+1 − 2 Si (3)

Comparing and calculating the semantic relationship between the blocks, we have been able to correct the shortcomings of Persiantiling method when coping with low number of words along with another subject with high number of repetitions of words and by smoothing the output graphs, we improved the identification of actual boundaries.

3 EVALUATION

3.1 Evaluation Criteria

Standard evaluation criteria is as follows that we used to evaluate the two algorithms: TSUL and Persiantiling.

$$Recall = \frac{number\ of\ borders\ system\ accuratly\ diagnozed}{number\ of\ system\ real\ borders} \quad (4)$$

$$\text{precision} = \frac{number\ of\ borders\ system\ accuratly\ diagnozed}{number\ of\ borders\ system\ diagnozed} \quad (5)$$

$$F\_measure = \frac{2 * recall * precision}{recall + precision} \quad (6)$$

3.2 Trial Set

To evaluate the proposed algorithms, we require test samples to examine the Precision and efficiency of algorithm to discover the border between issues. Reviewing the data set in English and considering the basic features such as a variety of topics, closeness of the topics to each other conceptually, the length of sentences (sentences with long and short length) and a plurality in the topics, we expanded the first data set made in the Persian language. The samples were extracted using the Fars news agency online newspaper due to observing the standards for writing. The text are extracted from different dates.

3.3 Test Results

We have tested TSUL and Persiantiling algorithms on 50 different samples and compared it with reality and calculated the value of 3 criteria, recall, Precision and F_measure. For comparison between methods, we calculated the average of 50 samples for 2, 3 and 4 sentence blocks and drew their charts.

Table 1 shows the output of TSUL method. According to the obtained values and results of Persiantiling algorithm output shown in Table 2 it specifies that the percentage of recall with the value of 67.97 was significantly increased in all the blocks and the Precision with a value of 87.25 percent had the maximum Precision in determining boundaries. The F_measure with 70.97 is the maximum result obtained in all blocks.

Table 1: Mean of TSUL method

	Mean of results		
Number of sentences	Recall	Precision	F measure
2	67.97	74.26	70.97
3	57.84	75.28	65.41
4	59.08	87.25	70.45

Table 2: Mean of Persiantiling method

	Mean of results		
Number of sentences	Recall	Precision	F measure
2	62.88	46.32	53.33
3	61	60.95	60.97
4	54.41	78.46	64.25

The results of the two methods, TSUL and Persiantiling are drawn in Figures 1 and 2. The horizontal axis is the number of sentences within each building block and vertical axis is the number of criteria as percent.

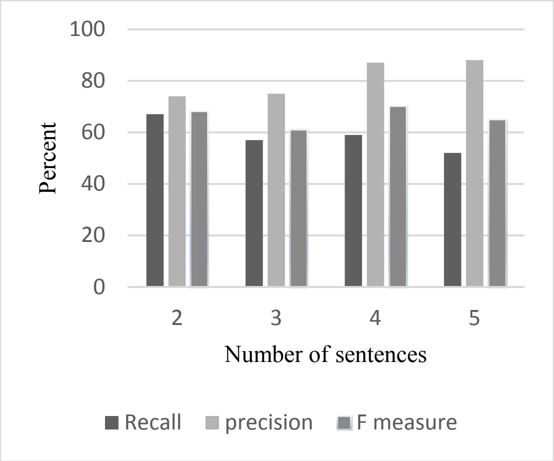

Fig. 1. Diagram of average measures for TSUL method

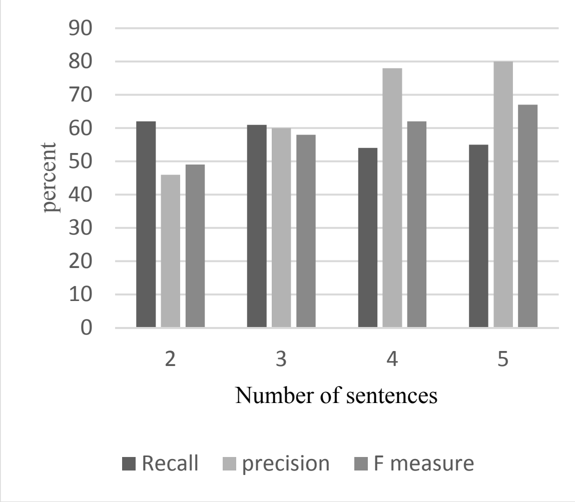

Fig. 2. Diagram of average measures for Persiantiling method

4 CONCLUSION

Given the importance of segmentation in the field of learning of computer analysis machine, many methods are presented to explore the boundary between the topics in the text. Having applied some changes in the hidden meaning analysis algorithm, we could present a new algorithm in accordance with the structure of Persian language. According to the results on 50 samples, we can understand that TSUL had a significant improvement compared to Persiantiling when we consider any building block as two 2 sentences.

5 REFERENCES

[1] A. Alemi, P. Ginsparg, Text Segmentation based on Semantic Word Embeddings. arXiv journal , Sydney, Australia,18 Mar 2015

[2] B. Dadachev, A. Balinsky, H. Balinsky, On Automatic Text Segmentation. Proceedings of the 2014 ACM symposium on Document engineering, Colorado USA, 16-19 sep 2014, pp: 73-80

[3] FY. Choi, P. Wiemer-Hastings, J. Moore, Latent semantic analysis for text segmentation, In Proceedings of EMNLP. Citeseer, 2001

[4] b. Fitzgerald, Implementation Of An Automated Text System Using Hearst's Texttiling Algorithm. 1 june 2000

[5] MA. Hearst, TextTiling : A quantitative approach to discourse segmentation Technical Report. University of California, Berkeley, Sequoia, 1993

[6] MA. Hearst, X. Parc, Texttiling: Segmenting Text Into Multi-Paragraph Subtopic Passages. Comput. Linguist. 1997, 23: 33–64

[7] S. Mokhtar zadeh shahraki, Unsupervised Persian Text Segmentation, M.Sc thesis, Islamic Azad University Science and Research Bushehr Branch Faculty of Computer, 2012

[8] s. Mokhtar zadeh shahraki, m. Sadeghzadeh, Ph.D., R. Dianat, Ph.D., A Method for Unsupervised Text Segmentation in Persian, First National Conference on Science and Computer Engineering, Shiraz Islamic Azad University of Shiraz, November 2012

[9] s. Mokhtar zadeh shahraki, m. Sadeghzadeh, Ph.D., M. Bahrani, A comparative study of algorithms for text segmentation based on lexical cohesion in Persian texts , Eighteenth Annual National Conference of Computer Society of Iran, Tehran, Sharif University of Technology, March 2012, pp: 24-22

[10] JC. Reynar, Topic segmentation: Algorithms and applications. Ph.D. thesis, Computer and Information Science, University of Pennsylvania, 1998.