Ciência e Natura, v. 37 Part 2 2015, p. 247–252

ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

ciência e natura

An Imperialist Competitive Algorithm for Persian Text Segmentation

Nooshin Shirali[1], Marjan Abdeyazdan[2]

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

Abstract

Text Segmentation has been used in different natural language processing tasks, such as information retrieval and text summarization. In this paper a novel Persian text segmentation algorithm is proposed. Our proposed algorithm applies the imperialist competitive algorithm (ICA) to find the optimal topic boundaries. It is the first time that an evolutionary algorithm applies in Persian text segmentation. The experimental results show that proposed algorithm is more accurate than other Persian text segmentation algorithms.

Keywords: *Text Segmentation, Persian Text Segmentation, Imperialist Competitive Algorithm, Natural Language Processing.*

Recebido: dia/mês/ano Aceito: dia/mês/ano

1 Introduction

Text segmentation can be defined as the task of breaking text into topically coherent multi-paragraph subparts. Text segmentation is important to several natural language processing tasks, such as information retrieval and text summarization. In information retrieval for example, having topically segmented documents can result in the retrieval of short relevant text segments that directly correspond to a user's query instead of long documents examined by a user carefully in order to find the object of his/her interest. In text summarization, linear text segmentation enables the system to summarize subtopics of a document instead of the whole document.

While extensive research has targeted this technique in English, few have studied it in other languages and almost no one except (Mokhtarzadeh et al., 2013a; Mokhtarzadeh et al., 2013b), has addressed it for Persian language. The lack of research in this topic pushed us to present an evolutionary based algorithm for such language.

In this paper a novel domain independent text segmentation algorithm is proposed. Imperialist competitive algorithm is used to find the optimal topic boundaries. To the best of our knowledge, ICA has not yet been applied to linear text segmentation. It is expected that ICA will do as well in linear text segmentation as in other research areas. We will prove this by some experiments in section 5. The experimental result show that proposed algorithm is more accurate than other Persian text segmentation algorithms.

2 Previous works

There are several classical approaches that have been proposed for text segmentation. Considering English and Persian languages, existing techniques can be divided into two categories: text segmentation for English language and text segmentation for Persian language.

2.1 Text segmentation for English language

Approaches that address the problem of text segmentation can be classified into knowledge-based approaches or word-based approaches. Knowledge-based systems (Grosz and Snider, 1986), require an extensive manual knowledge engineering effort to create the knowledge base (semantic network and/or frames) and this is only possible in very limited and well-known domains. To overcome this limitation, and to process a large amount of texts, word-based approaches have been developed. Word-based approaches can also fall into two main groups: lexical cohesion-based approaches and feature-based approaches. For example Morris and Hirst (1991) described a text segmentation algorithm based on lexical cohesion relations, In the TextTiling approach of Hearst (1997), blocks of words were represented as word count vectors and similarity between adjacent blocks was measured using dot-product in the vector space. In Kozima (1993), ''semantic similarity between words'' (called as Lexical Cohesion Profile (LCP)) was measured by manually designing a semantic network from a small English dictionary and spreading activation on this network. LCP was used for finding the segment boundaries. Richmond et el. (1997) describe a technique for locating topic boundaries. Their method weights the importance of words based on their frequency within a document and the distance between repetitions. They determine the similarity between neighboring regions of text by summing the weights of the words which occur in both regions and then subtracting the summed weights of words which occur only in one segment. They normalize this figure by dividing by the number of words in each section. C99 proposed by Choi (2000) is based on lexical cohesion and it uses the cosine metric to compute similarity between sentences. Later, Choi (2001) improved his algorithm by using Latent Semantic Analysis (LSA) to extract semantic knowledge from text. Lexical cohesion can also be added as a feature in the feature-based approach, as exemplified by work presented in (Reynar, 1998).

Moreover there are few remarkable works which used evolutionary algorithms to find optimal topic boundaries. Lamprier et el. (2007) proposed a genetic algorithm called SegGen, Wu et el. (2010) used a discrete particle swarm optimization algorithm for finding text segments, which called DPSO-SEG, Later they improved they algorithm by using hierarchical agglomerative clustering and proposed a hybrid algorithm called TSHAC-DPSO (Wu et al., 2014).

2.1 Text segmentation for Persian language

Due to importance of text segmentation in natural language processing, there are almost no remarkable work in Persian language except (Mokhtarzadeh et al., 2013a; Mokhtarzadeh et al., 2013b). Mokhtarzadeh et el. (2013.a), adapted Reynar dotplot algorithm (Reynar, 1998) to the Persian language. They performed their algorithm on 22 sample of Persian text. In (Mokhtarzadeh et al., 2013.b) Mokhtarzadeh et el., adapted TextTiling algorithm (Hearst, 1997) to the Persian language, their algorithm called Persiantiling.

Recebido: dia/mês/ano Aceito: dia/mês/ano

3 Imperialist competitive algorithm

Imperialist competitive algorithm (ICA) has been introduced by Atashpaz and Lucas (2007). ICA is one of the evolutionary population based optimization and search algorithms. The source of inspiration of this algorithm is the imperialistic competition. ICA has good performance in both convergence rate and better global optimum achievement. The ICA formulates the solution space of the problem as a search space. This means each point in the search space is a potential solution of the problem. The ICA aims to find the best points in the search space that satisfy the problem constraints.

An ICA algorithm begins its search and optimization process with an initial population. Each individual in the population is called a country. Then the cost of each country is evaluated according to a predefined cost function. The cost values and their associated countries are ranked from lowest to highest cost. Some of the best countries are selected to be imperialist states and the remaining form the colonies of these imperialists. All colonies of the population are divided among the imperialists based on their power. Obviously more powerful imperialists will have the more colonies. The colonies together with their relevant imperialists form some empires. The ICA contains two main steps that are assimilation and imperialistic competition.

During assimilation step, colonies in each empire start moving toward their relevant imperialist and change their current positions. The assimilation policy causes the powerful empires are reinforced and the powerless ones are weakened. Then imperialistic competition occurs and all empires try to take the possession of colonies of other empires and control them. The imperialistic competition gradually brings about a decrease in the power of weaker empires and an increase in the power of more powerful empires. In the ICA, the imperialistic competition is modeled by just picking some of the weakest colonies of the weakest empire and making a competition among all empires to possess these colonies. The assimilation and imperialistic competition are performed until the predefined termination conditions are satisfied.

4 The proposed algorithm

Imperialist competitive algorithm is widely used to solve different problems. In this paper we used ICA to identify the optimal topic boundaries of text segments in a document. At first basic preprocessing applied to the text, punctuation and the not useful words are eliminated by using a list of stop words specific to the Persian language (Davarpanah et al., 2009). After which each sentence is represented as a term frequency vector. Then the similarity between a pair of sentences is computed using cosine similarity, this is applied to all sentence pairs to generate a similarity matrix. Finally the optimal boundaries are created by ICA-SEG according to the sentence similarity matrix. In order to improve the performance of our algorithm we used the optimal boundaries which are found by Choi's C99 algorithm (which we applied on Persian language).

4.1 Sentence similarity matrix

After preprocessing, each sentence is represented as a term frequency vector. Let $f_{i,j}$ denote the frequency of word j in sentence i. The similarity between a pair of sentences x, y is computed using the cosine similarity:

$$sim(x,y)=\frac{\sum_j f_{x,j}\times f_{y,j}}{\sqrt{\sum_j f_{x,j}^2\times\sum_j f_{y,j}^2}} \quad (1)$$

Next, the sentence similarities are illustrated by considering a theoretical matrix – sentence-sentence similarity matrix. The sentence similarity matrix is constructed by computing all pair-wise sentence similarities.

4.2 Cost function definition

When dividing a long text into several topical segments, the sentences within each segment should cover the same subtopic. Moreover, sentences among different segments should belong to different subtopics. Therefore, both the average sentence similarity within each segment (inner similarity) and the average sentence similarity between two consecutive segments (outer similarity) are considered in the cost function. The inner similarity and the outer similarity are then combined as follows:

$$f=\frac{1}{\#seg}\times\sum_{i=1}^{\#seg}Internal(i)-\frac{1}{\#seg-1}\times\sum_{j=1}^{\#seg-1}External(j,j+1)+STVDEV\times\alpha+\frac{\#seg}{\#sentence}\times\beta \quad (2)$$

Where α and β are weighting factors, which we set to 1.5 and 0.5 respectively.

4.2.1 Internal cohesion

In a good segmentation, sentences within a segment should cover the same subtopic. Hence, the goal is to find the segmentation with the minimum internal (lexical) cohesion. The internal cohesion can be represented as:

$$Internal(i)=\frac{\sum_{S_m\in seg_i}\sum_{S_n\in seg_i}Sim(S_m,S_n)}{|seg_i|\times|seg_i|} \quad (3)$$

Recebido: dia/mês/ano Aceito: dia/mês/ano

Internal(i) represents the internal density of the segment i; seg_i denote segment i and $|seg_i|$ denote the number of sentences in seg_i; the numerator denote the total value in the sentence similarity matrix corresponding to the seg_i and $|seg_i| \times |seg_i|$ is the area of the sub-matrix.

4.2.2 External dissimilarity

In a good segmentation, sentences among different segments should belong to different subtopics. Hence, the goal is to find the segmentation with the maximum external (two consecutive segments) dissimilarity. The external dissimilarity is represented as:

$$External(j, j+1) = \frac{\sum_{S_m \in seg_j} \sum_{S_n \in seg_{j+1}} Sim(S_m, S_n)}{|seg_j| \times |seg_{j+1}|} \quad (4)$$

External(j, j+1) is the dissimilarity between segment j and segment j+1; the numerator denote the total value in the sentence similarity matrix corresponding to the adjacent segments seg_j and seg_{j+1}; $|seg_j| \times |seg_{j+1}|$ is the area of the sub-matrix.

4.2.3 Standard deviation

Generally speaking, it is usually the case that the number of sentences in each segment is similar. For this reason, the goal is to minimize the standard deviation of the segment length in a text. The standard deviation of the segment length is represented as:

$$STDEV(k) = \sqrt{\frac{\sum_{k=1}^{\#seg} \left(|seg_k| - \mu \right)^2}{\#seg}} \quad (5)$$

STDEV(k) is the standard deviation of the segment length in a text; μ is the mean length of the segments; #seg is the number of segments.

4.2.4 Number of segments

According to our observation, owing to the factor STDEV(k) is considered, the number of boundaries created by ICA-SEG is more than the actual number when the variance of segment length is greater. To cope with this problem, the number of segments is considered in the cost function. The number of segments is normalized as:

$$\frac{\#seg}{\#sentence} \quad (6)$$

4.3 Proposed algorithm

After constructing the sentence similarity matrix, ICA-SEG algorithm applies the principle of ICA as follows:

Step 1: The initial population for each empire should be generated randomly.
Step 2: Move the colonies toward their relevant imperialist.
Step 3: Exchange the position of a colony and the imperialist if the colony's cost is lower.
Step 4: Compute the cost function of all empires by.
Step 5: Pick the weakest colony and give it to one of the best empires.
Step 6: Eliminate the powerless empires.
Step 7: Repeating step 2 until the predefined number of iterations is exceeded.

4.4 Improving cost function

In order to improve the performance of ICA-SEG algorithm, we devise another algorithm called PersianICA-C99 which uses the distance between the optimal boundaries which founds by the C99 algorithm, Using the ranking method, and the optimal boundaries that founds by ICA-SEG as follows:

$$C = f + \lambda \frac{\sum D}{M} \quad (7)$$

Where λ is a constant which set to 0.1 and d is the distance between optimal boundaries which found by two mentioned algorithms.

After constructing a sentence similarity matrix, PersianICA-C99 algorithm applies the principle of ICA which is mentioned in section 4.3.

5 Experiments

Since there is no publicly available test corpus in Persian language, we used articles from a well-known Persian websites named farsnews, to evaluate the performance of our proposed algorithm. This sample consists of 1000 sentences which is randomly selected from the farsnews website.

The evaluation metrics we used is the Recall, Precision and F-measure metrics, which are defined as follows:

$$\text{Recall} = \frac{\text{Number of correctly system detected segments}}{\text{Total number of real segments}} \quad (8)$$

$$\text{Precision} = \frac{\text{Number of correctly system detected segments}}{\text{Total number of system generated segments}} \quad (9)$$

$$\text{F - measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (10)$$

Recebido: dia/mês/ano Aceito: dia/mês/ano

In order to evaluate the proposed algorithm, we compare it with C99 and ICA-SEG algorithms. The parameters in the ICA algorithm employed are set as Table 1.

Figure 1 shows the experimental result on each algorithm according to mentioned metrics. Table 2 compares these algorithms according to their F-measures. According to Table 2, PersianICA-C99 is demonstrated to provide a comparatively promising accuracy and it is more accurate than other presented algorithms.

Table 1. Parameter Settings of ICA

Parameter	Value
Population size	300
Number of Initial Imperialists	20
Number of iterations	500
Revolution Rate	0.1
Z	0.01
B	2

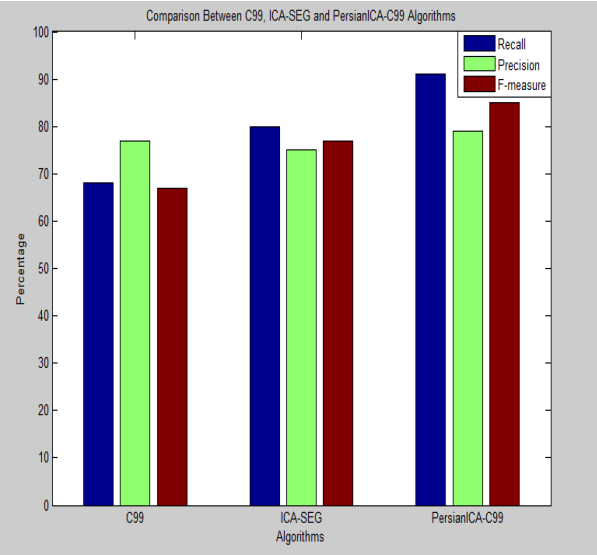

Fig. 1. Experimental result: C99, ICA-SEG and PersianICA-C99 algorithms

Table 2. Experimental result according to F-measure

Algorithm	F-measure
C99	67%
ICA-SEG	77%
PersianICA-C99	85%

6 Concolusion

In this paper, we presented a novel Persian text segmentation algorithm, PersianICA-C99. In Persian text segmentation, it is the first time that an evolutionary algorithm is applied. We evaluate the proposed algorithm with a data set based on articles from a Persian news website. The experimental results show that PersianICA-C99 algorithm provide a comparatively promising accuracy and it is more accurate than other presented algorithms

References

Atashpaz-Gargari, E., Lucas, C. (2007). Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. IEEE Congress on Evolutionary Computation, pp. 4661-4667.

Choi, F. Y. Y. (2000). Advances in Domain Independent Linear Text Segmentation. In Proceeding of the North American Chapter of the Association for Computational Linguistics, Seattle, USA, pp. 26-33.

Choi, F. Y. Y., Wiemer-Hastings, P. and Moore, J. (2001). Latent Semantic Analysis for Text Segmentation. In proceedings of EMNLP, Pittsburg, PA, USA.

Davarpanah, M. R., Sanji, M., Aramideh, M. (2009). Farsi Lexical Analysis and Stop Word List. journal of Library Hi Tech, 27(3), pp. 435-449.

Grosz, B., J., and Snider, C., L. (1986). Attention, Intentions and Structure of Discourse. Journal of Computational Linguistics, 12(3), pp. 175-204.

Hearst, M. A. (1997). Texttiling: Segmentation Text into Multi-Paragraph Subtopic Passages. Journal of Computational Linguistics, 23(1), pp. 33-64.

Kozima, H. (1993). Text Segmentation based on Similarity Between Words. In proceeding of the 31st Annual Meeting of the Association for Computational Linguistics, Student Session, pp. 286-288.

Lamprier, S., Amghar, T., Levrat, B. and Saubion, F. (2007). Seggen: A Genetic Algorithm for Linear Text Segmentation, in International Joint Conference on Artificial Intelligence, pp. 1647-1652.

Recebido: dia/mês/ano Aceito: dia/mês/ano

Mokhtarzadeh, S., Sadeghzadeh, M. and Bahrani. M. (2013.a). Comparative Study of Text Segmentation algorithms based on Lexical Cohesion in Persian Texts. The 18th Annual Conference of Computer Society of Iran, Sharif University, Tehran, Iran, pp. 269-274.

Mokhtarzadeh, S., Sadeghzadeh, M. and Dianat, R. (2013.b). An Approach for Unsupervised Text Segmentation in Persian Language. Proceedings of National Conference on Computer Science and Engineering, Islamic Azad University, Najafabad Branch, Iran, pp. 912-918.

Morris, J. and Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of The Structure of Text, Journal of Computational Linguistics, 17(1), pp. 21-42.

Reynar, G. C. (1998). Topic Segmentation: Algorithms and Applications, PHD thesis, University of Pennsylvania.

Richmond, K., Smith, A. and Amitay, E. (1997). Detecting Subject Boundaries within Text: A Language Independent Statistical Approach, The Second Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 47-54.

Wu, J., Tseng, J. and Tsai, W. (2010). A Discrete Particle Swarm Optimization Algorithm for Domain Independent Linear Text Segmentation. In Proceeding of IEEE International Conference on Granular Computing, pp. 519-524.

Wu, J., Tseng, J. and Tsai, W. (2014). A Hybrid Linear Text Segmentation Algorithm Using Hierarchical Agglomerative Clustering and Discrete Particle Swarm Optimization. Journal of Integrated Computer-Aided Engineering, vol 21, pp. 35-46.

Recebido: dia/mês/ano Aceito: dia/mês/ano