Ray Optimization Algorithm for Multi Document Summarization

Zuhair Hussein Ali*
zuhair72h@uomustansiriyah.edu.iq

Ahmed Kawther Hussein*
ahmedkawther@uomustansiriyah.edu.iq

Ali Shakir Mahmood*
asmjhm2006@uomustansiriyah.edu.iq

*Computer Science Department, College of Education, Mustansiriyah University, Baghdad, Iraq

Abstract

Multi Document summarization is the process of generating single document from a set of documents. In this paper a new model which based on computing readability and content coverage for each sentence. The Ray optimization algorithm used to choose the most suitable sentences to be included in the final summary. The TAC-2011 dataset used to evaluate the proposed model for English document only. The performance of the proposed model is evaluated using ROUGE metric. The calculated results support the effectiveness of the proposed model.

Keywords: Ray optimization, summarization, Readability, Content coverage.

خوارزمية الاشعة لتلخيص النصوص المتعددة

م.د. زهير حسين علي *   م.د. احمد كوثر حسين *   م.د. علي شاكر محمود *

* قسم علوم الحاسبات / كلية التربية / الجامعة المستنصرية / بغداد / العراق

المستخلص

تلخيص النصوص المتعددة هي تكوين نص واحد من مجموعة من النصوص. في هذا البحث تم استخدام طريقة بالاعتماد على قابلية القراءة وتغطية المحتوى. تم اعتمادخوارزمية الاشعة –TAC

2011لاستخراج أفضل الجمل لكي يتم أضافتها للملخص. تم استخدام قاعدة البياناتللغة
الانكليزية.اختبرت النتائج باستخدام ROUGE. أن النتائج المتحصلة تدعم كفاءة النظام المقترح.

1. Introduction

Due to the rapid growth of information on the World Wide Web a huge number of documents has been produced. Users need to get just the mostpertinent information on an exact subject as fast as possible. Automatic Text summarization (ATS) is the process to extract this significant information. ATS is the process of producing a single document from one document or more than one document with keeping the main ideas of the summarized document(s) [1]. Based on the number of documents to be summarized ATS can be classified as a Single Document Summarization (SDS) or a Multi Document Summarization (MDS). In an SDS only one document can be summarized into shorter ones, whereas in MDS a set of related documents with the same topic is summarized into one shorter summary [2]. ATS techniques can also be categorized as abstractive summarization and extractive summarization. Abstractive summarization required deep natural language processing techniques, whereas extractive summarization does not require [3].Also, summarization can be categorized as query summarization and generic summarization. In the query- based summarization, a summary was generated according to the user query, where the documents searched to match with the user query. While generic summarization creates a summary, which include the main content of the documents. One of the most challenges for the generic summarization is that no topic or query available for the summarization process [4].

In this paper a new model for extracting generic MDS is proposed. The proposed model is based on computing the readability and content coverage for each sentence in the document collection. The Ray Optimization Algorithm used to choose the best sentences to be include to final generate summary.

2. Related Works

In this section some of ATS methods will be investigated.In [5] at 2011 the authors proposed an approach for MDS for Arabic and English documents. The approach based on K-means clustering technique. The main idea of this approach based on treating all documents to summarize as a single bag of sentences, where the sentences are clustered using a single cluster, the sentences then ranked according to the similarity to the centroid of the cluster. The summary is created by choosing sentences in that ranked order. The selection process was repeated until reaching the expected limit. The TAC-2011 dataset was used in this approach. The results were embedded in the TAC-2011 dataset for both Arabic and English documents

In [6] at 2017 the authors proposed a method for extractive MDS based on supervised learning that used to compute the pyramid scores and uses them for MDS. Firstly, the genetic algorithm used for an automatic training data generation the pyramid used as a fitness function. This learning used later to estimate pyramid scores. In [7] at 2017 the author proposed a method based on a Cuckoo Search algorithm (CSA) as MDS. The model starts by calculating an informative score for each sentence in the document set. The calculation based on term frequency and inverse sentence frequency (TF-ISF), then the cosine similarity used to calculate the inter sentence similarity between the sentences. The sentences with less similarity selected to represent each document. All these selected sentences merged to produce a single document. At the last step the CSA parameters initialized to select most suitable sentences to be included in the final summary. In [8] at 2018 the authors proposed a method that based on calculating sentence score using two methods. For first method set of the text feature calculated for each sentence in the document, then a weightassigns for these features, then linearly combining these features scores with the weights to produce final sentence score. For the method the score of each sentence calculating the average score for text features. The two-score method used to rank all sentences, then the collaborative ranking has been adopted to choose the most important sentences.

3. Problem Statement

To produce a good summary for any MDS system three issues must be considered. These issues are

a. Readability: That indicates the selected sentences must relate to each other with a high degree of similarity.

b. Content coverage: summary should include significant sentences that cover the main ideas of the summarized documents

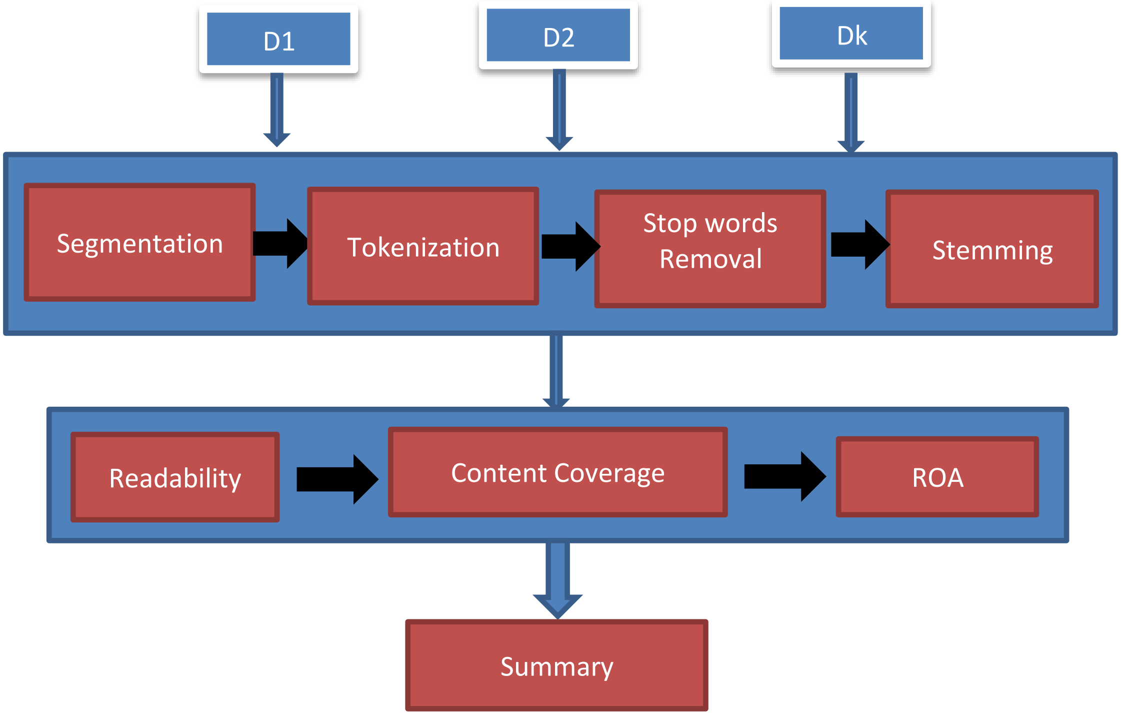
c. Redundancy: The generated summary should include less redundant information to cover most of the relevant topics.

Formally, given a corpus which consists of many clusters, each cluster contains a set of documents called D with the same topic. The set D can be defined as D= {d1, d2, dk} where k is the number of distinct documents in D. Each D can be represented by a set of sentences called Si, i.e D= {Si | 1<=i<=n} where n represents the total number of sentences in the set D. Our goal is to find a subset of set D called S i.e. S⊂ D that satisfies the three important issues of MDS.

4. The Proposed Method

Figure (1) shows the main phases of the proposed method based on the Artificial Bee Colony (ABC) for solving the problem of MDS. The proposed model is intended to satisfy the challenges of the text summarization problem being represented by readability, content coverage and reducing redundancy.

Figure (1): Phases of the proposed model



4.1 Preprocessing

There are four steps for preparing the data, these steps are:

a. Sentence segmentation: which can be performed by splitting sentences based on the dot between them.

b. Tokenization: The main goal of tokenization is separate sentences into words.

c. Stop Words Removal: is the manner of removing words that appear many times in the text and don't offer the required information for recognizing an important sense of the document. There are many strategies utilized for indicating such stop words list. Now, various English stop word list is generally utilized to the TC procedure.

d. Stemming: is the method of generating origin of the word.

4.2 Sentence Representation Phase

After applying the preprocessing phase, D can be represented by distinct terms. Let T'= {t1, t2, …, tp} represents distinct terms exist in D, where p is the number of terms in D. Every sentence (si) in the document D can be represented using the vector space model. As indicated by this model, every sentence (si) is located as a point in the p dimensional vector space, si={wi,1 ,wi,2,….,wi,p}, for i=1…M. Each element in the vector represents a term within a given sentence. The value of each element in the vector assigns a weight using term frequency-inverse-sentence-frequency as explained in Eq. (1).

$$W_{i,k} = TF_{i,k} \times \log\left(\frac{n}{n_k}\right) \qquad (1)$$

Where:

$TF_{i,k}$ is the term of frequency(represent how many term $t_k$occur in the sentence S_i).
n number of sentences in the document collection.
n_k number of sentences in which term t_k appears.
The weight $W_{i,k}$of term $t_k$should be zero if the term t_k does not appear in the sentence [9].

The vector space model requiresthe high dimensionality of the feature space, which affecting the performance of the summarization methods. Depending on the number of words in each sentence the specified vector dimension p is very large. Accordingly, the subsequent vectors would have numerous null elements. It is the major disadvantage of vector space model.

The value of every elementrelies upon the level of relationship between its related terms and the respective sentence. Terms weight can be calculated using term frequency (tf) and inverse-sentence-frequency (isf). Therefore the weight wi,k related to the term tk of sentence si calculated based on (tf_isf) scheme.

At the end, the center of document collection (o) can be calculated as the average of weights wi,k of term tk for all si in the document collection as in Eq. (2).

$$O_k = \frac{1}{n}\sum_{i=1}^{n} w_{i,k} \qquad for\ k = 1\ to\ p \qquad (2)$$

4.3 Similarity Measure

After applying sentence representation phase. Each sentence in the document collection is represented by the term weight si = {w1, w2, ..., wp}. The cosine similarity measure is used to compute the similarity between

To calculate the similarity between the center of document collection O and each sentence si , the calculations will be as follows: for a center vector O={o1,o2,..,op} and a sentence si={w1,w2,…,wp}for M sentences, cosine similarity is used as in Eq. (3)[10].

$$Y = \frac{\sum_{i=1}^{z} O_i S_i}{\sqrt{\sum_{i=1}^{z}(O_i)^2} * \sqrt{\sum_{i=1}^{z}(S_i)^2}} \qquad (3)$$

4.4 Sentences Readability and Content Coverage

The readability indicates how relate the selected sentences to each other's, while content coverage indicates the importance of the selected sentences that should cover the main ideas of the summarized document to be included to the final generated summary. The readability and content coverage can be computed according to Eq.(4)and Eq.(5) respectively.

$$R_s = \sum_{0 \le i < s} w(s_i, s_{i+1}) \qquad (4)$$

$$RF_s = \frac{R_s}{Max_{\forall i} \times R_i} \qquad (5)$$

RFs represents the readability factor of a summary with lengthS [11].

The content coverage of each sentence in the document collection D can be calculated as follows.

$$F_{coverage}(s_i) = sim(s_i, O) \qquad for\ i = 1\ to\ n \qquad (6)$$

Where O is the center vector of document collection.

The similarity between sentence s_i and O is evaluated to calculate the significance of each sentence. Higher similarity values relate to the higher content coverage. The similarity between sentences and center vector should be maximized.

4.5Ray Optimization Algorithm based Text Summarization

ROA is a metaheuristic algorithm proposed by Kaveh and Khayatazad at 2012. Inspired by the Snell's light refraction law. Since the light refracted, and it is direction change when it is moved from the light medium to the darker medium. ROA can be considered as a multi-agent technique. The agents searched the optimal solution when the light is reflected and the direction is changed, this action helps the agents to find the search space in early stages of the optimization operation and to make them converge in the last stages. Each agent in the RO is can be considered as a ray of light that moves in the search space to discover the global or near-global solution [12].

Algorithm: Ray optimization for Text summarization

```
Input: Document collection D
Output: summary of length L.
Step1: Do the preprocessing phase.let summary={}
Step2: Calculate the Readability for each sentence.
Step3: Calculate the Content coverage for each sentence.
Step4: Select number of agents randomly.
Step5: for each agent initialized the velocity and position randomly
Step6: set LBM and GBM /Local best memory and global best memory
Repeat
Step7: for each agent do
$New\_pos_i=current\_pos_i+movement\_vec_i$
$X_{i,j}=X_{j,min}+rand(X_{j,max}-X_{j,min})$
        End for
Step8: Update LBM and GBM.
Step9: For each agent do
$V_{i,j}=-1*2rand$
        End for
Step10: Select the best agent
Step11: Compare all the summary sentences with the selected sentence
Using Eq.(3)
if similarity <threshold add a sentence to the summary
else ignore the sentence/to remove redundant information
until reach max summary length.
```

5. Dataset and Evaluation Metrics

The dataset utilized in this model experimentisTAC-2011 which consists of a document set written in seven languages (English, Arabic, Hebrew French, Hindi, Greek, and Czech). Foreach language (10) topics are used each of (10) documents. Summarization of (10) pre evaluated documents were alsoprovided by the authors ofTAC-2011 [13]. Our proposed method deals with Englishlanguages only.

ROUGEwill beutilized to evaluate the performance ofthe proposed model. ROUGE package produces three numbers representing: Precision, Recall and F−score [14]. They are formulated as follows.

$$\text{Precision} = \frac{\text{system summary sentences } \cap \text{ ideal summary sentences}}{\text{number of sentences in the system summary}} \tag{7}$$

$$\text{Recall} = \frac{\text{system summary sentences } \cap \text{ ideal summary sentences}}{\text{number of sentences in the ideal summary}} \tag{8}$$

$$\text{F} - \text{measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{9}$$

6. Experimental Results

Table (1) and Table (2) shows the results ofthe proposed model in comparison with reference[5].

Table (1): Proposed MDS Model Results using ROUGE-1.

Model	Proposed MDS model1 Results			Ref.[5] Results		
Topic	Precision	Recall	F-Score	Precision	Recall	F-Score
ID1	0.44236	0.63137	0.52022	0.41253	0.40524	0.40776
ID2	0.35258	0.5122	0.41764	0.45655	0.46481	0.46062
ID3	0.36245	0.56316	0.44104	0.47909	0.43169	0.45404
ID4	0.4321	0.65322	0.52013	0.44966	0.44423	0.44691
ID5	0.42121	0.60121	0.49536	0.43513	0.41092	0.42243
ID6	0.37221	0.49545	0.42507	0.45122	0.3547	0.39617
ID7	0.42123	0.59765	0.49416	0.3953	0.39586	0.39547
ID8	0.49743	0.69752	0.58072	0.39265	0.38714	0.38985
ID9	0.46311	0.59592	0.52118	0.37726	0.38105	0.3791
ID10	0.45527	0.67428	0.54354	0.51806	0.52488	0.52141

Table (2): Proposed MDS Model Results using ROUGE-2.

Model / Topic	Proposed MDS model1 Results			Ref [5] Results		
	Precision	Recall	F-Score	Precision	Recall	F-Score
ID1	0.23211	0.35121	0.27950	0.12448	0.12125	0.12247
ID2	0.17122	0.19998	0.18448	0.16779	0.17052	0.16914
ID3	0.20130	0.19833	0.19980	0.19256	0.1733	0.18237
ID4	0.27221	0.41006	0.32720	0.15369	0.1517	0.15269
ID5	0.22334	0.33211	0.26707	0.14404	0.13605	0.13985
ID6	0.14011	0.17013	0.15366	0.1367	0.10655	0.11937
ID7	0.17521	0.25232	0.20681	0.09612	0.09662	0.09635
ID8	0.25200	0.35211	0.29376	0.12298	0.12144	0.12219
ID9	0.21324	0.28118	0.24254	0.10841	0.10962	0.109
ID10	0.25238	0.31837	0.28156	0.2483	0.25177	0.25

Table (1) and Table (2) shows that the results ofthe proposed MDS model1 are better than Ref [5] results for the three metrics. This improvement of the proposed model occurs for two reasons. First, the good performance of the ROA to obtain the best sentences. Second, the good definition of the three important issues of MDS: Readability, content coverage and remove redundancy that increases the performance of the proposed model.

7. Conclusions

MDS is the process of producing a single document from a set of documents with keeping the main idea of the summarized documents. This paper proposed an MDS based on ROA. The proposed model based on three important issues in MDS, which include readability, content coverage and reducing redundancy. The ROA used to select the most important sentences. Good results obtained from the proposed method due to the efficient formulating of the readability and content coverage and efficient use of ROA.

Acknowledgment

The author would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad-Iraq for its support in the present work.

8. References

1. Alguliev, R. Isazade, N. ,Abdi.A. & Idris,n. (2017). "A Model for Text Summarization", International Journal of Intelligent Information Technologies, Vol. 13,No.1.
2. Kumar,R. & Chandrakal,D. (2016)" A survey on text summarization using optimization algorithm," ELK Asia Pacific Journals, Vol. 2, No. 1.
3. Meena,Y. & Gopalani,D. (2015). "Evolutionary Algorithms for Extractive Automatic Text Summarization," Procedia Comput. Sci., Vol. 48, No.2, PP. 244–249.
4. Alguliev, R. , Aliguliyev, R. & Isazade, N. R. (2012)." DESAMC + DocSum : Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization". Knowledge-Based Systems, Vol.36, PP.21–38.
5. E l-Haj, M. (2012) "Multi-Document Arabic Text Summarization". Albert Sloman Library, Essex University. Thesis (PhD.), School of Computer Science and Electronic Engineering - University of Essex.
6. Peyrard, M. &Eckle-Kohler, J. (2017). Supervised Learning of Automatic Pyramid for Optimization-Based Multi-Document Summarization. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1084–1094.
7. Rautray,R., &Balabantaray,R.C.(2017). "An evolutionary framework for multi document summarization using Cuckoo search approach:MDSCSA". Applied Computing and Informatics, Vol.14 No.2 PP. 134-144.
8. Verma,p.,&Om,H.(2018)."CooaborativeEanking-based Text summarization using a Metaheuristic Approach. Emerging Technologies in Data mining and information security,PP.417-426. Doi:10.1007/978-981-13-1501-5_36.
9. Salton, G., & Buckley, C. (1988),"TERM-WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL", Information Processing & Management Vol. 24, No. 5, PP. 513-523.
10. Lin,Y. , Jiang.J & Lee,S.(2014)." A Similarity Measure for Text Classification andClustering". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 26, No. 7,PP.1575-1590.
11. Asgari, H., Masoumi, B., & Sheijani, O. S. (2014). Automatic text summarization based on multi-agent particle swarm optimization. 2014 Iranian Conference on Intelligent Systems (ICIS).
12. Kaveh, A. & Khayatazad, M. (2012). A new meta-heuristic method: ray optimization. Computers and Structures, Vol. 112-113, pp. 283–294.
13. Giannakopoulos,G. El-Haj M. Favre,B. Litvak, M. Steinberger, Josef. &Varma,V. (2011). "TAC11 MultiLing Pilot Overview". TAC2011 Workshop. Presented at the TAC 2011 Gaithersburg,MD, U.S.A.
14. Chin-Yew,L. (2004). "Rouge: A package for automatic evaluation of summaries." In Text summarization branches out: Proceedings of the ACL-04 workshop, Vol. 8.