Using Cellular Automata to Construct Sentence Ranking

Pouya Khosravian Dehkordi and Farshad Kiyoumarsi

Abstract—This work proposes an approach to address the problem of improving content selection in automatic text summarization by using some statistical tools. This approach is a trainable summarizer, which takes into account several features, for each sentence to generate summaries. First, we investigate the effect of each sentence feature on the summarization task. Then we use all features in combination to train cellular automata (CA), genetic programming approach and fuzzy approach in order to construct a text summarizer for each model. Furthermore, we use trained models to test summarization performance. The proposed approach performance is measured at several compression rates on a data corpus composed of 17 English scientific articles. This article shows that some features are more important to construct models rather than other.

***Index Terms*—Fuzzy, genetic programming, cellular automata, machine learning.**

I. Introduction

Automatic text summarization has been an active research area for many years. Evaluation of summarization is a quite hard problem. Often, a lot of manual labor is required, for instance by having humans read generated summaries and grading the quality of the summaries with regards to different aspects such as information content and text clarity. Manual labor is time consuming and expensive. Summarization is also subjective. The conception of what constitutes a good summary varies a lot between individuals, and of course also depending on the purpose of the summary.

Recently many experiments have been conducted for the text summarization task. Some were about evaluation of summarization using relevance prediction [1], and voted regression model [2]. Others were about single- and multiple-sentence compression using ‘‘parse and trim” approach and a statistical noisy-channel approach [3] and conditional random fields [4]. Other research includes multi-document summarization [5] and summarization for specific domains [6].

We employ an evolutionary algorithm, Cellular Automata (CA) [7], as the learning mechanism in our Adaptive Text Summarization (ATS) system to learn sentence ranking functions. Even though our system generates extractive summaries, the sentence ranking function in use differentiates ours from that of [8] who specified it to be a linear function of sentence features. We used CA to generate a sentence ranking function from the training data and applied it to the test data, which also differs from [9] who used decision tree, [10] who used bayes’s rule and [4] who implemented both na ïve bayes and decision tree.

In this work, sentences of each document are modeled as genetic programming of features extracted from the text. The summarization task can be seen as a two-class classification problem, where a sentence is labeled as ‘‘correct” if it belongs to the extractive reference summary, or as ‘‘incorrect” otherwise. We may give the ‘‘correct” class a value ‘1’ and the ‘‘incorrect” class a value ‘0’. In testing mode, each sentence is given a value between ‘0’ and ‘1’ (values between 0 and 1 are continuous). Therefore, we can extract the appropriate number of sentences according to the compression rate. The trainable summarizer is expected to “learn” the patterns which lead to the summaries, by identifying relevant feature values which are most correlated with the classes ‘‘correct” or ‘‘incorrect”. When a new document is given to the system, the ‘‘learned” patterns are used to classify each sentence of that document into either a ‘‘correct” or ‘‘incorrect” sentence by giving it a certain score value between ‘0’ and ‘1’. A set of highest score sentences are chronologically specified as a document summary based on the compression rate.

II. Background

A. Text Features

We concentrate our presentation in two main points: (1) the set of employed features; and (2) the framework defined for the trainable summarizer, including the employed classifiers.

A large variety of features can be found in the text-summarization literature. In our proposal we employ the following set of features:

(F1) sentence length [11]:

This feature is employed to penalize sentences that are too short, since these sentences are not expected to belong to the summary. We use the normalized length of the sentence, which is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

(F2) Sentence position [12]:

This feature can involve several items, such as the position of a sentence in the document as a whole, it’s the position in a section, in a paragraph, etc., and has presented good results in several research projects .

We use here the percentile of the sentence position in the document. The final value is normalized to take on values between 0 and 1.

(F3) Similarity to title [11]:

According to the vectorial model, this feature is obtained by using the title of the document as a “query” against all the

Manuscript received October 28, 2011; revised November 25, 2011.

Authors are with Islamic Azad University-Shahrekord Branch, Shahrekord from Iran (email: Khosravyan@iaushk.ac.ir; Kumarci-farshad@iaushk.ac.ir).

sentences of the document; then the similarity of the document's title and each sentence is computed by the cosine similarity measure.

(F4) Similarity to keywords [12]:

This feature is obtained analogously to the previous one, considering the similarity between the set of keywords of the document and each sentence which compose the document, according to the cosine similarity. For the next two features we employ the concept of text cohesion. Its basic principle is that sentences with higher degree of cohesion are more relevant and should be selected to be included in the summary. This feature must be introduced by expert person of that language.

(F5) Occurrence of proper nouns [13]:

The motivation for this feature is that the occurrence of proper names, referring to people and places, are clues that a sentence is relevant for the summary. This is considered here as a binary feature, indicating whether a sentence s contains (value "true") at least one proper name or not (value "false"). Proper names were detected by a part-of-speech tagger.

(F6) Indicator of main concepts [14]:

This is a binary feature, indicating whether or not a sentence captures the main concepts of the document. These main concepts are obtained by assuming that most of relevant words are nouns. Hence, for each sentence, we identify its nouns using part-of-speech software. For each noun we then compute the number of sentences in which it occurs. The fifteen nouns with largest occurrence are selected as being the main concepts of the text. Finally, for each sentence the value of this feature is considered "true" if the sentence contains at least one of those nouns, and "false" otherwise.

(F7) Occurrence of non-essential information [15]:

We consider that some words are indicators of non-essential information. These words are speech markers such as "because", "furthermore", and "additionally", and typically occur in the beginning of a sentence. This is also a binary feature, taking on the value "true" if the sentence contains at least one of these discourse markers, and "false" otherwise.

(F8) Sentence-to-centroid cohesion [13]:

This feature is obtained as follows: for each sentence s we first compute the similarity between s and each other sentence s of the document; then we add up those similarity values, obtaining the raw value of this feature for s; the process is repeated for all sentences. The normalized value (in the range 0 and 1) of this feature for a sentence s is obtained by computing the ratio of the raw feature value for s over the largest raw feature value among all sentences in the document. Values closer to 1.0 indicate sentences with larger cohesion.

B. Text Summarization Based on Genetic Programming

In order to implement text summarization based on Genetic Programming [16], we used GP since it is possible to simulate genetic programming in this software. To do so; first, we consider each characteristic of a text such as sentence length, location in paragraph, similarity to key word and etc., which was mentioned in the previous part, as the genes of GP. Then, we enter all the operators needed for summarization, in the knowledge base of this system (All those operators are formulated by several expends in this field). Afterward, a value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available operators in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary. To do these steps, we summarize the same text using genetic programming.

C. Text Summarization Based on Fuzzy Logic Approach

In order to implement text summarization based on fuzzy logic [17], we used MATLAB since it is possible to simulate fuzzy logic in this software. To do so; first, we consider each characteristic of a text such as sentence length, location in paragraph, similarity to key word and etc., which was mentioned in the previous part, as the input of fuzzy system. Then, we enter all the rules needed for summarization, in the knowledge base of this system (All those rules are formulated by several expends in this field).

Afterward, a value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary. To do these steps, we summarize the same text using fuzzy logic.

D. Cellular Automata

At the beginning of 1950, cellular automata (CA) have been proposed by Von Neumann. He was interested to male relation between new computational device - automata theory -and biology. His mind was preoccupied with generating property in natural events [18].

He proved that CA can be general. According to his findings, CA is a collection of cells with reversible states and ability of computation for everything. Although Van rules were complicated and didn't strictly satisfy computer program, but he continues his research in two parts: for decentralizing machine which is designed for simulation of desirable function and designing of a machine which is made by simulation of complicated function by CA [19].

Wolfram has conducted some research on problem modeling by the simplest and most practicable method of CA architecture too. In 1970,"The Game of Life" introduced by Conway and became very widely known soon. At the beginning of 1980, Wolfram studied one-dimension CA rules and demonstrated that these simple CAs can be used in modeling of complicated behaviors [20], [21].

CA is characterized by (a) cellular space (b) transfer rule [7]. For CA , cell, the state of cell in time t, sum of neighbors state at time t and neighborhood radius are denoted by i, , , and r, respectively. Also, the rule is function of CA and it is characterized by 1-cellular space 2-transfer rule [7].

For CA , cell, the state of cell in time t, sum of neighbors state at time t and neighborhood radius are denoted by i, S_i^t, η_i^t, and r, respectively. Also, the rule is function of $\varphi(\eta_i^t)$.

Each cell changes its state, spontaneously. The primary quality of cells depends on primary situation of problem. By these primary situations, CA is a system which has certain behavior by local rules. The cells which are not neighbors

have no effect on each other. CA has no memory, so present state defines the next state [22].

Quad rule CA is as CA= (Q, d, V and F), where Q, d, V and F are collection of possible state, CA dimension, CA neighborhood structure and local transferring rule, respectively.

For 1-d CA, amount of i cell ($1 \le i \le n$) at t is shown by $a_i(t)$ and is calculated by this formula:

$$a_i(t+1) = \varphi[a_{i-1}(t), a_i(t), a_{i+1}(t)]$$

In this formula, if φ is affected by the neighbors, it is general.

If φ is a function of neighbor's cell collection and central cell, it is totalistic:

$$a_i(t+1) = \varphi[a_{i-1}(t) + a_i(t) + a_{i+1}(t)]$$

III. The Proposed Automatic Summarization Model

Figure 1 shows the proposed automatic summarization model. We have two modes of operations:

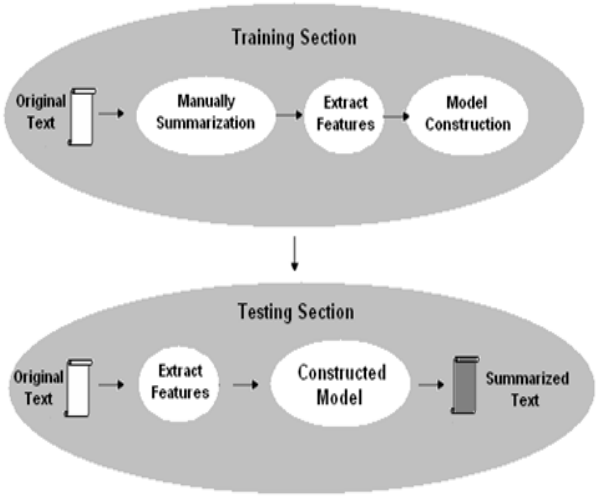

Fig. 1. The proposed automatic summarization model

1. Training mode where features are extracted from 16 manually summarized English documents and used to train Cellular Automata, Fuzzy and Genetic programming models.

2. Testing mode, in which features are calculated for sentences from one English document. (These documents are different from those that were used for training.) The sentences are ranked according to the sets of feature weights calculated during the training stage. Summaries consist of the highest-ranking sentences.

A. Cellular Automata Model

The basic purpose of Cellular Automata (CA) is optimization. Since optimization problems arise frequently, this makes CA quite useful for a great variety of tasks. As in all optimization problems, we are faced with the problem of maximizing/minimizing an objective function f(x) over a given space X of arbitrary dimension [20]. Therefore, CA can be used to specify the weight of each text feature.

The Cellular Automata (CA) is exploited to obtain an appropriate set of feature weights using the 17 manually summarized English documents.

Thousand states for each iteration were produced. Evaluate fitness of each state (we define fitness as the average precision obtained with the state when the summarization process is applied on the training corpus), and retain the fittest 8 state to mate for new ones in the next iteration. In this experiment, thousand iterations are evaluated to obtain steady combinations of feature weights. A suitable combination of feature weights is found by applying CA. All document sentences are ranked in a descending order according to their scores. A set of highest score sentences are chronologically specified as a document summary based on the compression rate.

IV. Experimental Result

A. Text Data

Seventeen English articles in the domain of science were collected from the Reading Book. Seventeen English articles were manually summarized using compression rate of 30%. These manually summarized articles were used to train the previously mentioned three models. The other one English article was used for testing. The average number of sentences per English articles is 85.8, respectively.

One of the sample articles as shown in the follow:

The Race of Man

If you stood in a busy place in big cosmopolitan city, like Times Square in New York City or Piccadilly Circus in London, and watched people go by, you would soon realize how the people of the modern world intermixed are Anthropologists speak of three major races of man.

These races are identified as there distinct group of people. Each group has certain physical characteristics that are inherited. These three groups belong to one human family, and all may have been the same originally.

However, as they moved to different parts of the earth, they developed different features adapted to the conditions of climate and food in the places where they lived for a long period of time. In more modern times, these groups of people have been intermixing. Some groups have been conquered; other groups have intermarried.

Nevertheless, if you watched the passers-by in Times Square carefully, you would probably recognize several major types of people. A man with yellowish skin, straight black hair, high cheekbones, and almond-shaped eyes probably belongs to the people of eastern Asia called the mongoloid race. American Indians, who live in America and have reddish-yellowish skin, also belong to this group,

If a man is from Africa south of the Sahara desert, he is likely to have a long head with black or dark brown skin, a broad, flat nose; thick, protruding lips; and tightly curled hair. He belongs to the negroid race. Other men like him can be found in the south pacific islands. a third group of men had their original home on the content of Europe. This group is known as the Caucasoid race, because the earliest skull of this type of human being was found in the Caucasus

Mountains region in southeastern Europe. The Caucasians are called the white race because their skin color is generally lighter than the yellow, brown, or black skin tones of the other races. People of the white race have a variety of head shapes. And their hair varies from silky straight to curly. People of the race living near the Mediterranean Sea are usually darker-haired and darker-eyes than people in areas farther north. Nowadays people of the Caucasoid people live ia all parts of the world.

Some scientists speak of a fourth group, the Australia, who are dark-skinned aborigines living on the content of Australia. There are also some people, like the ainu of Japan, who do not seem to belong to any one of the major races.

If you wanted to make a map showing the races of mankind, it could not have only three or four colors for the main racial group; it would have to show many tints and shades. As men mingle more and more in the modern world of easy travel, the races become more intermixed. It is sometimes difficult to label a man as belonging to one race or another. Is it even desirable to emphasize the differences between races? Mans great problem is to learn how to live peacefully with people different from himself. As members of one family, men must "live like brothers or die like beasts."

If you stood in a busy place in big cosmopolitan city, like Times Square in New York City or Piccadilly Circus in London, and watched people go by, you would soon realize how the people of the modern world intermixed are. If you wanted to make a map showing the races of mankind, it could not have only three or four colors for the main racial group; it would have to show many tints and shades. If a man is from Africa south of the Sahara desert, he is likely to have a long head with black or dark brown skin, a broad, flat nose; thick, protruding lips; and tightly curled hair.

Nevertheless, if you watched the passers-by in Times Square carefully, you would probably recognize several major types of people. Some scientists speak of a fourth group, the Australia, who are dark-skinned aborigines living on the content of Australia. This group is known as the caucasoid race, because the earliest skull of this type of human being was found in the Caucasus Mountains region in southeastern Europe.

B. Cellular Automata Configuration

We are going to exploit the CA approach of [7], for summarization and use it as a baseline approach. For a sentence s, a weighted score function, is exploited to integrate the eight feature scores mentioned in previous

Related parameters for the training and testing of the CA model like States, Rules, Neighbor and other are given in Table 1 and 2.

TABLE I: CA DATA

Independent Variables:	8
Training Samples:	1016
Testing Samples:	105

TABLE II: CA GENERAL SETTINGS

States	2,3
Rules	256, 7625597484987
Neighbor	Von Neumann

The selected neighbor was showed in figure 2:

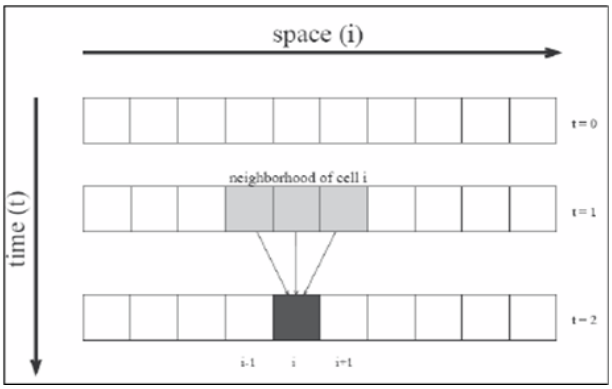

Fig. 2. Neighborhood space of von neumann in 1-D CA

C. The Result of Cellular Automata Model

We have exploited the CA approach of [7], for summarization as described above. Therefore, we have exploited the eight features for summarization. The system calculates the feature weights using Cellular Automata.

All document sentences are ranked in a descending order according to their scores. A set of highest score sentences are chronologically specified as a document summary based on the compression rate. To do CA concepts we using CA Classification model [20].

D. CA Model Explicit Formulation

By using CA rules and analyzing data we got set of rules are given in figure 3:

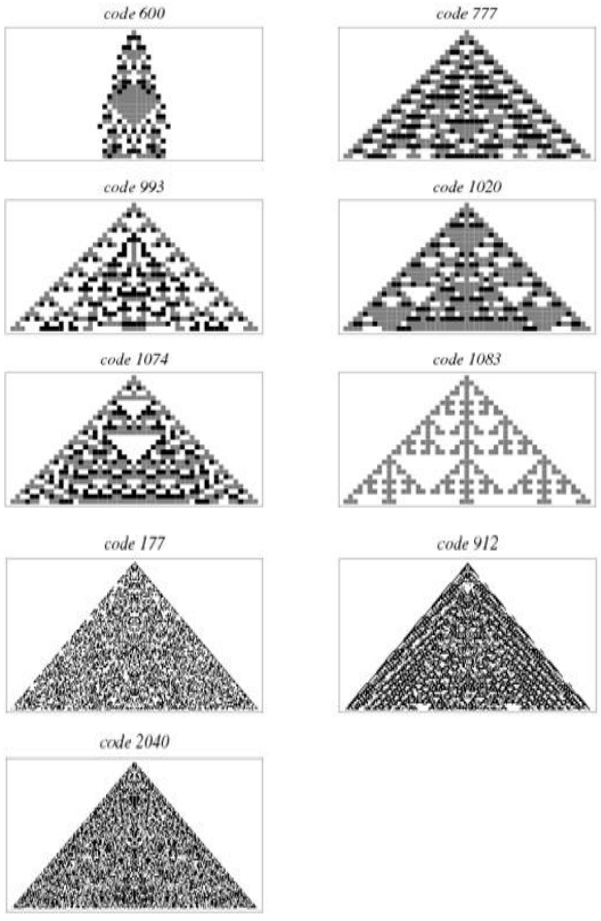

Fig. 3. Specify rules was produced by CA concepts for automatic text summarization

TABLE III: ALL MODELS PERFORMANCE EVALUATION BASED ON PRECISION

Compression Rate (CR)	10%	20%	30%
	Precision (P)	Precision (P)	Precision (P)
CA Model	28.18%	29.04%	31.88%
Fuzzy Model	36.36%	42.86%	46.88%
Genetic Programming Model	54.54%	57.14%	59.38%

TABLE IV: STATISTICS – TRAINING

Best Fitness:	679.43
Max. Fitness:	1000
Accuracy:	68.04%

TABLE V: STATISTICS – TESTING

Best Fitness:	425.96
Max. Fitness:	1000
Accuracy:	43.81%

E. Evaluation CA Model

We used 16 English text documents for training and one for testing CA model and the results are given in table 4 and 5:

	Precision (10%)	Recall (10%)	Precision (20%)	Recall (20%)	Precision (30%)	Recall (30%)
F1	0.2809	0.2911	0.2841	0.2953	0.2879	0.326
F2	0.2734	0.3113	0.2798	0.3177	0.2834	0.3213
F3	0.312	0.3564	0.3192	0.3636	0.3206	0.365
F4	0.3315	0.3732	0.3328	0.3745	0.3417	0.3834
F5	0.2609	0.2988	0.2658	0.3037	0.2764	0.3055
F6	0.3413	0.3834	0.3513	0.3934	0.3663	0.4084
F7	0.3005	0.3262	0.3114	0.3371	0.3127	0.3384
F8	0.3775	0.4209	0.3882	0.425	0.3923	0.4271

Fig. 4. The summarization precision and recall associated with each feature for different compression rates

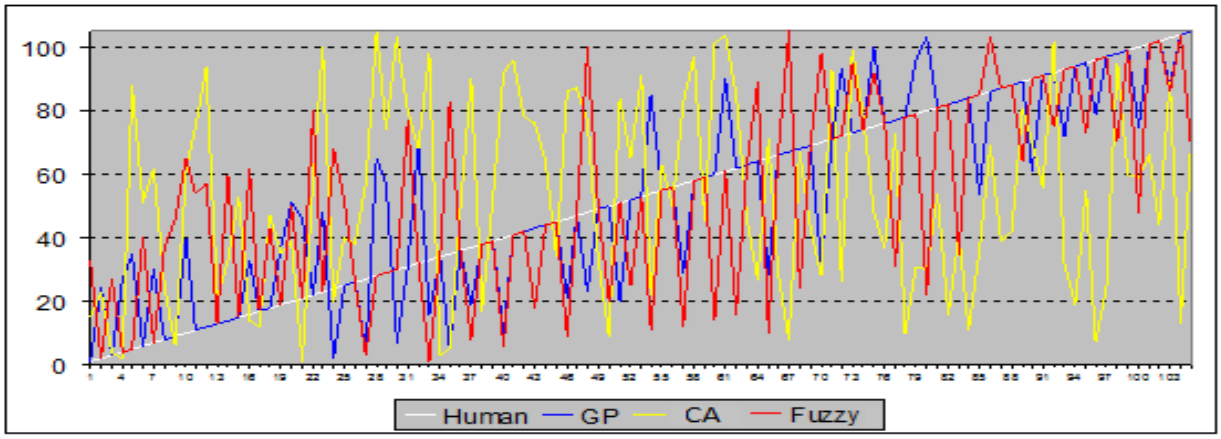

Fig. 5. Compare sentence priority of CA, fuzzy and genetic programming model with human priority

V. DISCUSSION

It is clear from Table 3 that this approach cannot be extended to the genre of newswire text and also Fig.4 shows that the most important text feature for summarization is F8 (sentence-to-sentence cohesion) since it gives the best results. It is reasonable, since the sentence that has a maximum number of branches should convey the most important part in the article. F6 (indicator of main concepts) also gives good results since it conveys the vocabulary overlap between this sentence and other sentences in the document. Usually, the document title conveys the main topic of this document. Therefore, F4 (similarity to title) which is the vocabulary overlap between this sentence and the document title gives good results. The lowest results are associated with F11 (occurrence of non-essential information) since most of reading texts do not contain many anaphors data. Therefore, the system ranks a sentence that does not contain anaphors data according to its position.

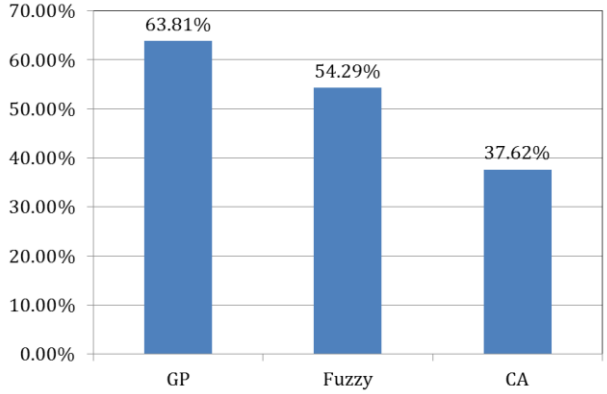

Fig. 6. The accuracy for all models

Fig.5 and Fig.6 shows the total system performance in terms of precision for in case of all models for English articles, respectively. It is clear from the figure that CA approach gives the lowest results since CA has a bad capability to model arbitrary densities. The Fuzzy model and GP has better precision than the CA model.

VI. Conclusion and Function Work

In this paper, we have investigated the use of cellular automata (CA), genetic programming approach and fuzzy approach for automatic text summarization task. We have applied our new approaches on a sample of 17 English scientific articles. Our approach results outperform the baseline approach results. Our approaches have been used the feature extraction criteria which gives researchers opportunity to use many varieties of these features based on the text type.

In the future work, we will extend this approach to multi-document summarization by addressing some anti-redundancy methods which are needed, since the degree of redundancy is significantly higher in a group of topically related articles than in an individual article as each article tends to describe the main point as well as necessary shared background.

References

[1] S. Hobson, B. Dorr, C. Monz, and R. Schwartz, "Task-based evaluation of text summarization using relevance prediction," *Information Processing and Management 43 (6)*, pp. 1482–1499, 2007.

[2] T. Hirao, M. Okumura, N. Yasuda, and H. Isozaki, "Supervised automatic evaluation for summarization with voted regression model," *Information Processing and Management 43 (6)*, pp. 1521–1535, 2007.

[3] D. Zajic, B. Dorr, J. Lin, and R. Schwartz, "Multi-candidate reduction: sentence compression as a tool for document summarization tasks," *Information Processing and Management 43 (6)*, pp. 1549–1570, 2007.

[4] J. L. Neto, A. A. Freitas, and C. A. A. Kaestner, "Automatic text summarization using a machine learning approach," *In Proc. 16th Brazilian Symp on Artificial Intelligence (SBIA-2002)* , springer-verlag, 2002, pp. 205-215.

[5] S. Harabagiu, A. Hickl, and F. Lacatusu, "Satisfying information needs with multi-document summaries," *Information Processing and Management 43 (6)*, pp. 1619–1642, 2007.

[6] M. Moens, "Summarizing court decisions," *Information Processing and Management 43 (6)*, pp. 1748–1764, 2007.

[7] A. Moore, *New Constructions in Cellular automata*, oxford university Press, 2003, pp. 21-32.

[8] S. Sekine and C. Nobata, "Sentence extraction with information extraction technique," *In Proc. Of ACM SIGIR'01 Workshop on Text Summarization*, New Orleans, pp. 1115-1129, 2001.

[9] C. Lin, "Training a selection function for extraction," *In the 8th International Conference on Information and Knowledge Management (CIKM 99)*, Kansa City, Missouri, pp. 112-129, 1999.

[10] C. Aone, J. Gorlinsky, B. Larsen, and M. E. Okurowski, "A trainable summarizer with knowledge acquried from robust nlp techniques," *Advances in Automatic Text Summarization*, The MIT Press, Cambridge, Massachusetts, 1999, pp. 71-80.

[11] L. Ferrier, "A maximum entropy approach to text summarization," *School of Artificial Intelligence*, Division of Informatics, University of Edinburgh, pp. 20-34, 2001.

[12] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development 2 (2)*, pp. 159–165, 1958.

[13] T. Nomoto, "Discriminative sentence compression with conditional random fields," *Information Processing and Management 43 (6)*, pp. 1571–1587, 2007.

[14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management 24*, Reprinted in: Sparck-Jones, pp. 513-523, 1988.

[15] S. J. Yeh, T. H. Ke, M. W. Yang, and L. I. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," *Information Processing and Management 41 (1)*, pp. 75–95, 2005.

[16] P. K. Dehkordi, H. Khosravi, and F. Kumarci, "Text summarization based on genetic programming," *International Journal of Computing and ICT Research (IJCIR)*, Vol.3, No.1, pp. 57-64, June 2009.

[17] F. Kyoomarsi, H. Khosravi, E. Eslami, and M. Davoudi, "Extraction-based text summarization using fuzzy analysis," *Iranian Journal of Fuzzy Systems*, Vol. 7, No. 3, pp. 15-32, 2010.

[18] V. Neumann, *Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components*, von neumann's collected works, A. Taub (Ed), 1993, pp. 102-112.

[19] V. Neumann, *The Theory of Self-Reproducing Automata*, A. W. Burks (ed), Univ. of illinois press, urbana and london, 1996, pp. 57-68.

[20] E. Wolfram, *A New Kind of Science*, wolfram media, inc., 2002, pp. 110-139.

[21] E. Wolfram, *Statistical Mechanics of Cellular Automata*, rev. mod. Phys., 1983, pp. 31-44.

[22] E. Wolfram, *Universality and Complexity in Cellular Automata*, physical D, 1984, pp. 78-88.

Pouya Khosravian Dehkordi was born in Isfahan, Iran. He is thirty years old. He got his MSc in Software Engineering form Islamic Azad University of Arak, Iran. His major field of study is Neutral language processing (NLP), Text Summarization, and Artificial Intelligent. He is a "Faculty Member" in Shahrekord Azad University (Iran). His contact information is: Email: khosravyan@iaushk.ac.ir, Tel: +983813361000

Farshad Kiyoumarsi was born in Shahrekord, Iran. in 1972. he is a Member of IACSIT . he got his PhD in Computer Science form Shahid Bahonar University of Kerman, Iran. His major field of study is Neutral language processing (NLP) , Text Summarization ,and Artificial Intelligent. He is a "Faculty Member" in Shahrekord Azad University (Iran). His contact information is: Email: kumarci-farshad@iaushk.ac.ir, tel: +98 381 3361000