Optimizing Machine Learning Approach Based on Fuzzy Logic in Text Summarization

Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami, and Pooya Khosravyan Dehkordy

[1,4] *Islamic Azad University(Shahrekord branch)*

[2]*International Center for Science & High Technology & Environmental Sciences , University of Shahid Bahonar Kerman*

[3]*Shahid Bahonar University of Kerman, The center of Excellence for Fuzzy system and applications*

[1]*Kumarci_farshad@graduate.uk.ac.ir,* [2]*Hkhosravi@mail.uk.ac.ir,*
[3]*Eeslami@mail.uk.ac.ir,* [4]*Pooya_khd@iaun.ac.ir*

Abstract

With the proliferation of the Internet and the huge amount of data it transfers, automatic text summarization is becoming more important. In this paper we first analyze some state of the art methods to text summarization., We also try to analyze one of the previous text summarization methods, "Machine learning Approach", and eliminate its shortcomings .Finally we present an approach to the design of an automatic text summarizer that generates a summary using fuzzy logic to obtain better results compared to previous methods.

1. Introduction

Nowadays, people need much more information in work and life, especially the use of internet makes information more easily gained. Extensive use of internet is one of the reasons why automatic text summarization draws substantial interest. It provides a solution to the information overload problem people face in this digital era.

Text Summarization is not a new idea. Research on automatic text summarization has a very long history, which can date back at least 40 years ago, from the first system built at IBM. Several researchers continued investigating various approaches to this problem through the seventies and eighties. Many innovative approaches began to be explored such as statistical and information-centric approaches, linguistic approaches and the combination of them. In next section, we will deal with the main approaches, which have been used or proposed.

Summary construction is, in general, a complex task which ideally would involve deep natural language processing capacities. In order to simplify the problem, current research is focused on extractive-summary generation [3]. An extractive summary is simply a subset of the sentences of the original text. These summaries do not guarantee a good narrative coherence, but they can conveniently represent an approximate content of the text for relevance judgment.

A summary can be employed in an indicative way – as a pointer to some parts of the original document, or in an informative way – to cover all relevant information of the text [5,7]. In both cases the most important advantage of using a summary is its reduced reading

time. Summary generation by an automatic procedure has also other advantages: (i) the size of the summary can be controlled; (ii) its content is determinist; and (iii) the link between a text element in the summary and its position in the original text can be easily established[10,13].

2. Summarization Approaches

The main steps of text summarization are identifying the essential content, "understanding" it clearly and generating a short text. Understanding the major emphasis of a text is a very hard problem of NLP[15]. This process involves many techniques including semantic analysis, discourse processing and inferential interpretation and so on. Most of the research in automatic summarization has Been focused on extraction. But as in [5,8] the author described, when humans produce summaries of documents, they do not simply extract sentences and concatenate them, rather they create new sentences that are grammatical, that cohere with one another, and that capture the most salient pieces of information in the original document. So, the most pressing need is to develop some new techniques that do more than surface sentence extraction, without depending tightly on the source type. These need intermediated techniques including passage extraction and linking; deep phrase selection and ordering; entity identification and relating, rhetorical structure building and so on. Here we discuss some main approaches which have been used and proposed.

Summarization methods can be roughly grouped into three categories : Statistical approach, general linguistic approach, and hybrid methods that use a combination of these two approaches.

2.1. Statistical Approach

The statistical approach summarizes without understanding; it relies on the statistical distribution of certain features [9]. Recent representative works include the classification-based method, the IR-based method [4] and the position-based method among others. For the classification-based method, it is to classify sentences as either summary-worthy or not, depending on the training data. A classifier can be used to perform this task. A training procedure that is an algorithm is used to calculate the parameters of the classifier. For the IR based methods, usually we make full use of techniques form Information Retrieval. But normally it is hard to give a satisfactory result, because the output summary is not very coherent. For the position-based method, we should notice that, important sentences are located at positions that are genre-dependent. Briefly speaking the statistical approach is domain-independent and fast, but has an inherent upper bound of performance.

2.2. Linguistic Approach

Summarization based on these method needs linguistic knowledge so that the computer can analyze the sentences semantically and then decides which sentences to choose considering the position of the verb, subject, noun and etc. these methods are more difficult than statistical methods.

3. A Review of Text Summarization Based on Machine Learning

An automatic summarization process can be divided into three steps: (1)In the preprocessing step a structured representation of the original text is obtained; (2) In the

processing step an algorithm must transform the text structure into a summary structure; and (3) In the generation step the final summary is obtained from the summary structure.

The methods of summarization can be classified, in terms of the level in the linguistic space, in two broad groups: (a) shallow approaches, which are restricted to the syntactic level of representation and try to extract salient parts of the text in a convenient way; and (b) deeper approaches, which assume a semantics level of representation of the original text and involve linguistic processing at some level.

In the first approach the aim of the preprocessing step is to reduce the dimensionality of the representation space, and it normally includes: (i) stop-word elimination – common words with no semantics and which do not aggregate relevant information to the task (e.g., "the", "a") are eliminated; (ii) case folding: consists of converting all the characters to the same kind of letter case - either upper case or lower case; (iii) stemming: syntactically-similar words, such as plurals, verbal variations, etc. are considered similar; the purpose of this procedure is to obtain the stem or radix of each word, which emphasize its semantics.

A frequently employed text model is the Machine Learning model. After the preprocessing step each text element – a sentence in the case of text summarization – is considered as a N-dimensional vector. So it is possible to use some metric in this space to measure similarity between text elements. The most employed metric is the cosine measure, defined as $\cos q = (<x.y>) / (|x| \cdot |y|)$ for vectors x and y, where $(<,>)$ indicates the scalar product, and $|x|$ indicates the module of x. Therefore maximum similarity corresponds to $\cos q = 1$, whereas $\cos q = 0$ indicates total discrepancy between the text elements.

4. The Used Attribute in Text Summarization

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:

(a) Mean-TF-ISF. Since the seminal work of Luhn, text processing tasks frequently use features based on IR measures . In the context of IR, some very important measures are term frequency (TF) and term frequency inverse document frequency (TF-IDF). In text summarization we can employ the same idea: in this case we have a single document d, and we have to select a set of relevant sentences to be included in the extractive summary out of all sentences in d. Hence, the notion of a collection of documents in IR can be replaced by the notion of a single document in text summarization. Analogously the notion of document – an element of a collection of documents – in IR, corresponds to the notion of sentence – an element of a document – in summarization. This new measure will be called term frequency ´ inverse sentence frequency, and denoted TF-ISF(w,s). The final used feature is calculated as the mean value of the TF-ISF measure for all the words of each sentence [6].

(b) Sentence Length. This feature is employed to penalize sentences that are too short, since these sentences are not expected to belong to the summary [11,14]. 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.

(c) Sentence Position. This feature can involve several items, such as the position of a sentence in the document as a whole, its 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, as proposed by Nevill-Manning; the final value is normalized to take on values between 0 and 1.

(d) Similarity to Title. According to the vectorial model, this feature is obtained by using the title of the document as a "query" against all the sentences of the document; then the similarity of the document's title and each sentence is computed by the cosine similarity measure [2].

(e) Similarity to Keywords. 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 .

(f) Sentence-to-Sentence Cohesion. 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, 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.

(g) Sentence-to-Centroid Cohesion. This feature is obtained for a sentence s as follows: first, we compute the vector representing the centroid of the document, which is the arithmetic average over the corresponding coordinate values of all the sentences of the document; then we compute the similarity between the centroid and each sentence, obtaining the raw value of this feature for each sentence. The normalized value in the range [0, 1] for s is obtained by computing the ratio of the raw feature value over the largest raw feature value among all sentences in the document. Sentences with feature values closer to 1.0 have a larger degree of cohesion with respect to the centroid of the document, and so are supposed to better represent the basic ideas of the document.

For the next features an approximate argumentative structure of the text is employed. It is a consensus that the generation and analysis of the complete rethorical structure of a text would be impossible at the current state of the art in text processing. In spite of this, some methods based on a surface structure of the text have been used to obtain good-quality summaries. To obtain this approximate structure we first apply to the text an agglomerative clustering algorithm. The basic idea of this procedure is that similar sentences must be grouped together, in a bottom-up fashion, based on their lexical similarity. As result a hierarchical tree is produced, whose root represents the entire document. This tree is binary, since at each step two clusters are grouped. Five features are extracted from this tree, as follows:

(h) Referring position in a given level of the tree (positions 1, 2, 3, and 4). We first identify the path form the root of the tree to the node containing s, for the first four depth levels. For each depth level, a feature is assigned, according to the direction to be taken in order to follow the path from the root to s; since the argumentative tree is binary, the possible values for each position are: left, right and none, the latter indicates that s is in a tree node having a depth lower than four.

(i) Indicator of main concepts. 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 a 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.

(k) Occurrence of proper names. 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.

(l) Occurrence of anaphors. We consider that anaphors indicate the presence of non-essential information in a text: if a sentence contains an anaphor, its information content is covered by the related sentence. The detection of anaphors was performed in a way similar to the one proposed by Strzalkowski: we determine whether or not certain words, which characterize an anaphor, occur in the first six words of a sentence. This is also a binary feature, taking on the value "true" if the sentence contains at least one anaphor, and "false" otherwise.

(m) Occurrence of non-essential information. 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.

The ML-based trainable summarization framework consists of the following steps:

1. We apply some standard preprocessing information retrieval methods [12] to each document, namely stop-word removal, case folding and stemming. We have employed the stemming algorithm proposed by Porter.

2. All the sentences are converted to its vectorial representation.

3. We compute the set of features described in the previous subsection. Continuous features are discredited: we adopt a simple "class-blind" method, which consists of separating the original values into equal-width intervals. We did some experiments with different discrimination methods, but surprisingly the selected method, although simple, has produced better results in our experiments.

4. A ML trainable algorithm is employed; we employ two classical algorithms, namely C4.5 and Naive Bayes. As usual in the ML literature, we employ these algorithms trained on a training set and evaluated on a separate test set.

5. The problems of machine learning method and their solution

Some of the features used in this method such as main concepts, the occurrence of proper nouns and non essential information have binary attributes such as zero and one which sometimes are not exact. For example, one of these features is the "main concepts" attributes; that is, if one sentence contain at least one of the given nouns, the value of that sentence is one and otherwise it is zero.

What is obvious is that the sentence containing one noun has less value than the sentence containing two nouns. But there is no deference between these two sentences in the ordinary methods. To solve this problem, we try to define these attributes as fuzzy quantities; that is each sentence, depending on the presence of each attribute, has the value ranging from zero to one. Also, to compare different sentences, we use COS formula which depends on cross product. Since all of the vector dimension (sentence attributes) are the same in cross product, each of these attributes are the same in the final result. What is clear is that some of the attributes have more importance and some have less and so they should have balance weight in computations and we use fuzzy logic to solve this problem.

6. Fuzzy logic

As the classic logic is the basic of ordinary expert logic, fuzzy logic is also the basic of fuzzy expert system. Fuzzy expert systems, in addition to dealing with uncertainty, are able to model common sense reasoning which is very difficult for general systems. One of the basic limitation of classic logic is that it is restricted to two values, true or false and its advantage is that it is easy to model the two-value logic systems and also we can have a precise deduction.
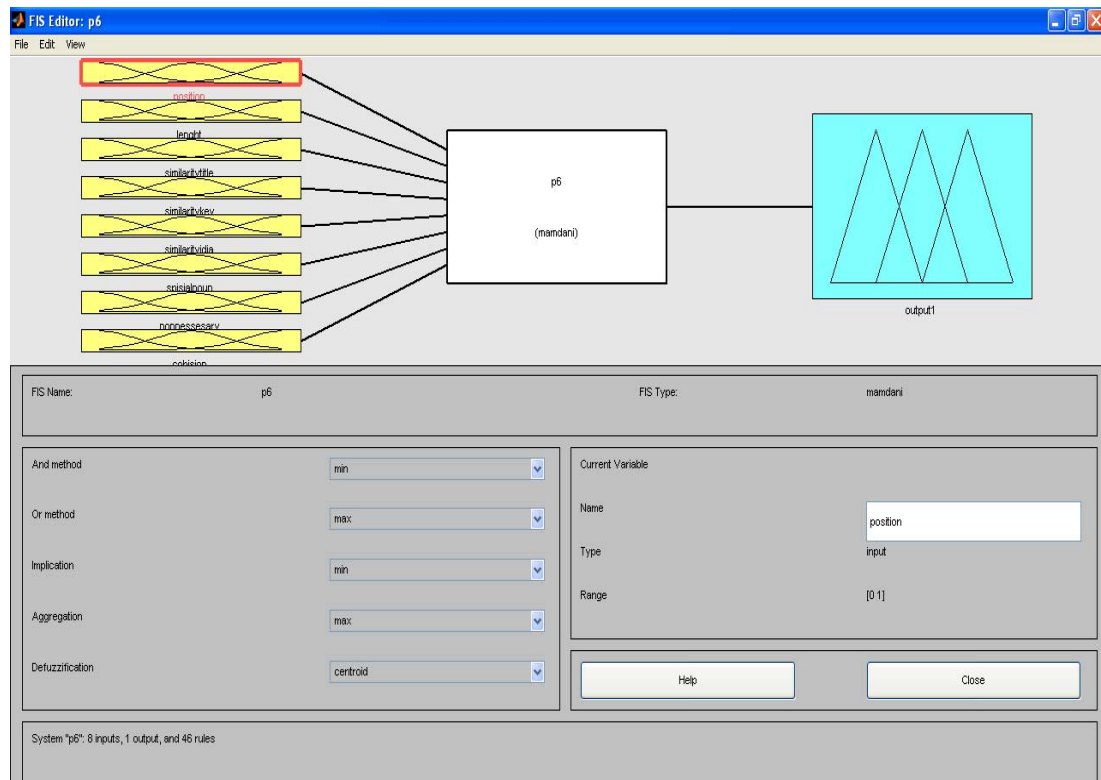

Figure 1. Producing goal function by attributes of Text Summarization

The major shortcoming of this logic is that, the number of the two-value subjects in the real world is few. The real world is an analogical world not a numerical one.

We can consider fuzzy logic as an extension of a multi-value logic, but the goals and application of fuzzy logic is different from multi-value logic since fuzzy logic is a relative reasoning logic not a precise multi-value logic. In general, approximation or fuzzy reasoning is the deduction of a possible and imprecise conclusion out of a possible and imprecise initial set [1].

7. Text summarization based on fuzzy logic

In order to implement text summarization based on fuzzy logic, 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, similarity too little, 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 like figure 1,2 and 3).
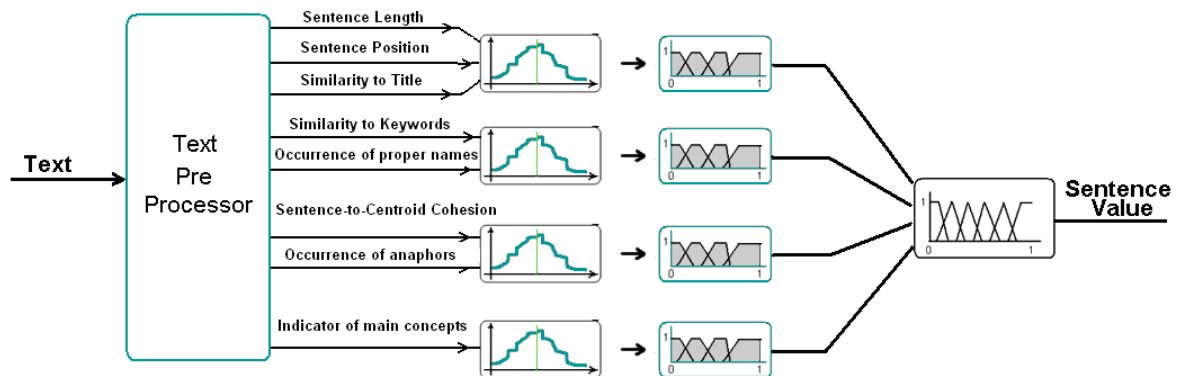
Figure 2. The whole block diagram of the proposed text summarization system

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.

8. Simulation Results

Title of original text file (to be used in text summarization) is "Dangers of racing cars" and the statements of this text file is in below table. And the summarization of these two methods (Machine Learning approach and Fuzzy approach) based on its priority is shown in this table.

Keyword: racing cars, risk, speed, dangerous, death

Dangers of racing cars
People who drive racing cars are a special breed. They risk their whenever they race, as the high the speed of their cars can be dangerous. Some have been badly injured on the racing circuit. Others have been killed. It takes a great deal of experience and skill to avoid injury or death.
Consequently, drivers must have lots of courage. They need nerves of steel and quick reflexes. They can't make errors in judgment. The slightest mistake, and the driver is doomed. One of the best drivers was Jackie Stewart. He won the world's Driving Championship three times. When he was 34, he decided to retire from racing.
Jackie's main reason for quitting was the death of his racing partner. His good friend, Francois Cevert, was killed racing at watking's Glen, New York . This incident caused Jackie to think about his future. Since he had his family to consider, he decided to race no more. Too many of his friends had lost their race with death.

Summarize with Machine learning approach

People who drive racing cars are a special breed. Some have been badly injured on the racing circuit. The slightest mistake and the driver is doomed. One of the best drivers was Jackie Stewart. He won the world's Driving Championship three times. Jackie's main reason for quitting was the death of his racing partner. His good friend, Francois Cevert, was killed racing at watking's Glen, New York . Too many of his friends had lost their race with death.
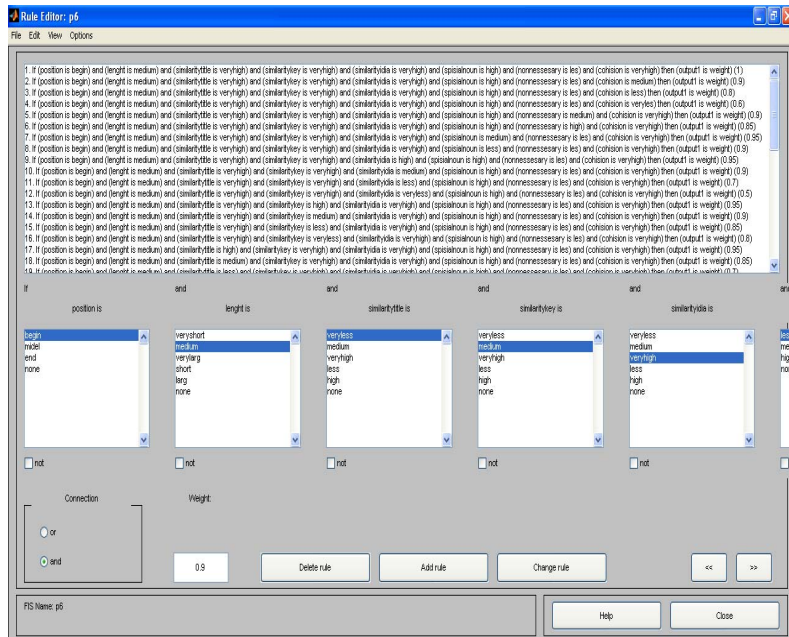
Figure 3. Rules definition of goal function

Summarize with Fuzzy logic approach

People who drive racing cars are a special breed. They risk their whenever they race, as the high the speed of their cars can be dangerous. They can't make errors in judgment. The slightest mistake, and the driver is doomed. One of the best drivers was Jackie Stewart. When he was 34, he decided to retire from racing. Jackie's main reason for quitting was the death of his racing partner. This incident caused Jackie to think about his future.

Table 1. The summarization of these two methods(Machine Learning approach and Fuzzy approach) based on its priority

Statement Number in main text	Number of statement	The priority of selected Statement in Machine Learning approach	The priority of selected Statement in Fuzzy approach	Statement
1	1	1	1	People who drive racing cars are a special breed.
2	1		2	They risk their whenever they race, as the high the speed of their cars can be dangerous.
3	1	2		Some have been badly injured on the racing circuit.
4	1			Others have been killed.

5	1			It takes a great deal of experience and skill to avoid injury or death.
6	2			Consequently, drivers must have lots of courage.
7	2			They need nerves of steel and quick reflexes.
8	2		3	They can't make errors in judgment.
9	2	3	4	The slightest mistake, and the driver is doomed.
10	2	4	5	One of the best drivers was Jackie Stewart.
11	2	5		He won the world's Driving Championship three times.
12	2		6	When he was 34, he decided to retire from racing.
13	3	6	7	Jackie's main reason for quitting was the death of his racing partner.
14	3	7		His good friend, Francois Cevert, was killed racing at watking's Glen, New York .
15	3		8	This incident caused Jackie to think about his future.
16	3			Since he had his family to consider, he decided to race no more.
17	3	8		Too many of his friends had lost their race with death.

9. Comparison

We chose 10 general text out of TOEFL text to compare the result of Machine Learning method with fuzzy method. We gave these texts and the summaries produced by both Machine Learning and fuzzy methods to 5 judges who had an M.A. in TEFL(teaching language as foreign language). We asked the judge to read the main texts and to score the summaries produced by the two methods considering the degree to which they represent the main concepts. This means that if a user has to read one of these summaries instead of reading the main text, which summary conveys concept of the main text. The given score by the judges using the two methods are shown in table No.3.

The results show that all the judge gave a better score to the summaries produced by fuzzy method. This indicates that fuzzy method worked better in parts of the sentence which contained uncertainty due to the use of fuzzy quantities. Therefore by using fuzzy approach in text summarization, we can improve the effect of available quantities for choosing sentences used in the final summaries. In order word, we can make the summaries more intelligent.

Table 3. the results of comparing Machine Learning and fuzzy logic method presented by judges

Judges methods	The First judge	The second judge	The third judge	The four judge	The five judge
Score of Machine Learning method	%55	%58	%48	%45	%65
Score of fuzzy method	%60	%65	%52	%48	%65

10. Conclusion

   In this paper, an approach has been presented for extracting sentences using fuzzy inference system. The analysis of the parameters important in summarization is done by designed fuzzy analyzers based on human perception of this problem. This text summarization system consists of (1) the text pre processor which extracts information needed for fuzzy analysis and (2) the analyzers which contain fuzzy inference systems. Weighted score of excellent sentences is computed for each sentence and the scores of relevance are ranked. Starting with the highest score, the sentences for which the relevance score is higher than the threshold value set are included in the summary. The process continues until the ratio of compression satisfies the limitation set initially. A MATLAB simulation based model has been developed for this approach. The advantage of this method is that linguistic variables and human perception are taken into consideration. Comparison of performance measures indicates the superiority of the approach as compared to the commercially available summarizers.

References

[1] Buckley,J.J and Eslami ,E. An introduction to fuzzy logic and fuzzy sets. Advances in Soft Computing. Physica-Verlag, Germany (2002).
[2] Fisher , S. and Roark , B.,Query-focused summarization by supervised sentences ranking and skewed word distribution , In proceedings of DUC(2006).
[3] Hirst , G., DiMarco , C., Hovy , E., and Parsons , K, Authoring and generating health-education documents that are tailored to the needs of the individual patient , 1997 .
[4] Hand , T.F,. A proposal for task-based evaluation of text summarization systems , in Mani , I., and Maybury , M., eds., Proceedings of the ACL/EACL97 Workshop on Intelligent Scalable Text Summarization, Madrid , Spain , 11 July 1997.
[5] Inderjcet Main , the MITRE corporation 11493 Sanset Hills noad , USA , 2003 .
[6] Louisa Ferrier , A Maximum Entropy Approach to Text Summarization, School of Artificial Intelligence , Division of Informatics , University of Edinburgh , 2001.
[7] Mani , I.: Automatic Summarization . John Benjamins Publishing Company, 2001.
[8] Mani , I.:Maybury , M.T.: Advances in Automatic Text Summarization . MIT Press, 1999.

[9] Passonneau , R., KuKich , K., Robin , J., Hatzivassiloglou , V., Lefkowitz , L., Jing , H , Generating summaries of work flow giagrams , In Proceedings of the International Conference on Natural Language Processing and Industrial Applications , New Brunswick , Canada . University of Moncton , 1996 .

[10] Passonneau , R., Kukich , K., McKeown , K., Radev , D., and Jing, H., summarizing web traffic : A portability exercise , Department of Computer Science , Columbia University , 1997 .

[11] Radev , D.R. Language Reuse and Regeneration : Generating Natural Language summaries from Multiple On-Line Sources . PhD thesis, Department of Computer Science , Columbia University , New York , 1998.

[12] Stairmond , Mark A. A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval . Ph. D. thesis , Center for Computational Linguistics , UMIST , Manchester , 1999 .

[13] Strzalkowski , T.; Stein , G.; Wang , J.; Wise , B. A Robust Practical Text summarizer . In Nani , I.; Maybury , M. (eds.) , Adv . in Autom . Text Summarization . The MIT Press , 1999 .

[14] Teufel , S,; Moens , M. Argumentayive classification of extracted sentences as a first step towards flexible abstracting. In Mani, I.; Maybury M . (eds). Advances in automatic text summarization . The MIT Press, 1999.

[15] V. Hatzivassiloglou , J. Klavans , and E. Eskin. Detecting text similarity over short passages : Exploring linguistic feature combinations via machine learning . In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora , 1999 .

Authors

Farshad Kyoomarsi
•PhD Student in Computer Science, thesis title:"Text summarization based on fuzzy logic", Mathematics and computer Group, Shahid Bahonar University, Kerman Iran.
•Head of Department of Computer Science, Islamic Azad University (Shahrekord Branch), Iran.
•**Research Area**: Natural Language Processing, Information Extraction
•**Email**: kumarci_farshad@graduate.uk.ac.ir
•**Phone**: home: +98-0381-2250262 **mobile**: 09133819862
Fax: +98- 0381-2226131

Dr. Hamid Khosravi
•PhD in Computer Science, thesis title: "Extracting Pragmatic Content from Email", Natural Language Processing Group, CS Department, University of Sheffield, UK.
•Assistant Professor, Shahid Bahonar University, Kerman Iran.
•**Research Area**: Natural Language Processing, Information Extraction
•**Email**: hkhosravi@mail.uk.ac.ir   **Phone**: 03412112865 **Office**: 0342-6226611-15
•**Fax**: 0342-6226617

Dr. Esfandiar Eslami is a Professor of mathematics in Faculty of Mathematics
and Computer Science of Shahid Bahonar University of Kerman, Kerman, Iran.
He is coauthor (with J.J. Buckley) of two books : An Introduction to Fuzzy Logic and Fuzzy Set Theory , and Fuzzy Mathematics in Economics and Engineering.
His recent interests are algebraic models of fuzzy logic and applications of fuzzy logic in modeling linguistic hedges.
•**Research Area**: Mathematics and Fuzzy logic
•**Email**: Eeslami@mail.uk.ac.ir
•**Phone**: 03412112865 **Office**: 0342-6226611-15
•**Fax**: 0342-6226617

Pooya Khosravyan Dehkordy
•MSC Student in Computer Science, thesis title :"Text summarization based on Genetic Programming", Computer Group, Islamic Azad University, (Shahrekord Branch) Iran.
•Scholarship of Computer Science Department, Islamic Azad University (Shahrekord Branch), Iran
•**Research Area**: Natural Language Processing, Genetic Programming, Cellular Automata , Software Engineering.
•**Email**: Pooya_khd@iaun.ac.ir
•**Phone**: +98-0381-3338410