2011 International Conference on Intelligent Building and Management
Proc .of CSIT vol.5 (2011) © (2011) IACSIT Press, Singapore

Optimizing Persian Text Summarization Based on Fuzzy Logic Approach

Farshad Kiyoumarsi[1 +], Fariba Rahimi Esfahani [2]

[1] Islamic Azad University-Shahrekord branch
[2] Islamic Azad University-Shahrekord branch

Abstract. With the sudden increase of information in the world and especially on the internet, text summarization has received great attention these days. This paper is an attempt to find a new method for summarizing Persian texts based on features available in Persian language and the use of fuzzy logic. We compare the new approach with other methods to show how effective the new method will be. 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.

Keywords: Text Summarization, Fuzzy Logic, Machine learning, Persian

1. Introduction

With the huge amount of information available electronically, there is an increasing demand for automatic text summarization systems. Text summarization is the process of automatically creating a compressed version of a given text that provides useful information for the user [8]. Text summarization addresses both the problem of selecting the most important portions of text and the problem of generating coherent summaries. There are two types of summarization: extractive and abstractive. Extractive summarization methods simplify the problem of summarization into the problem of selecting a representative subset of the sentences in the original documents. Abstractive summarization may compose novel sentences, unseen in the original sources. However, abstractive approaches require deep natural language processing such as semantic representation, inference and natural language generation, which have yet to reach a mature stage nowadays [7].

Automatic text summarization is the technique in which a computer automatically creates such a summary. This process is significantly different from that of human based text summarization since human can capture and relate deep meanings and themes of text documents while automation of such a skill is very difficult to implement. Document summarization refers to the task of creating document surrogates that are smaller in size but retain various characteristics of the original document [9].

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[5]. 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

+ Corresponding author. Tel.: + 98 381 2250262 ; fax: +98 381 2226131.
E-mail address: kumarci-farshad@iaushk.ac.ir.

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.

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| . |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 in English Language

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) Mean-TF-ISF. Since the seminal work of [2,3,4], 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. The final used feature is calculated as the mean value of the TF-ISF measure for all the words of each sentence.

(F2) Sentence Length. This feature is employed to penalize sentences that are too short, since these sentences are not expected to belong to the summary [9]. 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.

(F3) 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 [5]; the final value is normalized to take on values between 0 and 1.

(F4) 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.

(F5) 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 .

(F6) 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.

(F7) 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:

(F8) 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.

(F9) 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 [7]. 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.

(F10) Occurrence of proper nouns. 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 [9].

(F11) 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 [9]: 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.

(F12) 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.

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.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 to 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 and 2).[7]

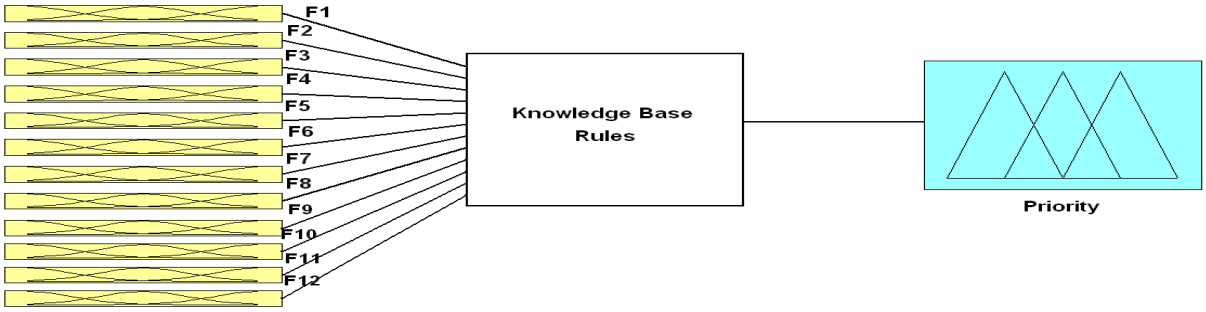

Figure 1. Producing goal function by attributes of Text Summarization

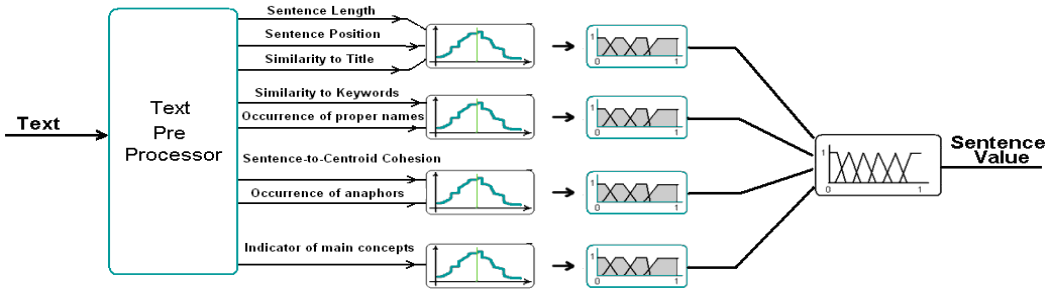

Figure2: the whole block diagram of the proposed text summarization system

After ward, 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. The Used Attribute in Text Summarization in Persian texts

In Persian, the attributes used for choosing important sentences in the final summary are a little different from English. For example, in some cases, the last sentences in the paragraph or text have higher semantic value, while in Persian the first sentences have higher value and in many cases , these attributes are the same for both languages. In this paper, we changed the previous proposed fuzzy models [9] based on their application in Persian. Then, we implement and simulate this model again.

9. Simulation Results and Comparison

In this experiment, we train all previously mentioned models on the twelve Persian features (using the same 100 Persian articles) and test these models by human judges to investigate the proposed system performance on a newswire data. Fig.3 shows the results of all models for the 100 Persian articles. Then we rank each document sentences based on this similarity value. A set of sentences is specified as a reference summary for each document based on the compression ratio.

We chose 10 general Persian 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 teaching 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.1.

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 no.1 :the results of comparing fuzzy logic **me**thod and other methods presented by judges.

	First judge	Second judge	Third judge	Forth judge	Fifth judge	Average judges
Score of fuzzy method	%83	%81	%89	%82	**%79**	%85
Score of Microsoft Word method	%65	%63	%70	%59	%58	%63
Score of Copernic method	%71	%73	%65	%68	%73	%72
Score of Pertinence method	%72	%75	%79	%79	%78	%76
Score of SweSum method	%78	%80	%83	%81	%76	%82

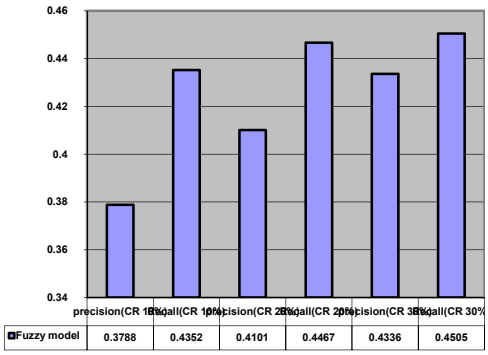

Fig.3: performance evaluation based on precision and recall for different compression rate (Persian testing data)

10. 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] Jure Leskovec, Marko Grobelnik, Natasa Milic-Frayling(2004), Learning Semantic Graph Mapping for Document Summarization, Italy.

[7] Kyoomarsi.F , Khosravi.h., Eslami.E and Davoudi.M, EXTRACTION-BASED TEXT SUMMARIZATION USING FUZZY ANALYSIS. Iranian Journal of Fuzzy Systems Vol. 7, No. 3, (2010) pp. 15-32

[8] Louisa Ferrier , A Maximum Entropy Approach to Text Summarization, School of Artificial Intelligence , Division of Informatics , University of Edinburgh , 2001.

[9] Rahimi Isfahani, Fariba. Kyoomarsi, Farshad. Khosravi, Hamid. Eslami, Esfandiar. Tajodin, Asgar, Khosravyan Dehkordy Pooya,. APPLICATION OF FUZZY LOGIC IN THE IMPROVEMENT OF TEXT SUMMARIZATION. IADIS International Conference Informatics 2008.