2012 International Conference on Software and Computer Applications (ICSCA 2012)
IPCSIT vol. 41 (2012) © (2012) IACSIT Press, Singapore

Comparing English and Persian Text SummarizationBased on Fuzzy Logic

Farshad Kiyoumarsi[1,a] , FaribaRahimi Esfahani[2,b], ParisaRahimi Esfahani[3,c]
[1,2,3]Islamic Azad University-Shahrekord branch, Iran
[a]Kumarci_farshad@iaushk.ac.ir, [b]Rahimi_fariba@yahoo.com,[c]Rahimi_parisa@yahoo.com

Abstract. Due to the great amount of information we are provided with and thanks to the development of Internet technologies, needs of producing summaries have become more and more widespread. One novel approach to text summarization is the use of fuzzy logic for extracting the most relevant sentences from an original document to form a summary. The approach utilizes fuzzy measures and inference to find the most significant sentences. This paper aims to compare the application of the most important features of the text in text summarization in English and Persian using fuzzy logic, considering the differences between these features in both languages and their importance in producing text summaries.

Keyword: Text Summarization, Fuzzy Logic, Comparison,English, Persian

1. Introduction

Document summarization refers to the task of creating document surrogates that are smaller in size but retain various characteristics of the original document [4].The document of ISO 215 standards in 1986 formally defines a summary as a "brief restatement within the document (usually at the end) of its salient findings and conclusions" that "is intended to complete the orientation of a reader who has studied the preceding text." [5].However, research into automatic text summarization has received considerable attention in the past few years due to the exponential growth in the quantity and complexity of information sources on the internet. Specifically, such text summarizer can be used to select the most relevant information from an abundance of text sources that result from a search by a search engine [5]. Many summarization models have been proposed previously. None of the models are entirely based on document structure, and they do not take into account of the fact that the human abstractors extract sentences according to the hierarchical document structure.While abstracts created by professionals involve rewriting of text, automatic summarization of documents has been focused on extracting sentences from text so that the overall summary satisfies various criteria: optimal reduction of text, coverage of document themes, and similar [5].The technique proposed here applies human expertise in the form of a set of fuzzy rules and a set of nonstructural features. Specifically, the parser is designed for selecting sentences based on their attributes and locations in the article using fuzzy logic inference system. The remarkable ability of fuzzy inference engines in making reasonable decisions in an environment of imprecision and uncertainty makes them particularly suitable [2] for applications involving risks, uncertainty, and ambiguity that require flexibility and tolerance to imprecise values. These features make them attractive for automatic text summarization.

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[4].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 [3,5] 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.

3. 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.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 [5]. 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, 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, 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 rhetorical 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 [5]. 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 .

(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 [5]: 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.

4. The Used Attribute in Text Summarization in Persian Language

Most of the main features in English texts, which were explained in previous sections, are used in Persian texts,too. However, in Persian, the attributes used for choosing important sentences in the final summary are a little different from English. For example, in some cases in English, the last sentences in the paragraph or text have higher semantic value, while in Persian the first sentences have higher value .However, in many cases; these attributes are the same for both languages. In this paper, we changed the previous proposed fuzzy models [6] based on their application in Persian. Then, we implemented and simulated this model again.

5. 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].

6. 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 1,2 and 3)[7].

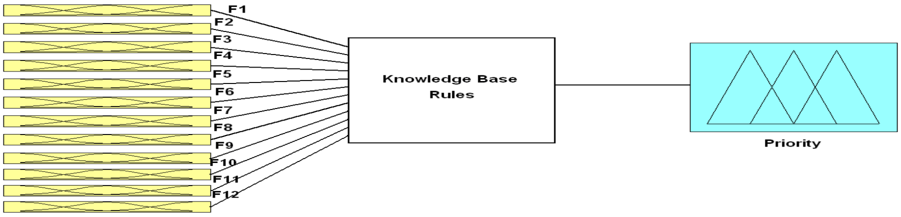


Figure 1. Producing goal function by attributes of Text Summarization

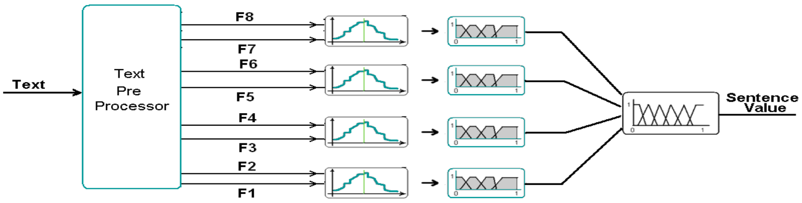
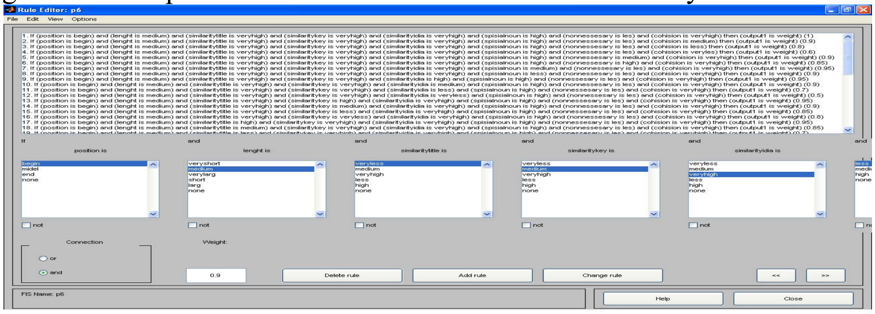


Figure2: the whole block diagram of the proposed text summarization system

Afterward, a value from zero to one is obtained for each sentence in the output based on sentencecharacteristics 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.



```
1. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (1)
2. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is medium) then (output1 is weight) (0.9)
3. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is less) then (output1 is weight) (0.8)
4. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryles) then (output1 is weight) (0.6)
5. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is medium) and (cohision is veryhigh) then (output1 is weight) (0.9)
6. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is high) and (cohision is veryhigh) then (output1 is weight) (0.85)
7. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is medium) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.95)
8. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is less) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.9)
9. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is high) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.95)
10. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is medium) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.9)
11. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is less) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.7)
12. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryhigh) and (similarityidia is veryless) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.5)
13. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is high) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.95)
14. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is medium) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.9)
15. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is less) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.85)
16. If (position is begin) and (lenght is medium) and (similaritytitle is veryhigh) and (similaritykey is veryless) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.8)
17. If (position is begin) and (lenght is medium) and (similaritytitle is high) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.95)
18. If (position is begin) and (lenght is medium) and (similaritytitle is medium) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.85)
19. If (position is begin) and (lenght is medium) and (similaritytitle is less) and (similaritykey is veryhigh) and (similarityidia is veryhigh) and (spisialnoun is high) and (nonnessesary is les) and (cohision is veryhigh) then (output1 is weight) (0.7)
```

If position is: begin (selected), midel, end, none — □ not

and lenght is: veryshort, medium (selected), verylarg, short, larg, none — □ not

and similaritytitle is: veryless (selected), medium, veryhigh, less, high, none — □ not

and similaritykey is: veryless, medium (selected), veryhigh, less, high, none — □ not

and similarityidia is: veryless, medium, veryhigh (selected), less, high, none — □ not

and less (selected), medi, high, none — □ no

Connection: ○ or ● and

Weight: 0.9

FIS Name: p6

Figure 3. Rules definition of goal function

7. Comparison

We chose 10 general English texts from TOEFL texts and 10 general Persian texts from several valid Persian texts. Then, we asked 5 Persian native speakers who were also fluent in English language, to read both the English texts and the summaries produced by Fuzzy method, and to score the extent of relatedness of the summaries to the text by percentage.

We alsoasked them to score the Persian texts and Persiansummaries produced by this method. In the following table, you can see the result of this judgment.Considering the result, it can be realized that, summarization based on fuzzy logic is valid and acceptablein both Persian and English texts from the point of view of the judges.(shown in Table no.1) However the summaries produced from English texts are more compatible with main texts and contain the main contents of the text.

Table no.1: the results of comparing fuzzy logic method in English Text and Persian Text Presented by judges.

	First judge	Second judge	Third judge	Forth judge	Fifth judge	Averagejudges
Score of fuzzy method In English Text	%75	%71	%71	%72	%72	%72
Score of fuzzy method In Persian Text	%78	%80	%85	%82	%79	%81

8. Conclusion

Comparing text summarization in English and Persian texts basedon Fuzzy method indicates that ,this method works betterin English texts in general; however, some changes in the main features of the Persian text can help the creation of better summaries from Persian documents. To achieve this end, some linguistic knowledge in Persian is neededto be familiar with basic features of Persian documentsfor producing better summaries. This issue is strongly recommended to researchers for further researches.

9. 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] Christopher C. Yang, Fu Lee Wang, Fractal Summarization: Summarization Based on Fractal Theory, SIGIR 2003, ACM 1-58113-646, Toronto, CA.

[3] Fisher , S. and Roark , B.,Query-focused summarization by supervised sentences ranking and skewed word distribution , In proceedings of DUC(2006).

[4] Khosravi, H., Eslami, E., Kyoomarsi, F., and Dehkordy, P., K., 2008 Optimizing Text Summarization Based on Fuzzy Logic. In: Book Series Studies in Computational Intelligence Publisher Springer Berlin / Heidelberg, ISSN 1860-949X (Print) 1860-9503 (Online), Volume 131/2008, Book Computer and Information Science, Copyright 2008 ,ISBN 978-3-540-79186-7, DOI 10.1007/978-3-540-79187-4_11, Pages 121-130, Subject Collection Engineering, SpringerLink Date Wednesday, May 07, 2008.

[5] KianiArman -B, M.-R. Akbarzadeh-T., M. H. Moeinzadeh, “Intelligent Extractive Text Summarization Using Fuzzy Inference Systems”, 1-4244-0457-6/06, IEEE

[6] Kyoomarsi.F ,Rahimi.F. Optimaizing Machine Learning Approach Based on Fuzzy Logic In Text Summarization. International Journal of Hybrid Information Technology (IJHIT). Vol. 2 No. 1, January 2009.

[7] RahimiIsfahani, Fariba. Kyoomarsi, Farshad. Khosravi, Hamid. Eslami, Esfandiar. Tajodin, Asgar, KhosravyanDehkordyPooya,.APPLICATION OF FUZZY LOGIC IN THE IMPROVEMENT OF TEXT SUMMARIZATION. IADIS International Conference Informatics 2008