IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 6 (Sep-Oct. 2012), PP 05-10
www.iosrjournals.org

Improvement of Text Summarization using Fuzzy Logic Based Method

[1]Rucha S. Dixit, [2]Prof. Dr.S.S.Apte
[1](Computer Science & Engg, W.I.T. Solapur/ Solapur University, India)
[2](Head, Computer Science & Engg, W.I.T. Solapur/ Solapur University, India)

ABSTRACT: *Automatic text summarization is undergoing wide research and gaining importance as the availability of online information is increasing. Automatic text summarization is to compress the larger original text into shorter text called as summary. Abstraction and Extraction are the two main methods to carry out text summarization. Our approach uses extractive method. Summarization by extraction involves identifying important features and extracting sentences based on their scores. 30 documents from news based URLs are used as input. After preprocessing the input document, eight features are used to calculate their score for each sentence. In this paper fuzzy logic method is proposed for improvement in the extraction of summary sentences.*
Keywords – *fuzzy logic, feature-based, fuzzy centroid, sentence score , text summarization.*

I. Introduction

With the exponential growth in the quantity and complexity of information sources on the internet, it has become increasingly important to provide improved mechanisms to user to find exact information from available documents. Text summarization has become an important and timely tool for helping and interpreting the large volumes of text available in documents.

Automatic text summarization is the summary of the source text created by machine to present the most important information in a shorter version of the original text while keeping its main content and helps the user to quickly understand large volumes of information. Text summarization can handle the problem of selecting the most important portions of text as well as the problem of generating coherent summaries. Automatic text summarization is significantly different from that of human based text summarization since humans can capture and relate deep meaning and themes of text documents while automation of such a skill is very difficult to implement.

Automatic text summarization can be carried out mainly by two methods: extraction and abstraction. Extraction summary method is the selection of sentences or phrases having highest score from the original text and put it together to a new shorter text without changing the source text. Abstraction summary method uses linguistic methods to examine and interpret the text. Automatic text summarization works best on well-structured documents such as news, reports, articles and scientific papers. The Extraction method for summarization involves indentifying the features such as sentence length, sentence location, term frequency, number of words occurring in title, number of proper nouns, number of numerical data and thematic word. Our approach uses feature fusion technique in order to decide which features are actually useful out of the available ones. This paper proposes text summarization based on fuzzy logic to extract important sentences as a summary.

II. Summarization Approaches

In early classic summarization system the important summaries were created according to the most frequent words in the text. Luhn created the first summarization system [1] in 1958. Rath et al. [2] in 1961 proposed empirical evidences for difficulties inherent in the notion of ideal summary. Both studies used one thematic feature called term frequency, thus they are characterized by surface level approaches. In the early 1960's new approaches called entity level approaches appeared; the first approach of this kind used syntactic analysis [3]. The location features were used in [4], where key phrases that are used dealt with three additional components: pragmatic words; title and heading words and structural indicators.

In this paper the proposed method uses fuzzy rules and fuzzy set for selecting sentences based on their features. Fuzzy logic technique in the form of approximate reasoning provides decision support and expert system with powerful reasoning capabilities. The permissiveness of fuzziness in human thought processes suggests that much of the logic behind human reasoning is not only a traditional two-values or multi-valued logic, but also logic with fuzzy truths, fuzzy connectives, and fuzzy rules of inference [5]. Fuzzy set proposed by Zadeh [6] is a mathematical tool for dealing with uncertainty, imprecision, vagueness and ambiguity. Fuzzy logic in text summarization needs more investigation. A few studies were done in this area, Witte and Bergler [7] presented a fuzzy-theory based approach to co-reference resolution and its application to text summarization. Automatic determination of co-reference between noun phrases is fraught with uncertainty. Kiani and

Akbarzadeh [8] proposed technique for summarizing text using combination of Genetic Algorithm (GA) and Genetic Programming (GP) to optimize rule sets and membership function of fuzzy systems.

The feature extraction techniques are used to obtain the important sentences in the text. In feature extraction technique some of the features have more importance and some have less so they should have balance weight in computations and we use fuzzy logic to solve this problem by defining membership function for each feature.

III. Experiment

3.1. Input Data and Preprocessing

We used news articles from news based URLs as an input to summarization system. The text portion of news article fetched from URL is saved in a text document that acts as an input document to the summarizer.
Input document is of plain text format. In this paper, preprocessing involves four main activities: Sentence Segmentation, Tokenization, Removing Stop Words and Word Stemming. Sentence segmentation is boundary detection and separating source text into sentence. Tokenization is separating the input document into individual words. Next, Removing Stop Words means removing the words which appear frequently in document but provide less meaning in identifying the important content of the document such as 'a', 'an', 'the', etc. The last step for preprocessing is Word Stemming; Word stemming is the process of removing prefixes and suffixes of each word.

3.2. Sentence Features

After this preprocessing, each sentence of the document is represented by a vector of features. We use eight features for each sentence. Each feature is given a value between '0' and '1'. There are eight features as follows.

3.2.1 Title Feature

This feature gives the measure of the similarity between the title sentence and every other sentence of the document. This is determined by counting the number of matches between the content words in a sentence and the words in the title. The score for this feature is the ratio of the number of matches between a sentence and title sentence over the number of words in title.

$$S\_F1(S) = \frac{\text{No of title words in sentence } S}{\text{No of words in title}} \quad (1)$$

3.2.2. Sentence Length

We use this feature to filter out short sentences such as datelines and author names. The short sentences are not expected to belong to the summary. Here first the length of the sentence is calculated by counting the number of words in it and then it is normalized. The score for the feature is given by the ratio of length of sentence over the length of the longest sentence in a document.

$$S\_F2(S) = \frac{\text{Length of sentence } S}{\text{Length of longest sentence in a document}} \quad (2)$$

3.2.3. Term weight

This feature uses the concept of term frequency [10] which has often been used to calculate the importance of a sentence. Here by term frequency we mean occurrence of a term within a document. We sum up the term frequency of all the term in a sentence. The score of this feature is given by the ratio of summation of term frequencies of all terms in a sentence over the maximum of summation values of all sentences in a document.
Weight of sentence 'i' is calculated by following equation.

$$W_i = \sum_{i=1}^{k} (TF_i) \quad (3)$$

Where 'k' is the number of words in a sentence.
Score for this feature is calculated as follows.

$$S\_F3(S) = \frac{W_i(S)}{\text{Max}[W_i(S)]_{i=1}^{N}} \quad (4)$$

3.2.4. Sentence Position

Position of the sentence in the text, decides its importance. This feature can involve several items such as the position of a sentence in the document, section and paragraph etc. For the score of this feature we consider the first 5 sentences in a document.

The feature score is calculated as follows.

$$S\_F4(S) = 1st\ sentence = \frac{5}{5}\ ;\ 2nd\ \ sentence = \frac{4}{5}\ ;\ 3rd\ sentence = \frac{3}{5}; 4th\ sentence = \frac{2}{5};$$
$$5th\ \ sentence = \frac{1}{5}; other\ \ sentences = \frac{0}{5} \qquad (5)$$

3.2.5. Sentence to Sentence Similarity

This feature is similarity between sentences. For each sentence S, the similarity between S and every other sentence is computed by the method of token matching. The [N][N] matrix is formed where N is the total number of sentence in a document. The diagonal elements of a matrix are set to zero as the sentence should not be compared with itself. The similarity of each sentence pair is calculated (6) as follows.

$$Sim(S_i, S_j) = TM[(t_i)_1^n, (t_j)_1^m] \qquad (6)$$

Where 'TM' stands for token matching method. The score of this feature is the ratio of summary of similarity of sentence S with every other sentence over the maximum of summary and is calculated (7) as follows.

$$S\_F5(S) = \frac{\sum[Sim(S_i, (S_j)]_1^N]}{Max(\sum[Sim(S_i, (S_j)]_1^N])_{i=1}^N} \qquad (7)$$

3.2.6. Proper Noun

The sentence that contains more proper nouns (name entity) is an important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of proper nouns that occur in sentence over the sentence length.

$$S\_F6(S) = \frac{\text{No of proper nouns in sentence S}}{\text{Sentence Length (S)}} \qquad (8)$$

3.2.7. Thematic Word

Thematic word means the word that occurs frequently in a document and it is probably related to the topic. Thematic word is the word with maximum possible relativity. We computed the frequency of occurrence of each word in a document and used the top 5 most frequent content words for consideration as thematic. The score for this feature is calculated as the ratio of the number of thematic words that occurs in a sentence over the maximum number of thematic words in a sentence.

$$S\_F7(S) = \frac{\text{No of thematic words in sentence S}}{\text{Max(No of Thematic words)}} \qquad (9)$$

3.2.8. Numerical Data

This feature is the number of numerical data in sentence. The sentence that contains numerical data is important and it is most probably included in a summary. The score for this feature is calculated as the ratio of the number of numerical data that occur in sentence over the sentence length.

$$S\_F8(S) = \frac{\text{No of numerical data in sentence S}}{\text{Sentence Length (S)}} \qquad (10)$$

3.3 Text Summarization based on Fuzzy logic

The goal of text summarization based on extraction is sentence selection. Our system consists of the following main steps.

a. Read the source document into the system;
b. For preprocessing step, the system extracts the individual sentences of the original document. Then, separate the input document into individual words. Next, remove stop words. The last step for preprocessing is word stemming;
c. Each sentence is associated with vector of eight features that described in section 3.2, whose values are derived from the content of the sentence;

d. The score for each sentence is derived from its features based on fuzzy logic method;
e. A set of highest score sentences are extracted as document summary based on compression rate.

To extract important sentences we used fuzzy logic method. Fuzzy logic usually implicates selecting fuzzy rules and membership function. The selection of fuzzy rules and membership functions directly affect the performance of the fuzzy logic system.

The fuzzy logic system consists of four components: fuzzifier, inference engine, defuzzifier and the fuzzy knowledge base. In the fuzzifer, crisp inputs are translated into linguistic values by using a membership function. After fuzzification, the inference engine refers to the rule base containing fuzzy IF-THEN rules to derive the linguistic values. In the last step, the output linguistic variables from the inference engine are converted to the final crisp values by the defuzzfier using membership function for representing the final sentence score. Fig. 1 shows text summarization based on fuzzy logic system architecture.

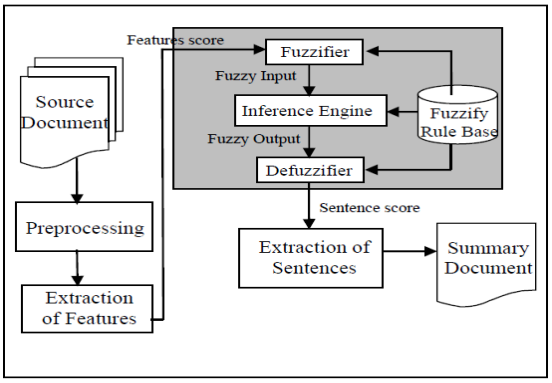

Figure 1: Text summarization based on fuzzy logic system

In order to implement text summarization based on fuzzy logic, first, the eight features extracted in the previous section are used as input to the fuzzifier. We used triangular membership functions and fuzzy logic to summarize the document. The input triangular membership function for each feature is divided into five fuzzy sets which are composed of unimportant values (low (L) and very low (VL), average values (medium (M) and important values (high (H) and very high(VH). The generalized triangular membership function depends on the three parameters a, b, and c where, 'a' and 'c' are left and right feet of a triangle and 'b' is at the peak of a triangle as shown Fig. 2. We used fuzzy centroid method to calculate a score for each sentence of a document. A value from zero to one is obtained for each sentence in the output based on sentence features and knowledge base. The obtained value in the output determines the degree of importance of the sentence in the final summary. The simplified fuzzy centroid calculation (11) is given by the following formula.

$$C(x,y) = \left(\frac{a+b+c}{3}, \frac{l+m+n}{3}\right) \qquad (11)$$

Where a, b, c are standard values of low, medium and high respectively and l, m, n are calculated values of low, medium and high respectively.

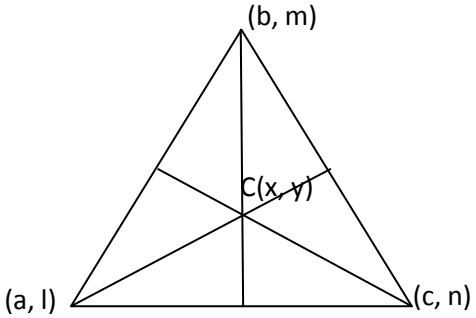

Figure 2: Fuzzy Centroid Calculation

In inference engine the most important part is the definition of fuzzy IF-THEN rules. The important sentences are extracted from these rules according to our features criteria. Sample of IF-THEN rules shows as the following rule.

IF (NoWordInTitle is VH) and (SentenceLength is H) and (TermFreq is VH) and (SentencePosition is H) and (SentenceSimilarity is VH) and (NoProperNoun is M) and (NoThematicWord is VH) and (NumericalData is M) THEN (Sentence is important)

Likewise the last step in fuzzy logic system is the defuzzfication. We used the output membership function which is divided into three membership functions: Output {Unimportant, Average, Important} to convert the fuzzy results from the inference engine into a crisp output for the final score of each sentence.

3.4. Extraction of Sentences

In the fuzzy logic method described above, each sentence of the document is represented by the sentence score. Then all the sentences of a document are ranked in a descending order based on their scores. A set of highest score sentences are extracted as document summary based on 20% compression rate. Finally the summary sentences are arranged in the original order.

4. Evaluation and Results

The evaluation of the summaries is done based on two factors mentioned in Fig. 5. We used 30 news documents from news based URLs as an input to the system. Here the human generated summaries are used as reference summaries for evaluation of our results. The human generated summary acts as a gold standard summary since humans can capture and relate deep meanings of the text as compared to machines. We received human generated summaries for our input documents from five different Experts. Here we call the summaries of Fuzzy summarizer, Copernic summarizer and MS Word summarizer as the candidate summaries and human generated summaries as reference summaries.

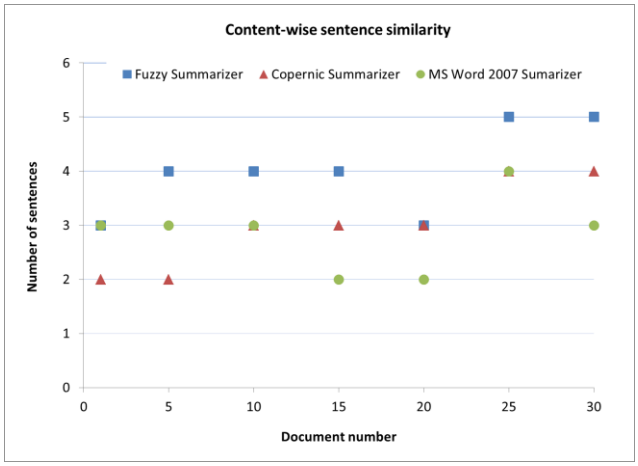

Figure 3: Content-wise sentence similarity between candidate summary and reference summary

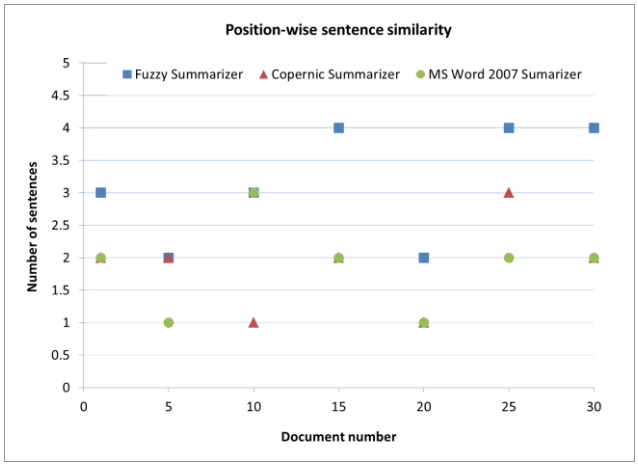

Figure 4: Position-wise sentence similarity between candidate summary and reference summary

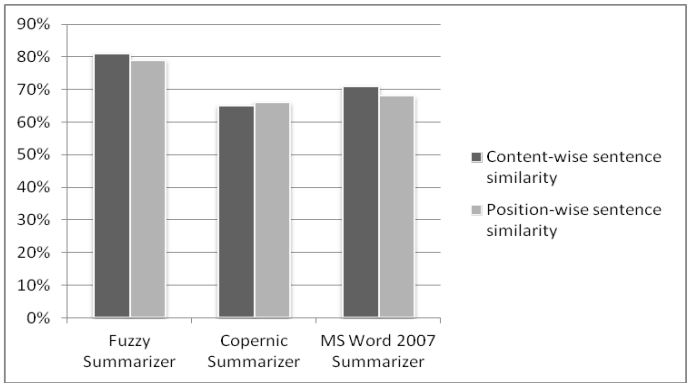

Figure 5: Average similarity between candidate summary and reference summary

The first factor in the evaluation as mentioned in Fig. 3 is the number of sentences from candidate summary that are similar to the sentences in reference summary. For this factor fuzzy summarizer exhibits the highest average 81% resemblance to the reference summary as shown in Fig. 5. The second factor in the evaluation as mentioned in Fig. 4 is the similarity between positions of the sentences in candidate summary and reference summary. For this factor fuzzy summarizer has average 79% resemblance to the reference summary which is highest among others as shown in Fig. 5. The results clearly show that fuzzy summarizer approach under consideration performs better than Copernic summarizer and MS Word 2007 summarizer.

IV. Conclusion and Future Work

In this paper we implement automatic text summarization which involves feature based extraction of important sentences using fuzzy logic method. The system is tested with 30 news documents and compared with Copernic summarizer and MS Word 2007 summarizer. The results show that the use of fuzzy logic in text summarization improves the quality of summaries. We applied our method for single document summarization which could be extended for multi-document summarization. In the input we used 30 news documents. The news have categorizes like sports, politics, weather etc. It is necessary to incorporate automatic selection of fuzzy inference rules, based on the type of input news in order to generate the best summaries. Our method could be extended for automatic selection of fuzzy inference rules.

V. Acknowledgement

We would like to thank Solapur University and WIT Faculty for supporting us.

References

Journal Papers:

[1] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, vol. 2, pp.159-165.1958.

[2] G. J. Rath, A. Resnick, and T. R. Savage, "The formation of abstracts by the selection of sentences" American Documentation, vol. 12, pp.139-143.1961.

[4] H. P. Edmundson., "New methods in automatic extracting" Journal of the Association for Computing Machinery 16 (2). pp.264-285.1969.

[5] A. D. Kulkarni and D. C. Cavanaugh, "Fuzzy Neural Network Models for Classification" Applied Intelligence 12, pp.207-215. 2000.

[6] L. Zadeh, "Fuzzy sets. Information Control" vol. 8, pp.338–353.1965.

[10] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval" Information Processing and Management 24, 1988. 513-523. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) Readings in I.Retrieval. Morgan Kaufmann. pp.323-328.1997.

Books:

[3] Inderjeet Mani and Mark T. Maybury, editors, Advances in automatictext summarization (MIT Press. 1999)

Proceedings Papers:

[7] R Witte and S. Bergler, "Fuzzy coreference resolution for summarization" In Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS). Venice, Italy: Università Ca' Foscari. pp.43–50. 2003.

[8] Arman Kiani and M.R. Akbarzadeh, "Automatic Text Summarization Using: Hybrid Fuzzy GA-GP" In Proceedings of 2006 IEEE International Conference on Fuzzy Systems, Sheraton Vancouver Wall Center Hotel, Vancouver, BC, Canada. pp.977-983.2006.

[9] J. Kupiec. , J. Pedersen, and F. Chen, "A Trainable Document Summarizer" In Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Seattle, WA, pp.68-73.1995.