AUTOMATIC TEXT SUMMARIZATION USING FEATURE-BASED FUZZY EXTRACTION

Ladda Suanmali[1], Naomie Salim[2], Mohammed Salem Binwahlan[3]

[1]Faculty of Science and Technology
Suan Dusit Rajabhat University
295 Rajasrima Rd, Dusit, Bangkok, Thailand 10300

[2,3]Faculty of Computer Science and Information System
Universiti Teknologi Malaysia
81310 Skudai, Johor

E-mail: [1]ladda_sua@dusit.ac.th, [2]naomie@utm.my, [3]moham2007med@yahoo.com

Abstract: Automatic text summarization is to compress the original text into a shorter version and help the user to quickly understand large volumes of information. This paper focuses on the automatic text summarization by sentence extraction with important features based on fuzzy logic. In our experiment, we used 6 test documents in DUC2002 data set. Each document is prepared by preprocessing process: sentence segmentation, tokenization, removing Stop Word and stemming Word. Then, we use 8 important features and calculate their score for each sentence. We propose a method using fuzzy logic for sentence extraction and compare our result with the baseline summarizer and Microsoft Word 2007 summarizers. The results show that the highest average precision, recall, and F-mean for the summaries are conducted from fuzzy method.

Keyword: Automatic text summarization, Fuzzy logic, Sentence extraction

1. INTRODUCTION

Automatic text summarization is the summary of the source text by machine to present the most important information in a shorter version of the original text while still keeping its main semantic content and helps the user to quickly understand large volumes of information. Text summarization addresses both the problem of selecting the most important portions of text and the problem of generating coherent summaries. 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. A number of researchers have proposed techniques for automatic text summarization which can be classified into two categories: extraction and abstraction. Extraction summary is a selection of sentences or phrases from the original text with the highest score 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. Most of the current automated text summarization system use extraction method to produce summary. Automatic text summarization works best on well-structured documents, such as news, reports, articles and scientific papers.

In this paper, we propose text summarization based on fuzzy logic aided method to extract important sentences as a summary of document. The rest of this paper is organized as follows. Section 2 presents the summarization approach. Section 3 and 4 describes our proposed, followed by experimental design, experimental results and evaluation. Finally, we conclude and suggest future work that can be carried out in Section 5.

2. SUMMARIZATION APPROACH

Automatic text summarization dates back to the Fifties, when 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 thematic features such as term frequency, thus they characterized by surface-level approaches. In the early 1960s, 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 are used dealt with three additional components: pragmatic words (cue words, i.e., words would have positive or negative effect on the respective sentence weight like significant, key idea, or hardly); title and heading words; and structural indicators (sentence location, where the sentences appearing in initial or final of text unit are more significant to include in the summary.

In this paper, we propose important sentence extraction used fuzzy rules and a set for selecting sentences based on their features. Fuzzy set proposed by Zadeh [10] is a mathematical tool for dealing with uncertainty, vagueness and ambiguity. Its application in text representation for information retrieval was first proposed by Buell [11], in which a document can be represented as a fuzzy set of terms. Miyamoto [12] investigated applications of fuzzy set theory in information retrieval and cluster analysis. Witte and Bergler [13] 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 [15] 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 locate the important sentences in the text. For instance, Luhn looked at the frequency of word distributions as frequent words should indicate the most important concepts of the document. Some of features are used in this research such as sentence length. Some sentences are short or some sentences are long. 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 by defining the membership functions for each feature.

3. EXPERIMENT

3.1 Data Set

We used 6 documents from DUC2002. Each document consists of 16 to 56 sentences with an average of 31 sentences. The DUC2002 collection provided [10]. Each document in DUC2002 collection is supplied with a set of human-generation summaries provided by two different experts. While each expert was asked to generate summaries of different length, we use only 100-word variants. DUC2002 for automatic single-document summarization create a generic 100-word summary.

3.2 Preprocessing

Currently, input document are of plain text format. In this paper, we use Microsoft Visual C# 2008 for preprocessing data. There are four main activities performed in this stage: Sentence Segmentation, Tokenization, Removing Stop Word, and Stemming Word. Sentence segmentation is boundary detection and separating source text into sentence. Tokenization is separating the input document into individual words. Next, Removing Stop Words, stop words are 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 foı preprocessing is Stemming word; Stemming word is the process of removing prefixes and suffixes of each word.

3.3 Features in Text Summarization

In order to use a statistical method it is necessary to represent the sentences as vectors of features. These features are attributes that attempt to represent the data used for the task. We concentrate our presentation in eight features for each sentence. Each feature is given a value between '0' and '1'. Therefore, we can extract the appropriate number of sentences according to compression rate. There are eight features as follows:

(1) **Title feature:** The number of title word in sentence, words in sentence that also occur in title gives high score [6]. This is determined by counting the number of matches between the content words in a sentence and the words in the title. We calculate the score for this feature which is the ratio of the number of words in sentence that occur in the title over the number of word in title.

$$Score\ (S_i) = \frac{No.Title\ word\ in\ S_i}{No.Word\ in\ Title} \qquad (1)$$

(2) **Sentence length:** The number of word in sentence, this feature is useful to filtering out short sentences such as datelines and author names commonly found in news articles. The short sentences are not expected to belong to the summary [5]. We use 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.

$$Score\ (S_i) = \frac{No.Word\ occurring\ in\ S_i}{No.Word\ occurring\ in\ longest\ sentence} \qquad (2)$$

(3) **Term weight:** Calculating the average of the TF-ISF (Term frequency, Inverse sentence frequency). The frequency of term occurrences within a document has often been used for calculating the importance of sentence [7].

$$Score\ (S_i) = \frac{Sum\ of\ TF\text{-}ISF\ in\ S_i}{Max(Sum\ of\ TF\text{-}ISF)} \qquad (3)$$

(4) **Sentence position:** Whether it is the first and last sentence in the paragraph, sentence position in text gives the importance of the sentences. This feature can involve several items

such as the position of a sentence in the document, section, paragraph, etc., [14] proposed first and last sentence highest ranking. The score for this feature: 1 for first and last sentence, 0 for other sentence.

$$Score\ (S_i) = 1\ for\ First\ and\ Last\ sentence,\ 0\ for\ other\ sentences \qquad (4)$$

(5) **Sentence to sentence similarity:** Similarity between sentences, for each sentence *s*, the similarity between *s* and each other sentence is computed by the cosine similarity measure. The score of this feature for a sentence *s* is obtained by computing the ratio of the summary of sentence similarity of sentence *s* with each other sentence over the maximum of summary

$$Score\ (S_i) = \frac{Sum\ of Sentemce\ Similarity\ in\ S_i}{Max(Sum\ of\ Sentence\ Similarity)} \qquad (5)$$

(6) **Proper noun:** The number of proper noun in sentence, sentence inclusion of name entity (proper noun). Usually the sentence that contains more proper nouns is an important one and it is most probably included in the document summary [17]. The score for this feature is calculated as the ratio of the number of proper nouns in sentence over the sentence length.

$$Score\ (S_i) = \frac{No.\ Proper\ nouns\ in\ S_i}{Length\ (S_i)} \qquad (6)$$

(7) **Thematic word:** The number of thematic word in sentence, this feature is important because terms that occur frequently in a document are probably related to topic. The number of thematic words indicates the words with maximum possible relativity. We used the top 10 most frequent content word for consideration as thematic. The score for this feature is calculated as the ratio of the number of thematic words in sentence over the sentence length

$$Score\ (S_i) = \frac{No.\ Thematic\ word\ in\ S_i}{Length\ (S_i)} \qquad (7)$$

(8) **Numerical data:** The number of numerical data in sentence, sentence that contains numerical data is important and it is most probably included in the document summary [16].The score for this feature is calculated as the ratio of the number of numerical data in sentence over the sentence length

$$Score\ (S_i) = \frac{No.\ Numerical\ data\ in\ S_i}{Length\ (S_i)} \qquad (8)$$

3.4 Text Summarization based on Fuzzy Logic

In order to implement text summarization based on fuzzy logic, we use MATLAB since it is possible to simulate fuzzy logic in this software. First, the features extracted in previous section are used as input to the fuzzy inference system. We used Bell membership functions. The generalized Bell membership function depends on three parameters a, b, and c as given by (9)

$$f\ (x; a, b, c) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}} \qquad (9)$$

where the parameter *b* is usually positive. The parameter *c* and *a,* locate the center and width of the curve.

For instance, membership function of sentence to sentence similarity is show in Figure 1.

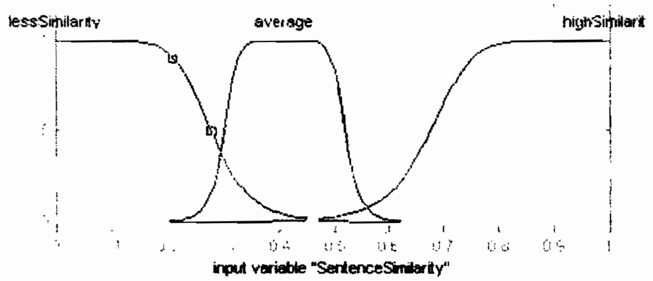

Figure 1. Membership function of sentence to sentence similarity

Afterword, we use fuzzy logic to summarize the document. 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.

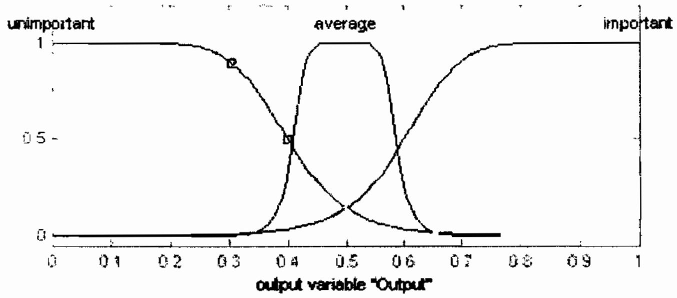

Figure 2. Membership function of Output

The input membership function for each feature is divided into three membership functions which are composed of insignificant values, average and significant values. For example, membership functions for title feature: SetenenceSimilarity {LessSimilarity, Average, and HighSimilarity}. Likewise, the output membership function is divided into three membership functions: Output {Unimportant, Average, and Important}. The most important part in this procedure is the definition of fuzzy IF-THEN rules. The important sentences are extracted from these rules according to our features criteria. For example our rules are showed as follow.

IF (NoWordInTitle is many) and (SentenceLength is long) and (TermFreq is very much) and (SentencePosition is first-last position) and (SentenceSimilarity is highSimilarity) and (NoProperNoun is many) and (NoThematicWord is many) and (NumbericalData is many) THEN (Sentence is important)

Figure 3. Sample of IF-THEN Rules

4. EVALUATION AND RESULT

We use the ROUGE, a set of metrics called Recall-Oriented Understudy for Gisting Evaluation, evaluation toolkit [8] that has become standards of automatic evaluation of summaries. It compares the summaries generated by the program with the human-generated (gold standard) summaries. For comparison, it uses n-gram statistics. Our evaluation was done using n-gram setting of ROUGE, which was found to have the highest correlation with human judgments, namely, at a confidence level of 95%. It is claimed that ROUGE-1 consistently correlates highly with human assessments and has high recall and precision significance test with manual evaluation results. So we choose ROUGE-1 as the measurement of our experiment results. In the table 1, we compare fuzzy summarizer with baseline summarizer form DUC2002 data set and Microsoft Word 2007 Summarizer.

Table 1. The result of comparing Fuzzy Summarizer and other Summarizers using Document set D061

Document	Fuzzy Sumarizer			Baseline			MS-Word Summarizer		
	P	R	F	P	R	F	P	R	F
AP880911-0016	0.59223	0.60396	0.59804	0.41748	0.40952	0.41346	0.55556	0.42857	0.4838
AP880912-0095	0.45484	0.48001	0.47607	0.43636	0.41379	0.42478	0.49231	0.44545	0.4166
AP880912-0137	0.48039	0.47573	0.47805	0.46602	0.47059	0.46829	0.47525	0.47525	0.4705
AP880915-0003	0.49038	0.48571	0.48803	0.44330	0.40952	0.42574	0.48571	0.48113	0.4834
AP880916-0060	0.50816	0.46714	0.48095	0.32642	0.32642	0.32222	0.31148	0.33929	0.3247
WSJ880912-0064	0.49524	0.51485	0.50485	0.49515	0.50495	0.50000	0.44231	0.42593	0.4339
Average	**0.50354**	**0.50457**	**0.50433**	**0.43079**	**0.42247**	**0.42575**	**0.46044**	**0.43260**	**0.4355**

The results are show in Table 1. Baseline reaches an average precision of 0.43079, average recall of 0.42247 and average F-mean of 0.42575; while Microsoft Word 2007 summarizer reaches an average precision 0.46044, recall of 0.43260 and F-mean of 0.43555. The fuzzy summarizer achieves an average precision of 0.50354, recall of 0.50457 and F-mean of 0.50433.

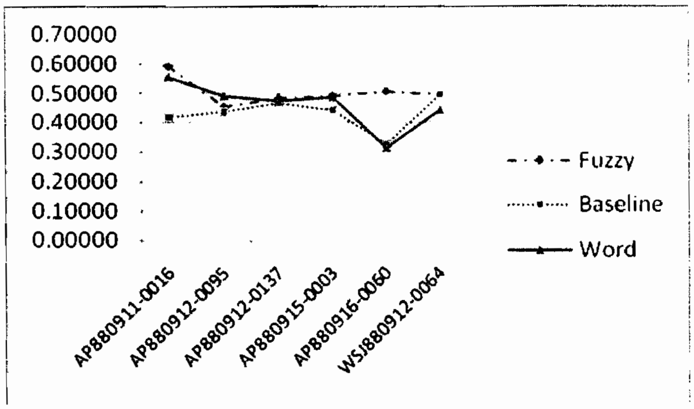

Figure 4. Precision result under difference summarizer using Document Set D061

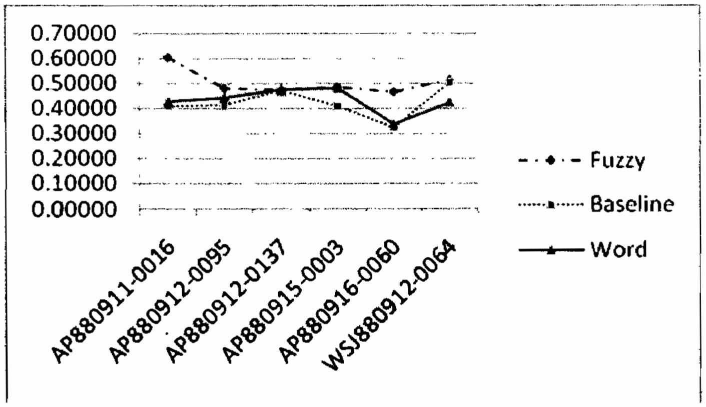

Figure 5. Recall result under difference summarizer using Document Set D061

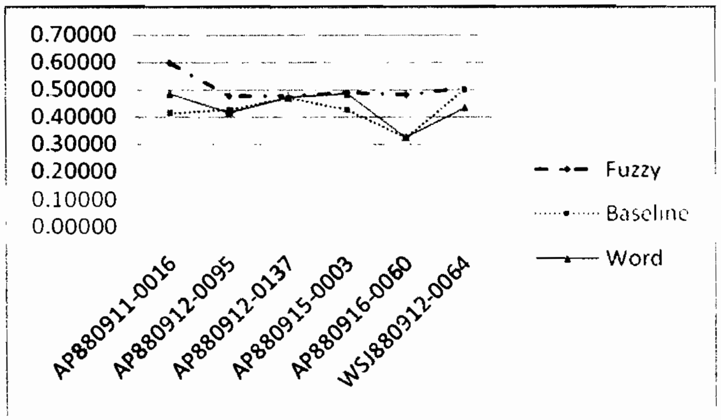

Figure 6. F-mean result under difference summarizer using Document Set D061

The results clearly show that fuzzy summarizer approach under consideration perform better than baseline summarizer and Microsoft Word 2007 summarizer. We further compare the performance of the fuzzy summarizer and other summarizer by examining their precision, recall and f-mean results. In this case, the best precision, recall and f-mean from Figure 4, 5, and 6 shows that the judges from fuzzy summarizer are the highest score. The score are as followed: 0.59223, 0.60396, and 0.59804. The significant performance improvement over fuzzy logic provides strong evidence of its feasibility in text summarization applications

5. CONCLUSION AND FUTURE WORK

In this paper, we propose automatic text summarization for important sentence extraction with important features based on fuzzy logic; title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word and numerical data. We choose 6 documents from DUC2002 data set and compare our summarizer with the baseline summarizer and Microsoft Word 2007 summarizers. The results show that the judge gave a better average precision, recall and f-mean to summaries produced by fuzzy method. Our method is intent to be used for single document summarization as well as multi documents summarization. We conclude that we need to extend the proposed method for multi document summarization and combine fuzzy logic and other learning methods in a large data set.

REFERENCES

[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.

[3] Inderjeet Mani and Mark T. Maybury, editors., "Advances in automatic text summarization," MIT Press. 1999.

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

[5] S. D. Afantenos, V. Karkaletsis and P. Stamatopoulos., "Summarization from Medical Documents: A Survey," Artificial Intelligence in Medicine, vol. 33, pp.157-177. 2005.

[6] 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.

[7] G. Salton., "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," Addison-Wesley Publishing Company. 1989.

[8] C.Y. Lin., "ROUGE: A Package for Automatic Evaluation of Summaries," In Proceedings of Workshop on Text Summarization of ACL, .Spain. 2004.

[9] DUC. Document understanding conference 2002 (2002), http://www-nlpir.nist.gov/projects/duc

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

[11] D. Buell., "An analysis of some fuzzy subsets application to information retrieval systems," Fuzzy Sets and Systems, vol. 7, no. 1, pp.35–42.1982.

[12] S. Miyamoto., "Fuzzy Sets in Information Retrieval and Cluster Analysis," Kluwer Academic Publishers, 1990.

[13] 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. http://rene-witte.net.

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

[15] 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.

[16] C.Y. Lin., "Training a selection function for extraction," In Proceedings of the eighth international conference on Information and knowledge management, Kansas City, Missouri, United States. pp.55–62. 1999.

[17] 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.