Automatic Text Summarization Using Fuzzy Inference

Mehdi Jafari
Eyvanakey Higher Institute Education
Jafari.teach@gmail.com

Amir Shahab Shahabi
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
shahabi_amir@azad.ac.ir

Jing Wang, Yongrui Qin
School of Computing and Engineering
University of Huddersfield, United Kingdom
{j.wang2, y.qin2}@hud.ac.uk

Xiaohui Tao
Faculty of Health, Engineering and Sciences
University of Southern Queensland, Australia
xtao@usq.edu.au

Mehdi Gheisari
Young Researchers and Elite club, Parand Branch, Islamic Azad University, Parand, Iran
Mehdi.gheisari61@gmail.com

***Abstract*—Due to the high volume of information and electronic documents on the Web, it is almost impossible for a human to study, research and analyze this volume of text. Summarizing the main idea and the major concept of the context enables the humans to read the summary of a large volume of text quickly and decide whether to further dig into details. Most of the existing summarization approaches have applied probability and statistics based techniques. But these approaches cannot achieve high accuracy. We observe that attention to the concept and the meaning of the context could greatly improve summarization accuracy, and due to the uncertainty that exists in the summarization methods, we simulate human like methods by integrating fuzzy logic with traditional statistical approaches in this study. The results of this study indicate that our approach can deal with uncertainty and achieve better results when compared with existing methods.**

Keywords— text summarization; fuzzy logic; sense relation

I. INTRODUCTION

The summarization of a context is the process of recognizing and indicating the most important components of a document or a set of documents. These identified components can be associated in a much smaller volume in comparison with the original context. These components should have a high accordance and relationship with the subject or concept of that is detailed in the document [1].

Considering the scenario described above, it is very useful for us to have an automatic system that enables us to summarize texts and helps in the analysis of large sets of documents [2,3]. Generally, summarization methods can be divided into two categories: the extractive and the abstractive summarization.

In the extractive method, some properties for each part of the text are determined; and based on these properties and characteristics, the level of importance of each part of the text is determined and finally best of them will be selected as a component of the final summary. But in the abstractive methods, usually the techniques of the natural language processing are used. And because of difficulties in general natural language processing, it is usually difficult to achieve satisfied results. The abstractive methods are usually a combination of both extractive and abstractive methods [4]. Because of a lack of certainty in the text summarization and to deal with uncertainty, fuzzy logic [5] is used. Fuzzy logic is used to measure the degree of importance and correlation and also to highlight the important phrases to create summarization. In this work, we integrate fuzzy logic with traditional extractive and abstractive approaches for text summarization. Through such technique, we obtain a summary of the original text, which can best reflect the main delivery of the original context.

The remainder of the paper is organized as follows. In Section II, we review related work on text summarization. Then our approach is detailed in Section III. In Section IV, experimental evaluations are presented to shown the benefits our approaches over exiting approaches. Finally, we conclude our work in Section V.

II. RELATED WORK

In [6], [7], methods for text-based summarization by using the fuzzy logic and fuzzy inference system are presented. Genetic Algorithm and Genetic Programming are also used to optimize the rule sets and membership functions of the fuzzy system.

The results of the comparison of the fuzzy-based methods and other methods have been achieved [8]. The main problem of previous work is that only syntactic parameters and semantic parameters are used. But the semantic relationship between words is ignored. Ignorance of the words semantic relations and sentences and the sole attention to the syntactic structures of the words and the sentences is mainly because that they could not help improve the accuracy of text summarization. On the other hand, if we just pay attention to semantic relations of expressions but ignore syntactic structures, it does not improve the accuracy of text summarization either [9]. In this paper, we propose applying both semantic relations and syntactic structures of expressions in producing text summarization and using the combination of these two aspects to improve the quality of summarization.

III. Fuzzy Inference Based Text Summarization

The overall flowchart of text summarizer proposed in this paper is given in Figure 1. In the preprocessing phase, to convert the original text into the text that will be used as an input parameter in the summarization system, the following steps should be taken:

- **Remove redundant words**. Ordinary words that do not have any specific data and do not show any value will be deleted, such as "the", "an", "a", etc.
- **Case folding**. Either uppercase letters are converted into lowercase or all lowercase are converted into uppercase. Here we convert all characters into the lowercase.
- **Stemming**. The derived words are converted into their stemming. For example, the names that are in the plural forms should be converted into their single forms and the verbs should be converted into their original forms. It should be noted that here we do not include the stemming phase to extract the semantic parameters, because we are using databases such as Wordnet [10] and we want the main words to be protected.

After the preprocessing phase is finished, the parameter extraction phase will start. In the following, syntactic and semantic parameters that are used in this paper are described.

A. Syntatic Parameters

TF/ISF (Term Frequency–Inverse Sentence Frequency**):** This parameter is actually inspired by the information retrieval fields, where the TF/IDF (Term Frequency–Inverse Document Frequency) concept is introduced. Generally the parameter TF/IDF is used in multiple documents summarization but in this paper, it is used for a single document that is converted into TF/ISF. To calculate this parameter, first the parameter TF is calculated for each word. The value of the TF of each word equals to the number of occurrences of the word in a document, divided by the total number of the words. Then we get the value of ISF for each word that is equal to the number of statements that include that word, e.g., N_{term} dividing the total number of the sentences N. Thus the value of TF/ISF for each word is given in Equation (1):

$$TF/ISF_{term} = TF_{term} * \lg (N_{term}/N) = TF_{term} * ISF_{term} \quad (1)$$

Then, to compute the TF/ISF values of each statement, the values of TF/ISF for all words are added up together and we obtain the TF/ISF values of a sentence. We normalize the TF/ISF by dividing this using the maximum value of TF/ISF among all the sentences.

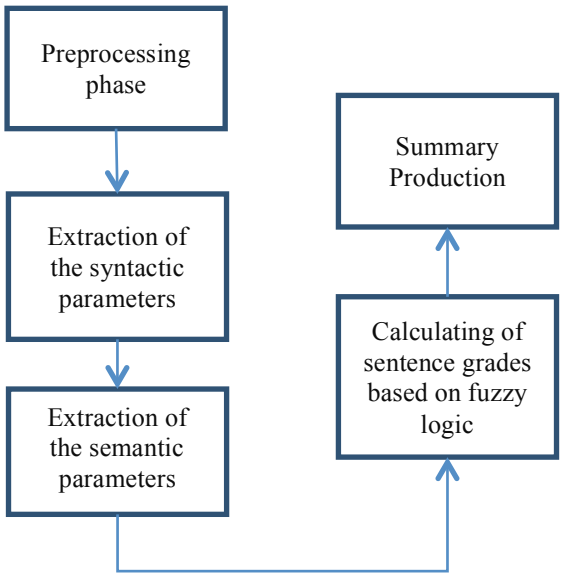

Figure 1: Flowchart summarizing the proposed approach

Sentence Length: This parameter is in fact the total number of main words. But very short sentences will be removed by the penalty and then they will not be placed in the summary. Dividing the largest sentence can help to normalize other parameters.

Location of Sentence: We can consider the relative position of a sentence in the whole document or in the section or in the paragraph, and so on. Here, the entire document will be considered in Cartesian Space and each sentence has an "x" and a "y" value.

The value of "y" is the number of paragraph that sentence is placed and the value of "x" is the number of sentence in that paragraph. Only five sentences of each paragraph are considered important. The location value of a sentence in the paragraph is the inverse order of the sentence number i.e. the first sentence has the value of the 5/5 and the second sentence has the value of 4/5, and so on [11].

Similarity to the Title: We use the vector method to split the title words and separate the words of the sentence. Then we use the cosine measure [12] to measure similarity between sentences and the title of the document. More details about how to convert sentences and titles into vectors should be referred to [12]. Then similarity values will be calculated as in Equation (2) (according to reference [12]):

$$SimToTitle_{sen} = \frac{\overrightarrow{titleVect} \cdot \overrightarrow{senVect}}{|titleVect| * |senVect|} \quad (2)$$

Similarity to Keywords: This parameter is the same as the previous parameters and to calculate it, the cosine measure between the desired vector and the vector of keywords can be calculated. The ten words that have the highest TF/ISF throughout the document are considered keywords.

Text-to-text coherence: For each sentence, we add all values of that sentence with other sentences in the same document. In order to calculate the similarity value between two sentences, we again use the cosine measure similar to the way of computing *Similarity to the Title* and *Similarity to Keywords*. Then we normalize this value by dividing it using the obtained maximum value for all sentence similarities.

Integrated text-to-center: Initially, we get the central sentence by using the sentence location parameter as described above. Then, for each sentence, the similarity value of that sentence is obtained and normalized by dividing it using the maximum obtained value among all sentences.

Key Concepts: The main concepts actually are 15 distinct words with the highest TF/ISF that have the following property: all words should be nouns. If a sentence contains those words it should be included in the final summary, otherwise not. But this does not consider sentences that contain only some of the 15 key words. For example, a sentence may contain a word of the main concepts and another sentence may contain 5 words of the concept words. To distinguish all these cases, we use fuzzy logic.

The nouns: Specific names/nouns may refer to people, places, and so on. These nouns are critical in generating a better summary. If a sentence contains specific names/nouns, it should be included in the final summary, otherwise not. However, we need to be selective during this process. If a sentence contains more nouns, it would tend to be a better sentence. Here, we apply fuzzy logic again to determine the selection of the final sentences. To identify the specific nouns, if the word is also not in the Wordnet database, it will be considered as a specific noun.

Non-basic information: If expressions (speech markers), such as because, while, in addition, and so forth, which usually occur at the beginning of the sentence, are in a sentence, we say that sentence contains non-basic information. If any sentence includes such expressions, the sentence is likely to be the emphasizing sentences, explaining sentences, or proverbial sentences. Such sentences are that we should avoid to include in the summary.

Anaphors: This parameter is the repetition of an expression in the successive sentences. The anaphors actually contain non-basic information and if a sentence contains an anaphor, its contents are covered by other related sentences. To compute this parameter, we will examine each sentence by its adjacent sentences. If up to six words are recurring, then the examined sentence is an anaphor.

After obtaining the syntactic parameters, then we need to go through the semantic parameters that are described below.

B. Semantic Parameters

The linguistic parameters use semantic knowledge such as semantic relationships between words and their combinational syntax. Based on such information, we can determine which terms are similar to each other.

Semantic similarities between sentences: Semantic similarity of two sentences is well discussed in [13]. To calculate this parameter, we first calculate the vector value properties for each sentence, then the maximum value of the semantic similarity between the words in the vector properties. The words in the same sentence are used to determine the words weight. Only the word similarity is used as the words weight. Further, the two words that are in the set of a part-of-speech are compared. Here, to obtain the semantic similarity between sentences, we use the approach that is presented in [14]. Similarity of two sentences is calculated according to Equation (3).

$$\mathrm{Sim}_{\mathrm{semantic}}(S1,S2)=\frac{\sum w1\epsilon s1 \; and \; w2\epsilon s2 \; MaxSim(w1,w2)}{|s1|+|s2|} \qquad (3)$$

The value of semantic similarity of each word in the first sentence with the all words that belong to the same part-of-speech in the second sentence is calculated and we retain the words that achieve maximum similarity. Then the summation of all maximum similarities is calculated and divided by the summation of the lengths of the two sentences and finally gets normalized.

The process of the calculation of semantic similarity between two words is as follows. Here, similar to [15], semantic similarity between two words is computing using Equation (4) in our approach.

$$\mathrm{WordSim}_{(x1,x2)}=\frac{2*N3}{N1+N2+2*N3} \qquad (4)$$

Where N1 is the number of the links with the father of sentence x1 until receiving the first common father with sentence x2. And N2 is also the number of links of x2's fathers until receiving the first common father of x1 and N3 is also the number of links with the common father to the root. For more details, interested readers are referred to [15].

The Order of Words: The composition of the sentence words plays an important role in the concepts of the sentence. This results in that the order of the words in the sentences is also important. For example, consider the sentences {**the sale manager hits the office worker**} and {**the office manager hits the sale worker**}. Indeed, these two sentences contain the same set of words but different sequences of words lead to different meanings. Recognition of the sequence of the words in a sentence is easy to the human but it is very challenging in natural language processing. Here, we use the method that has been proposed in [16]. In order to calculate the similarity of sequence of the words between the 2 sentences, at first we create the Join Set through the 2 sentences. First the words of the shorter sentence and then the words of the larger one are inserted into the Join Set. For example, the join set of the previous two sentences is {**the sale manager hits the office worker**}. Then the sequence vector of each of the two sentences is calculated. For example, for each word in the set of Join Set, if the same word in the same sentences exists in the corresponding place of that word in the sequence vector, the evidence index of that word in the sentence will be used. If that

word does not exist, the index of the similar word of that word that is larger than the threshold value will be used. For all other cases, the value of that index is 0.

After obtaining the sequence vectors of the two sentences r1 and r2, by using Equation (5) the difference value between the two vectors can be calculated.

$$\text{Sim}_{\text{wordOrder}}(S1,S2)=1-\frac{\|r1-r2\|}{\|r1+r2\|} \quad (5)$$

The value is also normalized. One special case is that the word that is in the Join Set does not exist in the corresponding indices. But by using corpus-based knowledge bases such as Wordnet the fact is that, we can use the method proposed in [17] to achieve the computation of similarity on top of word orders. In [17], the LCS[1] is used to determine the similarity between two words. A method for normalization is dividing the length of LCS into the longer string but because small strings are not used, it is not a suitable solution. This is because in some cases the small strings are very important. Therefore, we use Equation (6) for computing the normalization.

$$\text{NLCS}_{(r,s)}=\frac{Lenght(LCS(r,s))^2}{Lenght(r)*Lenght(s)} \quad (6)$$

The LCS method is case-sensitive but because of converting all letters into lowercase in the preprocessing phase, the computation is ensured to be valid.

C. Combination of Parameters

In this paper, after calculating the semantic similarity between the sentences and the similarity of word sequences in the sentences, the overall semantic similarity of the two sentences can be obtained via Equation (7).

$$\text{TotalSim}_{\text{semantic}}(S1,S2)=a*\text{Sim}_{\text{semantic}}(S1,S2)+(1-a)\ \text{Sim}_{\text{WordOrder}}(S1,S2) \quad (7)$$

In [18] and [19] it is proven that the combination of parameters is greater than the syntax parameter, and thus is the best evaluation of sentence similarity. The parameter *a* in our experiments takes 0.8 in the above equation.

D. Fuzzy Logic

A lot of work has been done on summarization based on syntax parameters. Also the several semantic parameters are discussed in several papers. Also in [20] [7] [6], it has been shown that the use of fuzzy logic achieves positive results in improving the summarization results. But our goal in this paper is to combine these cases and according to our knowledge, there is not any similar work in this area. The results also show that combining both syntactic and semantic parameters, and using the fuzzy logic can improve the results. That indicates the quality of our hybrid approach.

As classic logic is the base of the logic-based expert systems, fuzzy logic is also the base of the fuzzy expert systems. In fact, in facing the uncertainty when we use the fuzzy logic, modeling processes that use such techniques are more favored compared with conventional logic.

One of the problems of conventional logic is in the constraints of both values: True or False. However, conventional logic is well suited for a two-state system, but it is not suitable for systems that are not certain. The most important disadvantage of the conventional logic in a summarization process is that some parameters that are mentioned above are not two-state. As an example, to say that two sentences are semantically similar, or not, is not correct but in fuzzy logic, we can say that how much the two sentences are similar to each other. In our work, fuzzy logic is used for measuring the degree of importance and correlation and also used to identify the important sentence to create summarization.

IV. Evaluations

In this paper, we verify our approach using a real-world dataset. In proceedings of JAIR, 50 randomly articles was selected. The average number of words in an article is 436 words. First, all tables, images and formulas are eliminated in the preprocessing phase. The evaluation of summarization, in both internal and external evaluation, is carried out. In internal evaluation the focus is on the quality of the summarization, while in the external evaluation, most of focus is on the performance of the system in a particular problem. In fact, there is not consensus on which method is better or is preferred. But since the most recent work has been on internal evaluation and because of requirement of fair comparisons, in this paper we use the internal evaluation.

The quality of evaluation in this paper that is summarizing by the machine is compared with the summarizing of the proceeding JAIR issues that is summarized by human. Our assessment is also based on the F[2] scale, which uses the values of scales P (Precision) and R (Recall). The general formula for P, R and also F that we use, are shown in Equations (8) to (10).

$$P=\frac{(number\ of\ correct\ sentences\ that\ extracted\ by\ summarizer\ system)}{(Total\ number\ of\ sentences\ that\ extracted\ by\ summarizer\ system)} \quad (8)$$

$$P=\frac{(number\ of\ correct\ sentences\ that\ extracted\ by\ summarizer\ system)}{(Total\ number\ of\ sentences\ that\ extracted\ by\ summarizer\ system)} \quad (9)$$

$$F=\frac{2*P*R}{P+R} \quad (10)$$

Here, P measures the accuracy of the measure (Precision) and R is Recall.

Table 1 Comparing the results of summarizing methods

	MS Word			Copernic			Presented Method		
	Avg R	Avg P	Avg F	Avg R	Avg P	Avg F	Avg R	Avg P	Avg F
JAIR	0.28	0.31	0.29	0.46	0.55	0.51	0.58	0.6	0.59

[1] Longest Common Subsequence

[2] Fitness

Table 2: Comparison of proposed method with quality Huang

	Huang Method	Presented Method
	Avg F	Avg F
JAIR	0.46	0.58

Initially, measure F is calculated separately for each summarized essay then the average of F is obtained for all sets of the essay. The results of the summarizing method presented in this paper are shown in the Table 1. As can be seen the presented summarizing method shows better quality in comparison with other summarizing methods.

Moreover, comparisons with the mechanical summarizer that is presented in [21] are also performed on this set of essays. The comparison is shown in Table 2. We call the proposed method in [21] in Table 2 as the Huang Method. As can be seen in Table 2 the proposed summarizing method works better than the existing method, and shows better quality, although since the summarization process is not certain and is related to individual interest of the linguistic experts, the summarizing quality is not close to 100% correctness.

Overall, the proposed method shows better quality in comparison with existing summarizing methods.

V. CONCLUSIONS

The perception of a context from a computational perspective is still an unsolved problem. Many existing summarizing methods are mainly based on statistical and probabilistic extractions. These methods have no perception from the concept and meanings of the context. Summarizing most of the texts needs deeper understanding of the concepts to be meaningful.

We observe that attention to the concept and the meaning of the context could greatly improve summarization accuracy, and due to the uncertainty that exists in the summarization methods, we simulate human like methods by integrating fuzzy logic with traditional statistical approaches in this study. The results of this study indicate that our approach can deal with uncertainty and achieve better results when compared with existing methods. Our proposed method has been demonstrated to be able to provide better summarizing quality in comparison with existing methods.

Acknowledgement

The authors wish to thanks from deputy of research of Islamic Azad University Science and Research Branch for its support of this project and Dr Mohammad Mehdi Esnaashari from ITRC.

REFERENCES

[1] Mani, Inderjeet. Automatic summarization. Vol. 3. John Benjamins Publishing Company, 2001.

[2] E. Liddy, "Advances in automatic text summarization," Information Retrieval, vol. 4, no. 1, pp. 82–83, 2001.

[3] P. P. Balage Filho, T. A. Salgueiro Pardo, and M. das Gracas Volpe Nunes, "Summarizing Scientific Texts: Experiments with Extractive Summarizers," in Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on, 2007, pp. 520–524.

[4] J. Steinberger and K. Ježek, "Update summarization based on latent semantic analysis," in Text, Speech and Dialogue, 2009, pp. 77–84.

[5] L. A. Zadeh, "From circuit theory to system theory," Proceedings of the IRE, vol. 50, no. 5, pp. 856–865, 1962.

[6] F. Kyoomarsi, H. Khosravi, E. Eslami, P. K. Dehkordy, and A. Tajoddin, "Optimizing Text Summarization Based on Fuzzy Logic," Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008), pp. 347–352, May 2008.

[7] a. Kiani-B and M. R. Akbarzadeh-T, "Automatic Text Summarization Using Hybrid Fuzzy GA-GP," 2006 IEEE International Conference on Fuzzy Systems, pp. 977–983, 2006.

[8] F. Kyoomarsi, H. Khosravi, E. Eslami, and P. Khosravyan, "Optimizing Machine Learning Approach Based on Fuzzy Logic in Text Summarization," International Journal of Hybrid Information Technology, vol. 2, 2009.

[9] P. Achananuparp, X. Hu, and X. Shen, "The evaluation of sentence similarity measures," Data Warehousing and Knowledge Discovery, pp. 305–316, 2008.

[10] M. M. Stark and R. F. Riesenfeld, "Wordnet: An electronic lexical database," in Proceedings of 11th Eurographics Workshop on Rendering, 1998.

[11] L. Suanmali, M. S. Binwahlan, and N. Salim, "Sentence Features Fusion for Text Summarization Using Fuzzy Logic," 2009 Ninth International Conference on Hybrid Intelligent Systems, pp. 142–146, 2009.

[12] S. Fisher and B. Roark, "Query-focused summarization by supervised sentence ranking and skewed word distributions," in Proceedings of the Document Understanding Conference, DUC-2006, New York, USA, 2006.

[13] Mojtaba Sedigh Fazli, Jean-Fabrice Lebraty. A solution for forecasting pet chips prices for both short-term and long-term price forcasting, using genetic programming. WorldComp'2013. The 2013 International Conference on Artificial Intelligence, Jul 2013, Las Vegas, Nevada, United States. CSREA Press, II, pp.631-637. <hal-00859457>.

[14] R. Malik, L. V Subramaniam, and S. Kaushik, "Automatically selecting answer templates to respond to customer emails," in Proceedings of the 20th international joint conference on Artifical intelligence, 2007, pp. 1659–1664.

[15] Mojtaba Sedigh Fazli, Jean-Fabrice Lebraty. A comparative study on forecasting polyester chips prices for 15 days, using different hybrid intelligent systems. IEEE & International Neural Network Society. International Joint Conference on Neural Networks, Aug 2013, Dallas, Texas, United States. pp.1869-1875, 2013. <hal-00859445>

[16] Manaf Sharifzadeh, Saeed Aragy, Kaveh Bashash, Shahram Bashokian, and Mehdi Gheisari. A Comparison with two semantic sensor data storages in total data transmission. SoftConf, March 2013, Sanfransico

[17] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," ACM Transactions on Knowledge Discovery from Data, vol. 2, no. 2, pp. 1–25, Jul. 2008.

[18] T. K. Landauer, D. Laham, B. Rehder, and M. E. Schreiner, "How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans," in Proceedings of the 19th annual meeting of the Cognitive Science Society, 1997, pp. 412–417.

[19] P. Achananuparp, X. Hu, X. Zhou, and X. Zhang, "Utilizing sentence similarity and question type similarity to response to similar questions in knowledge-sharing community," in 17th international conference on World Wide Web, 2008.

[20]Mehdi Gheisari, Ali Akbar Movassagh, Yongrui Qin, Jianming Yong, Xiaohui Tao, Ji Zhang, Haifeng Shen, "NSSSD: A New Semantic Hierarchical Storage for Sensor Data", IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2016), May 4-6, 2016, Nanchang, China

[21] Huang, Hsun-Hui, Yau-Hwang Kuo, and Horng-Chang Yang. "Fuzzy-rough set aided sentence extraction summarization." Innovative Computing, Information and Control, 2006. ICICIC'06. First International Conference on. Vol. 1. IEEE, 2006.

[22] Porter, Martin. "The Porter stemming algorithm, 2005." See http://www. tartarus. org/~ martin/PorterStemmer.

[23] M.Gheisari," Design, Implementation, and Evaluation of SemHD: A New Semantic Hierarchical Sensor Data Storage", Indian J. Innovations Dev., Vol. 1, No. 3 (Mar 2012) ISSN 2277 – 5390