INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 2)

Available online at: www.ijariit.com

An automatic text summarization using simplified Lesk Algorithm

Vaidehi Tare
tare.vaidehi1@gmail.com
Shivajirao S Jondhale College of Engineering, Mumbai, Maharashtra

Priya Shukla
shuklap2796@gmail.com
Shivajirao S Jondhale College of Engineering, Mumbai, Maharashtra

Vaibhavi Vichare
vaibhavivichare23@gmail.com
Shivajirao S Jondhale College of Engineering, Mumbai, Maharashtra

Saroja T. V
sarojatv2005@gmail.com
Shivajirao S Jondhale College of Engineering, Mumbai, Maharashtra

ABSTRACT

Text Summarization is the method by which the noteworthy segments of a text are recovered. Diverse philosophies are created till now relying on a few parameters to discover the summary based on the position, configuration and sort of the sentences in an information content, organizations of various words, and recurrence of a specific word in content and so on. As indicated by various information sources, these predefined limitations enormously influence the outcome. The proposed approach generates the outline by undertaking an unsupervised learning method. The significance of a sentence in an information content is assessed by the assistance of Simplified Lesk Algorithm. As an online semantic dictionary WordNet is utilized. To start with, this approach assesses the weights of the considerable number of sentences of a content independently utilizing the Simplified Lesk Algorithm's calculation and organizes them in diminishing request as indicated by their weights. Next, depending upon the given level of the synopsis, a specific number of sentences are chosen from that requested rundown. The proposed approach gives best outcomes up to half outline of the original content.

Keywords: *Automatic Text Summarization, Extract, Abstract, Lesk algorithm, WordNet.*

1. INTRODUCTION

Automatic Text Summarization [1-2] is the method by which the consolidated data of a content is recovered. As the volume of electronic data is expanding step by step, it turns into a period and space expending matter to handle such gigantic volume of information.

In various extraordinary Natural Language Processing applications like Information Retrieval [3], Question Replying, Text Comprehension and so on, Automatic Text Summarization plays an inescapable part by delivering pertinent as well as particular data from a lot of information. For the most part, two kinds of summaries are generated from a content. The initial one is Extract [4-5], where the parts (words, sentences and so on.) of the content are reused and the second one is Abstract[7-8], where the summarized portion are regenerated. The majority of the calculations discover the concentrates by some hand labeled principles, as the position [6] of a sentence in a content, the arrangement of words (bold, italic and so forth.) in a sentence, recurrence of a word in a content and so forth. However, the exhibitions of these methodologies are extraordinarily influenced when input sources vary. These methodologies infer the significance of a data in a content by the example and position of that data in that content, instead of its semantic importance. The proposed approach recovers the significant data from the text by performing a semantic investigation of the original text. In this approach Simplified Lesk Algorithm [9-11] is utilized to separate the significant sentences from the content by performing semantic analysis on the sentences and WordNet [10-12] is utilized as an online semantic lexicon.

The rest of the paper is organized as Section 2 is about the Theoretical Background of the proposed approach; Section 3 describes the Proposed Approach; Section 4 depicts Experimental Results along with comparison; Section 5 represents Conclusion of the paper.

2. THEORETICAL BACKGROUND

In the proposed approach, the importance of a sentence in the content is extricated from its semantic analysis. Lesk calculation manages the semantic analysis of a word utilizing an online semantic dictionary WordNet. In this dictionary, words are arranged semantically unlike in other dictionaries where words are arranged alphabetically. The proposed approach suggests an improvement on Lesk algorithm to manage the semantic analysis of a word regarding the content it has been placed in.

2.1 Introduction to Lesk Algorithm

The Typical Lesk approach stresses on finding the actual meaning of a word in a specific context in order to remove the ambiguity between words. The Lesk calculation finds the real sense of that uncertain word by the following way:

To begin with, it chooses a short portion of a sentence containing an ambiguous (equivocal) word. Then, dictionary definition (gloss) of each of the senses of the ambiguous word is compared with glosses of the other words in that particular phrase. An equivocal word is allotted in that specific sense, whose gloss has a most astounding recurrence (number of words in like manner) with the gloss of different words of the expression.

2.2 Simplified Lesk Approach

The proposed approach holds the normal Lesk approach while suggesting a change for finding the significance of a sentence in a content.

The proposed approach embraces the normal Lesk approach also, suggests a change for finding the significance of a sentence in a content.

Next, the lexicon definitions (glosses) of all these significant words are considered and intersection activity is performed between every one of these glosses and the content itself as opposed to the glosses of alternate words.

Add up to a number of cover for each sentence represents the weight of the sentence in the text. These weights represent the significance of the sentences in the text, which go about as a key factor in Summarization process.

3. PROPOSED APPROACH

In the proposed approach, a single record input text is summarized by the given level of percentage of summarization utilizing an unsupervised learning method. To begin with, the Simplified Lesk Approach processes every single sentence to discover the weight of each sentence.

Next, the sentences with derived weights are organized in descending order according to their weights. Presently, as indicated by a particular level of percentage of summarization, certain numbers of sentences are chosen to be included in the summary.

In conclusion, the chosen sentences are rearranged by their original sequence in the information content.

3.1 Algorithm

The proposed approach compresses a text without depending upon the organization of the text and the position of a sentence in the text, as opposed to the semantic data lying in the sentence. This approach is language independent. To separate the semantic data from a sentence, just a semantic lexicon in that language is required.

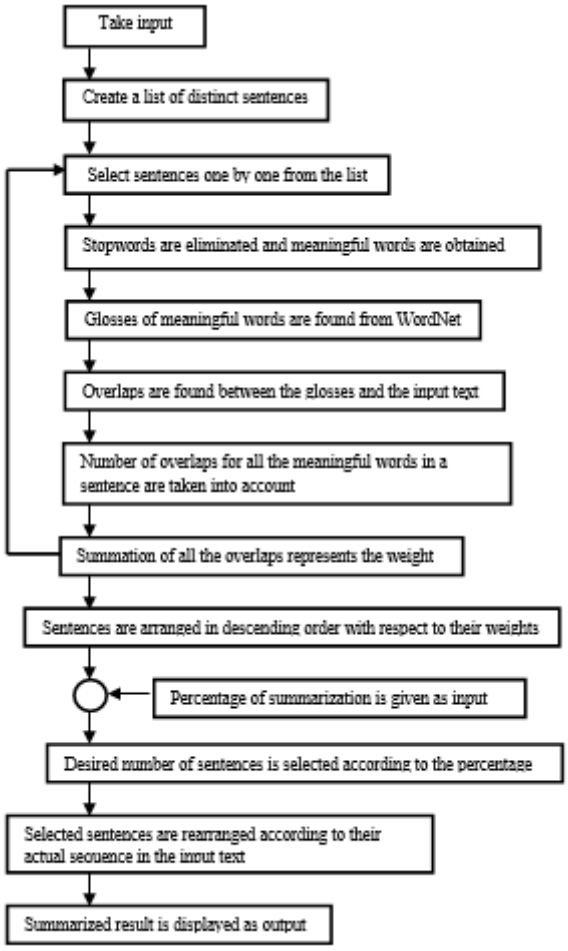

Fig 1: Flow of system

This algorithm evaluates the weights of the sentences of a text using Simplified Lesk algorithm and WordNet (refer Figure 2). Time Complexity of the algorithm is O (n^3), as finding the total number of overlaps between a particular sentence and the gloss is of O (n^2) complexity and this procedure is performed for all the n number of sentences.

Input: Input text
Output: Final summarized result

The lists of distinct sentences are created. Then each sentence is considered one by one. For each sentence stop, words are removed from the sentence as they do not participate directly in sense evaluation procedure. Glosses of all the meaningful words are extracted using the WordNet. The intersection is performed between the glosses and the input text itself. Summation of all the intersection results represents the weight of the sentence. Weight assigned sentences are arranged in descending order with respect to their weights. Desired numbers of sentences are selected according to the percentage of summarization. Selected sentences are re-arranged according to their actual sequence in the input text.

4. OUTPUT AND DISCUSSION

The output of this project is based on calculations that are derived by trying this approach on a large number of writings. The writings consist of various categories such as diverse specialized reports, distinctive news paper articles, diverse travel portrayals, short stories, etc. Text with different lengths and number of sentences are taken to see the efficiency of the algorithm in different cases and all are in the English language, as the semantic dictionary WordNet, used here, is in English.

We evaluate the correctness of the summary generated by comparing manual summary with the text summarized by the system. This evaluation is based on frequently used parameters which are: - Accuracy (A), Evoke (E) and F- Measure (F-M). The parameters are calculated in the following way:

Accuracy (A) = C / (C + W)

Evoke (E) = C / (C + M)

F-Measure (F-M) =2*A*E/ (A+E)

Where,

C (correct) = the number of sentences extracted by the system and the human;

W (wrong) = the number of sentences extracted by the system but not by the human;

M (missed) = the number of sentences extracted by the human but not by the system.

We consider 8 sample texts and their corresponding results are in such a manner:

Table-1: Performance measurement of the algorithm on sample text

Text	C	W	M	A	E	F-M
1	18	3	3	0.8571	0.8571	0.8571
2	34	5	5	0.8717	0.8717	0.8717
3	15	3	3	0.8333	0.8333	0.8333
4	29	4	4	0.8787	0.8787	0.8787
5	26	6	6	0.8125	0.8125	0.8125
6	33	5	5	0.8684	0.8684	0.8684
7	16	3	3	0.8421	0.8421	0.8421
8	22	4	3	0.8461	0.88	0.8627

In the above test, all the texts are summarized to 50% of their originals by both, human and system. So, the numbers of sentences in the summarized text for both the cases (system and person) are exactly same. For this reason in the above table, the majority "W" and the "M" columns show the same results.

It is already tested that, the algorithm gives good results for large texts. It is also seen that the calculation gives an acceptable outcome at 25% rundown given that the significance of the sentences is calculated from their semantic data.

5. CONCLUSION AND FUTURE WORK

The proposed approach depends on the semantic data of the extracts in a text. In this way, other factors like formats, places of various units in the text are not considered. But in barely any cases, there are dominating numbers of named elements in a content.

In those cases, hybridization of the proposed approach with some particular rules with respect to Named Entity Recognition should give more successful outcomes.

It is simple, highly efficient and comparatively less complex coding than other coding. This application can be used on a professional level as well as educational level also.

6. ACKNOWLEDGEMENT

We wish to express our deep gratitude to our project guide Prof. Saroja T.V. and project coordinator Prof. Uttara Gogate in the Department of Computer Engineering for all the advice, encouragement and constant support she has given us throughout our project work. This work would not have been possible without her support and valuable suggestions.

We are grateful to Prof. P.R.Rodge, Head of the Department of Computer Engineering and members of Project Review Committee for their valuable suggestions.

We are also grateful to Dr. J.W.Bakal, Principal for giving us the necessary facilities to carry out our project work successfully. We would like to thank all our friends for their help and constructive criticism during our project.

7. REFERENCES

[1] H. Dalianis, "SweSum – A Text Summarizer for Swedish," Technical report TRITA-NA-P0015, IPLab-174, NADA, KTH, October 2000.
[2] M. Hassel, "Evaluation of Automatic Text Summarization: A practical implementation," Licentiate Thesis, University of Stockholm, 2004.
[3] G. Salton, "Automatic Text Processing: The Transformation Analysis and retrieval of Information by Computer", Addison Wesley Publishing Company, 1989.
[4] R. Barzilay, M. Elhadad, "Using lexical chains for text summarization,"In:Inderjeet Mani and Mark Maybury, editors, Advances in AutomaticText Summarization, MIT Presspp 111–122, 1999.
[5] C. Nobata, S. Sekine, H. Isahara, and R. Grishman, "SummarizationSystem Integrated with Named Entity Tagging and IE pattern Discovery,"Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002); Las Palmas, Canary Islands, Spain.
[6] C-Y. Lin and E. Hovy, "Identify Topics by Position," Proceedings of the 5th Conference on Applied Natural Language Processing, 1997.
[7] I. Mani, M. Maybury (Eds.),"Advances in Automatic Text Summarization," MIT Press, Cambridge, MA, 1999.
[8] H.P. Edmundson, "New methods in automatic abstracting," In: Journal of the Association for Computing Machinery 16 (2). pp. 264-285, 1969.Reprinted in: I. Mani, M.T. Maybury,"Advances in Automatic Text Summarization," Cambridge, Massachusetts: MIT Press. pp. 21-42.
[9] S. Banerjee, T. Pedersen,"An adapted Lesk algorithm for word sense disambiguation using WordNet," In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February 2002.

[10] H. Seo, H. Chung, H. Rim, S. H., Myaeng, S. Kim, "Unsupervised word sense disambiguation using WordNet relatives," Computer Speech and Language, Vol. 18, No. 3, pp. 253-273, 2004.
[11] Gaizauskas, "Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs," Computer Speech and Language, Vol. 12, No. 3, pp. 453-472, Special Issue on Evaluation of Speech and Language Technology.
[12] A. J. Cañas , A. Valerio, J. Lalinde-Pulido, M. Carvalho, M. Arguedas, "Using WordNet for Word Sense Disambiguation to Support Concept Map Construction," String Processing and Information Retrieval, pp. 350-359, 2003.

