# An Approach To Automatic Text Summarization Using Simplified Lesk Algorithm And Wordnet

Alok Ranjan Pal,[1] Projjwal Kumar Maiti[1] and Diganta Saha[2]

[1]Dept. of Computer Science and Engineering College of Engineering and Management, Kolaghat, India
[2]Dept. of Computer Science and Engineering, Jadavpur University Kolkata, India

***ABSTRACT***

*Text Summarization is a way to produce a text, which contains the significant portion of information of the original text(s). Different methodologies are developed till now depending upon several parameters to find the summary as the position, format and type of the sentences in an input text, formats of different words, frequency of a particular word in a text etc. But according to different languages and input sources, these parameters are varied. As result the performance of the algorithm is greatly affected. The proposed approach summarizes a text without depending upon those parameters. Here, the relevance of the sentences within the text is derived by Simplified Lesk algorithm and WordNet, an online dictionary. This approach is not only independent of the format of the text and position of a sentence in a text, as the sentences are arranged at first according to their relevance before the summarization process, the percentage of summarization can be varied according to needs. The proposed approach gives around 80% accurate results on 50% summarization of the original text with respect to the manually summarized result, performed on 50 different types and lengths of texts. We have achieved satisfactory results even upto 25% summarization of the original text.*

***KEYWORDS***

*Automatic Text Summarization, Extract, Abstract, Lesk algorithm & WordNet*

## 1.INTRODUCTION

The volume of electronic information available on Internet is increasing day by day. As a result, dealing with such huge volume of data is creating a big problem in different real life data handling applications. Automatic Text Summarization [1-12] is the procedure to infer a condensed information from a large volume of data. The Automatic Text Summarization task makes the job easier for different Natural Language Processing (NLP) applications, such as Information Retrieval[13], Question Answering or Text Comprehension etc. These application can save time and resources, having their actual input text in condensed form.

Several types of summaries can be inferred from a text. As, **Extracts** [14-16]are summaries created by reusing portions (words, sentences etc.) of the input text and **Abstracts**[17-21] are created by re-generating the extracted content.

Most of the research works base on finding the extracts from a given text depending on few hand tagged rules, as the position[22] of a sentence in a text, format of words (bold, italic etc.) in a sentence, frequency of a word in a text etc. But the drawback of this approach is, it greatly

depends on the format of the text. As a result, importance of a sentence bases on its format and position in the text rather than its semantic information.

In the proposed approach we have extracted the relevant sentences from a single-document text based on the semantic information of the sentence using Simplified Lesk Algorithm [23-25], as an unsupervised learning algorithm and WordNet [26-28], as an online semantic dictionary.

Organization of rest of the paper is as follows: 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 relevance of a sentence in the text, where it belongs is extracted from its semantic information. Lesk algorithm deals with the semantic information of a word using an online dictionary WordNet. In this dictionary, words are arranged semantically rather than alphabetically. The proposed approach implies a modification on Lesk algorithm to deal with the semantic information of a word with respect to the text, it belongs.

### 2.1 Preliminaries of Lesk algorithm

The Typical Lesk approach emphasizes on finding the actual sense of a single word in a particular context, where the word can have more than one senses. That type of word is called ambiguous word. The Lesk algorithm finds the actual sense of that ambiguous word by the following way:

First, it selects a short phrase from the sentence containing an ambiguous 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 ambiguous word is being assigned with that particular sense, whose gloss has highest frequency (number of words in common) with the glosses of other words of the phrase.

Example 1: “Ram and Sita everyday go to bank for withdrawal of money.”
Here, the phrase is taken depending on window size (number of consecutive words). If window size is 3, then the phrase would be “go bank withdrawal”. All other words are being discarded as “stop words”.

Consider, the glosses of all words presented in that particular phrase are as follows:
The number of senses of “Bank” is 2 such as ‘X’ and ‘Y’ (refer Table 1).
The number of senses of “Go” is 2 such as ‘A’ and ‘B’ (refer Table 2).
and the number of senses of “Withdrawal” is 2 such as ‘M’ and ‘N’ (refer Table 3).

| Key word | Probable sense |
|---|---|
| Bank | X |
| | Y |
| | |

Table 1. Probable Sense of “Bank”.

| Word | Probable sense |
|---|---|
| Go | A |
| | B |

Table 2. Probable Sense of "Go".

| Word | Probable sense |
|---|---|
| Withdrawal | M |
| | N |

Table 3. Probable Sense of "Withdrawal".

Consider the word "Bank" as a keyword. Number of common words is measured in between a pair of sentences.

| Pair of Sentences | Common number of Words |
|---|---|
| X and A | A' |
| X and B | B' |
| Y and A | A'' |
| Y and B | B'' |
| X and M | M' |
| X and N | N' |
| Y and M | M'' |
| Y and N | N'' |

Table 4. Comparison Chart between pair of sentences and common number of words within particular pair.

Table 4 shows all possibilities using sentences from Table 1, Table 2, Table 3, and number of words common in each possible pair.

Finally, two senses of the keyword "Bank" have their counter readings (refer Table 4) as follows:
X counter, $X_C$ = A' + B' + M' + N'.
Y counter, $Y_C$ = A" + B" + M" + N".

Therefore, higher counter value would be assigned as the sense of the keyword "Bank" in particular sentence. This strategy believes that surrounding words have same senses as of the keyword.

## 2.2 Simplified Lesk approach

The proposed approach adopts the typical Lesk approach and implies a modification for finding the importance of a sentence in a text.

Here, sentences are picked up one by one from the text.

Then, after discarding the stop words(as they don't participate directly in sense disambiguation) from the sentence, only the meaningful words are considered for further operation.
Next, the dictionary definitions(glosses) of all these meaningful words are considered and intersection operation is performed between each of these glosses and the text itself rather than the glosses of the other words.

Total number of overlap for each sentence represents the weight of the sentence in the text. These weights represent the importance of the sentences in the text, which act as a key factor in summarization process.

## 3. PROPOSED APPROACH

In the proposed approach, a single-document input text is summarized according to the given percentage of summarization using unsupervised learning. First, the Simplified Lesk Approach is applied to each of the sentences to find the weight of each sentence (refer section 2.2). Next, the sentences with derived weights are arranged in descending order with respect to their weights. Now, according to a specific percentage of summarization at a particular instance, certain numbers of sentences are selected as a summary.

Lastly, the selected sentences are rearranged according to their original sequence in the input text.

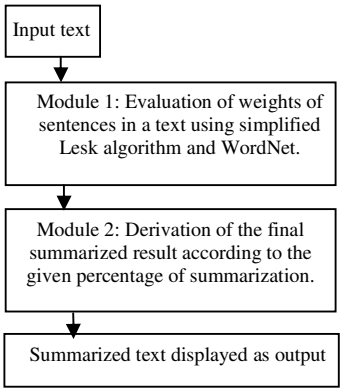


Figure 1. Modular representation of the overall approach.

**Algorithm 1:** This algorithm summarizes a single-document text using unsupervised learning approach (refer Figure 1). In Module 1, the weight of each sentence in a text is derived using Simplified Lesk algorithm and WordNet. In Module 2, the summarization process is performed according to the given percentage of summarization.
Input: Single-document input text.

Output: Summarized text.

Step 1: Input text is passed to Module 1, where the weights of each of the sentences of the text are derived using Simplified Lesk Algorithm and WordNet. In this module, the semantic analysis of the extracts are performed.

Step 2: Weight assigned sentences are passed to Module 2, where the final summarized result is evaluated and displayed.
Step 3: Stop.

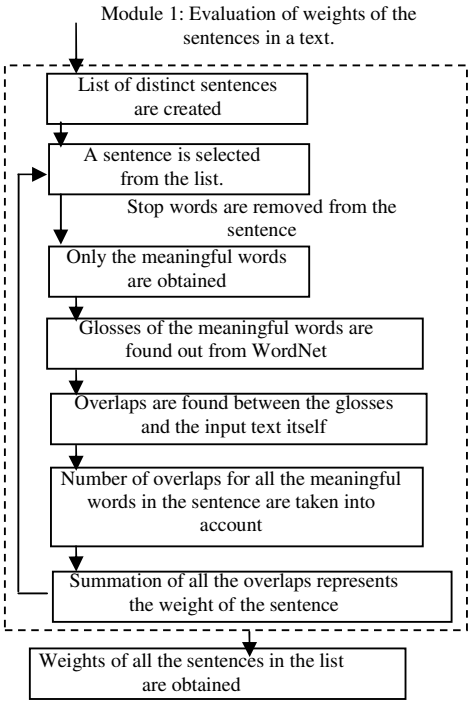


Figure 2. Evaluation of weights of sentences in a text using Simplified Lesk and WordNet.

**Module 1: Algorithm 2:** 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: Sentences of the text with assigned weight to each of it.
Step 1: The list of distinct sentences of the text is prepared.
Step 2: Repeat steps 3 to 7 for each of the sentences.
Step 3: A sentence is picked up from the list.
Step 4: Stop words are removed from the sentence as they do not participate directly in sense evaluation procedure.
Step 5: Glosses(dictionary definitions) of all the meaningful words are extracted using the WordNet.
Step 6: Intersection is performed between the glosses and the input text itself.
Step 7: Summation of all the intersection results represents the weight of the sentence.
Step 8: Stop.

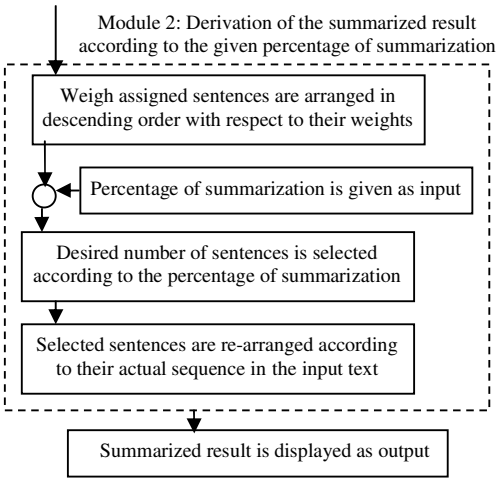


Figure 3. Derivation of the final summarized result according to the given percentage of summarization.

**Module 2: Algorithm 3:** This algorithm evaluates the final summarized result and displays (refer Figure 3). Time Complexity of the algorithm is $O(n^2)$, which is evaluated at the time of arranging the sentences.

Input: a) List of sentences of the input text with evaluated weights.
b) Percentage of summarization.

Output: Final summarized result.

Step 1: Weight assigned sentences are arranged in descending order with respect to their weights.
Step 2: Desired number of sentences are selected according to the percentage of summarization.
Step 3: Selected sentences are re-arranged according to their actual sequence in the input text.
Step 4: Stop.

The proposed approach summarizes a text without depending on the format of the text and the position of a sentence in the text, rather than the semantic information lying in the sentence. In addition to, this approach is language independent. To extract the semantic information from a sentence, only a semantic dictionary in that language is needed.

## 4. OUTPUT AND DISCUSSION

This algorithm is tested on total number of fifty texts. The texts are of five categories, where each category contains ten number of texts. The categories are legend personalities, such as sacred soul, writer, patriot, singer and sports personalities, different technical reports, different news paper articles on sports, politics, different travel narrations and short stories.

The length of the texts are taken different to see the efficiency of algorithm in different cases and all are in English language, as the semantic dictionary WordNet, used here, is in English. First, the texts are summarized by an expert person. At the same time, the texts are summarized by the system. Then the two results are compared using the mostly used parameters- Precession(P), Recall(R) and F-Measure(F). The parameters are expressed in the following way:

Precision(P) = correct / (correct + wrong),
Recall(R) = correct / (correct + missed),

F-Measure(F-M)=2*P*R/(P+R).
where,

correct = the number of sentences extracted by the system and the human;
wrong = the number of sentences extracted by the system but not by the human;
missed = the number of sentences extracted by the human but not by the system.
For example, 10 sample texts related to the legend personalities and their corresponding results are given below:

| Text | Correct | Wrong | Missed | P | R | F-M |
|---|---|---|---|---|---|---|
| Text 1 | 16 | 2 | 2 | 0.8889 | 0.8889 | 0.8889 |
| Text 2 | 36 | 5 | 5 | 0.8780 | 0.8780 | 0.8780 |
| Text 3 | 16 | 2 | 2 | 0.8889 | 0.8889 | 0.8889 |
| Text 4 | 16 | 2 | 2 | 0.8889 | 0.8889 | 0.8889 |
| Text 5 | 22 | 4 | 4 | 0.8461 | 0.8461 | 0.8461 |
| Text 6 | 26 | 6 | 6 | 0.8125 | 0.8125 | 0.8125 |
| Text 7 | 28 | 5 | 5 | 0.8484 | 0.8484 | 0.8485 |
| Text 8 | 32 | 6 | 6 | 0.8421 | 0.8421 | 0.8421 |
| Text 9 | 14 | 2 | 2 | 0.875 | 0.875 | 0.875 |
| Text 10 | 20 | 4 | 4 | 0.8333 | 0.8333 | 0.8333 |

Table 5: Performance measurement of the algorithm on sample texts.

In the above test, all the texts are summarized to 50% of their originals by the system as well as the expert person. So, the number of sentences in the summarized text for both the cases(system and person) are exactly same. For this reason in Table 5 the "Wrong" and the "Missed" columns show the same results.

*Kaili Müürisep, Pilleriin Mutso: ESTSUM - Estonian newspaper texts summarizer(2005)*[29]presented their test result as 60% average overlapping with handmade summaries using rule based approach. This unsupervised approach performs better (refer Table 5) compared to the rule based approach.

It is already tested that, the algorithm gives good results for large texts(more than 60 sentences), as well as for small texts(less than 20 sentences). It is also tested that the algorithm gives a satisfactory result at 25% summarization because the relevance of the sentences are derived from their semantic information.

It is observed that the proposed approach gives very good results for the technical reports as that type of texts contain a less number of named entity, because the less number of named entity in a sentence increases the number of meaningful words in the sentence. As much as the number of meaningful words in a sentence is increased, more number of glosses are obtained to be intersected with the text. As a result, the weight of a sentence is evaluated more effectively.

## 5. CONCLUSION AND FUTURE WORK

The proposed approach is based on the semantic information of the extracts in a text. So, different parameters like formats, positions of different units in the text are not taken into account. But in few cases, there are dominating numbers of named entities in a text. In those cases, hybridization

of the proposed approach with some specific rules regarding Named Entity Recognition should give more effective results.

## REFERENCES


[1] H. Dalianis, "SweSum – A Text Summarizer for Swedish," Technical report TRITA-NA-P0015, IPLab-174, NADA, KTH, October 2000.

[2] M. Hassel,"Resource Lean and Portable Automatic Text Summarization. PhD thesis, Department of Numerical Analysis and Computer Science," Royal Institute of Technology, Stockholm, Sweden 2007.

[3] K. Spärck Jones,"Automatic summarising: The state of the art," Information Processing & Management 43(6), , pp 1449–1481, 2007.

[4] R. Barzilay, M . Elhadad, "Using lexical chains for text summarization," In: Inderjeet Mani and Mark Maybury, editors, Advances in Automatic Text Summarization, MIT Press , , pp 111–122, 1999.

[5] M. Hassel, "Exploitation of Named Entities in Automatic Text Summarization for Swedish," In the Proceedings of NODALIDA '03 - 14th Nordic Conference on Computational Linguistics, May 30-31 2003, Reykjavik, Iceland.

[6] C. Nobata, S. Sekine, H. Isahara and R. Grishman, "Summarization System 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.

[7] I. Mani, M. Maybury (Eds.),"Advances in Automatic Text Summarization," MIT Press, Cambridge, MA, 1999.

[8] I. Mani," Automatic text summarization," John Benjamins, 2001.

[9] I. Mani, and M.T. Maybury, (eds), "Advances in Automatic Text Summarization," Cambridge, MA: MIT Press, 1999.

[10] H. Dalianis, and E. Åström,"SweNam-A Swedish Named Entity recognizer. Its construction, training and evaluation," Technical report TRITA-NAP0113, IPLab-189, NADA, KTH, June 2001.

[11] M. Hassel,"Exploitation of Named Entities in Automatic Text Summarization for Swedish," Proceedings of NoDaLiDa ´03, May 2003.

[12] M. Hassel, "Evaluation of Automatic Text Summarization: A practical implementation," Licentiate Thesis, University of Stockholm, 2004.

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

[14] H.P. Edmundson,"New methods in automatic extracting," In: Inderjeet Mani and Mark Maybury, editors, Advances in Automatic Text Summarization, MIT Press, pp 23–42,1969.

[15] H.P. Edmundson, "New Methods in Automatic Extraction," Journal of the ACM 16(2), pp 264-285, 1969.

[16] R. Mihalcea,"Graph-based ranking algorithms for sentence extraction, applied to text summarization," In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 2004.

[17] H.P. Luhn,"The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development pp 159-165, 1959.

[18] H.P. Luhn, "The automatic creation of literature abstracts," In: IRE National Convention, pp. 60-68. Also in: IBM J. Res. Dev., vol. 2, p. 159, April 1958.

[19] H.P. Luhn,"The automatic creation of literature abstracts," In: Inderjeet Mani and Mark Maybury, editors, Advances in Automatic Text Summarization, MIT Press , pp 15–22,1958.

[20] L.A. Ramshaw, M.P. Marcus,"Text chunking using transformation-based learning," In: Proceedings of the Third ACL Workshop on Very Large Corpora, Cambridge MA, USA. 1995.

[21] H.P. Edmundson, "New methods in automatic abstracting," In: Journal of the Association for Computing Machinery 16 (2). 264-285, 1969. Reprinted in: I. Mani, M.T. Maybury,"Advances in Automatic Text Summarization," Cambridge, Massachusetts: MIT Press. 21-42

[22] C-Y. Lin and E. Hovy, "Identify Topics by Position," Proceedings of the 5th Conference on Applied Natural Language Processing, 1997.

[23] 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.

[24] M. Lesk,"Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," Proceedings of SIGDOC, 1986.

[25] 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.
[26] 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.
[27] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, "WordNet An on-line lexical database," International Journal of Lexicography, Vol. 3, No. 4, pp. 235-244, 1990.
[28] 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.
[29] Kaili Müürisep, Pilleriin Mutso, "ESTSUM - Estonian newspaper texts summarizer," Proceedings of The Second Baltic Conference on Human Language Technologies. April 4-5, 2005. Tallinn

## Authors

Alok Ranjan Pal has been working as an a Assistant Professor in Computer Science and Engineering Department of College of Engineering and Management, Kolag hat since 2006. He has com pleted his Bachelor's and Master's degree under WBUT. Now, he is working on Natural Language Processing.

Mr. Projjwal Kumar Maiti is a student of Computer Science and Engineering Department of College of Engineering and Management, Kolaghat. His field of interest is AI, Soft Computing and NLP.

Dr. Diganta Saha is an Associate Professor in Department of Computer Science & Engineering, Jadavp ur University. His field of specialization is Machine Translation/ Natural Language Processing/ Mobile Computing/ Pattern Classification.