EXTRACTION BASED AUTOMATIC TEXT SUMMARY GENERATOR

Kritika Bakshi Isha Jain Tania

IT Department, ADGITM
Shastri Park, New Delhi 110053

Abstract. The paper lays emphasis on TextRank algorithm, a graph based approach used to tackle the automatic article summarization problem and proposing a variation to the similarity function used to compute scores during sentence extraction. The paper also emphasizes on the role of title of an article (if provided) in extracting an optimal, normalized score for each sentence.
Keywords: TextRank, Similarity, PageRank, Lexemes

1. Introduction

Ranking algorithms for undirected graph such as Google's PageRank algorithm [1] have been successfully able to establish their importance and use in social networks and especially the WWW (World Wide Web). Computing the values along the vertices and edges helps one to decide the path which may be an optimal solution to any query.

A similar approach is quite applicable in the field of Natural Language Processing wherein lexical or semantic graphs [8] have been used to extract useful and important phrases from the text available. One such prominent example is the Text Rank Algorithm [8], which is a text oriented ranking based method. The graph based TextRank algorithm is used to extract useful paraphrases and construct a useful and meaningful summary of the text/article available. The algorithm has been a center of research for a long time and has its own limitations too.

A multi document summarization model to reduce the redundancy is discussed in [11]. This model uses the statistical and linguistics for overcoming the information diversity problem. Paper [12] discussed the DBPedia for topic abstraction from clusters of online comments to news. Graph based technique for tweet summarization is used in paper [13]. K-mean clustering algorithm for extraction and text summarization is used in paper [14]. Text categorization for classifying a document in different categories is discussed in [15]. The authors have used KNN based machine learning model for this task. In paper [16] all the text summarization techniques have been discussed.

In this paper, we shall draw our focus towards the TextRank Algorithm and its limitations. Further, we shall present our modifications to the TextRank algorithm and factors (such as the title of an article) that can be incorporated while extracting a meaningful and coherent summary. The whole paper is divided into five sections. Section 1 being an introduction section and related work done in this direction. Section 2 details the textrank algorithm and how sentence is extracted. Section 3 emphasizes our proposed approach for automatic text summarization. Section 4 shows the implementation part and the results evaluated using our approach. Section 5 summarizes the whole work and direction to future work.

2. Literature Review

Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive techniques". In this paper author has explained the over-all idea of extractive summarization and all possible extractive summarization challenges in the paper. He have also explained about the features that are used to generate a summary and features that were earlier used and have described most of the extractive summarization techniques to solve the challenges and to obtain summary using these techniques. According to author "the importance of sentences is decided based on statistical and linguistic features of sentences"[2]. N.Moratanch and S.Chitrakala, "A Survey on Extractive Text Summarization". In this paper, author has described the word level features and sentence level features. In this paper author have categorized all extractive summarization methods into unsupervised and supervised methods and have explained each method and have depicted few evaluation metrics

Rajvardhan Oak, "Extractive Techniques for Automatic Document Summarization: A Survey". Author has described different extractive summarization methods and also describes a comparative study of different extractive summarization methods explaining each method's advantages and disadvantages. He have also explained two summarizer tools MEAD and summarist [4].

Selvani Deepthi Kavila and Dr.Radhika Y, " Extractive Text Summarization Using Modified Weighing and Sentence Symmetric Feature". In this paper author has mainly laid emphasis on summarization of different research papers of various fields. In this paper, three distinctive algorithms for summarization are shown and results are detected for each algorithm. Author has perceived that sentence score and feature scores used for the summarization process are determined on the basis of the statistical approaches. In this paper author has overcome few challenges like working with huge amount of data to summaries and including unnecessary sentences in the summary while using extractive methodologies by introducing compression ratio that will help to find out importance of each sentence [22].

Deepali K. Gaikwad and C. Namrata Mahender, " A Review Paper on Text Summarization". In this paper author has described both extractive summarization technique and abstractive summarization technique and have described text summarizers and summarization tools for Indian languages and have exhibited the comparison between performance of different methods [6].

Aysa Siddika Asa, Sumya Akter, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, Masud Ibn Afjal, "A Comprehensive Survey on Extractive Text Summarization Technique". The up-to-date extractive summarization techniques for different languages have been described in this paper. Here in this paper author have mainly laid emphasis to generate a summarization system for Bengali language using different types of features [7].

3. TextRank Algorithm

In this section, we will talk about the TextRank algorithm and its salient features. TextRank is an unsupervised machine learning algorithm. It is a type of Extraction based summarization [7], that is, it is used to extract relatively important sentences from each paragraph and arrangement of such sentences fo build a relevant summary. The application of TextRank [7] is found in both keyword extraction from a large pool of words and sentence extraction from a body of documents or a single document. It uses a graph based ranking approach wherein each sentence/word represents a node/vertices while the weighted edges represent the degree of similarity of between the vertices. The TextRank algorithm is an extension of the PageRank algorithm where the modified formula is used to calculate the cumulative score of each vertex representing a sentence. However, we propose to modify the similarity function and normalize the scores in order to produce better results. Major advantages of the TextRank [7] algorithm are as follows:

- It is unsupervised, therefore does not require any training set.
- No dependence on language.

The TextRank algorithm is based purely on the frequency of occurrence of words and does not require any prior knowledge of grammar. This eliminates the requirement of any particular tools dedicated to any particular languages. However, this may draw certain limitations to the algorithm particularly in cases of lexemes. Let G = (V, E)
be an undirected graph [8] consisting of set of vertices V and set of edges E (E V ×⊆V). For a given vertex V_i, let In (V_i) represent set of vertices pointing towards the former vertices and Out (V_i) represent the set of vertices that point to the next-inline vertices. The Score [1] of each vertex is calculated by the formula:

$$PR(Vi) = (1 - d) + d * X\,Vj \in In(Vi)\, PR(Vj)\,|Out(Vj)| \quad \textbf{(1)}$$

3.1 Sentence Extraction

Applying TextRank consists of building a graph associated to the pool of sentences/text available where the vertex represents an individual sentence which is to be ranked. All the sentences are ranked in the same way.
For applying TextRank to our text, we first need to build a graph associated [8] with the text, wherein the vertices depict the sentences to be ranked. In order to pick up relevant sentences, we need to rank all of them, and a vertex is created for each sentence in the text.
An edge is added in the graph between two sentences on the basis of the degree of similarity between them which is measured by the degree of words common between the sentences. An edge is connected between a pair of sentences that have common words. In order to avoid promoting long sentences [8], the formula uses a normalizing factor to divide the magnitude of overlapping content between two corresponding sentences. Let there be two sentences, Si and Sj [8], where a sentence is represented by Ni words that form a sentence:
$S_i = W_{k_1}, W_{k_2}, \ldots\ldots\ldots, W_{k_N}$, the similarity [8] between Si and Sj is calculated using the below mentioned formula:

$$Similarity\,(S_i, S_j) = \frac{|W_k | W_k \epsilon\, S_i \,\&\, W_k \epsilon\, S_j|}{log\,(|S_i|) + log\,(|S_j|)} \quad \text{(2)}$$

4. Proposed Approach

The title may be a name representing the subject in the article, or it may be used to describe a particular situation or description. It is believed that article titles are unique, that is, no two articles can have the same title. Articles titles can also add necessary distinguishing
information to elaborate the meaning of the same. The title of an article, if available, can further help us extract a more meaningful and precise sentence during the process of extraction based summarization of an article. We, therefore, propose to add title as another important factor in the process of article summarization. While we traditionally used to calculate the similarity between two sentences based on their degree of content overlap between them, the method can be employed while calculating the similarity between each individual sentence and the title of the article as well. The degree of similarity can, therefore be added which shall incorporate the importance of the title for an article as well, and hence, making sentence extraction more
meaningful and coherent. The modified similarity function for comparing two
sentences is given by:

$$Similarity_{sentences}\,(S_i, S_j) = \frac{|W_k | W_k \epsilon\, S_i \,\&\, W_k \epsilon\, S_j|}{(|S_i| + |S_j|\ \ )/2}$$

(2)
Similarly, the modified function for comparing each individual sentence is given by:

$$Similarity_{title}\,(S_i, S_{title}) = \frac{|W_k | W_k \epsilon\, S_i \,\&\, W_k \epsilon\, S_{title}|}{(|S_i| + |S_{title}|\ \ )/2} \quad (3)$$

Therefore the cumulative score of any sentence, S_i, say, S_1 is given by:

$$Similarity_{title}(S_1, S_{title}) + \{\sum_{i=1, j=2}^{j=N} Similarity_{sentences}(S_i, S_j)\} - Similarity_{sentences}(S_1, S_1) \quad (4)$$

5. Implementation & Results

Implementation is the developmental stage of the theoretical design [10]. At this stage, the project is turned into a working system. The total implementation of the project is divided into two important modules:
Module1: Uploading of input file which contains the article, processing on the input text file & calculation of scores

The file is uploaded using a dialog box and the title of the article is specified in the text input. Splitting of article into sentences and two stages of comparison take place:

- Comparison of each sentence with every other sentence.
- Comparison of each sentence with title.

We applied the modified sentence extraction formula on multiple articles for summarization task and evaluated the results. We took nearly 4 sample articles for article summarization and evaluated our results using ROUGE 2.0 Evaluation Technique. This method is found to be precisely related to human evaluation as it is based on Ngram statistics [9]. The Results are further mentioned in a table given below:

Table 1. Results of summary evaluation using ROUGE 2.0 Evaluation Toolkit. The summary is generated using the modified sentence extraction formula

Rouge Type	Task Name	Average Recall	Average Precision	Average FScore	Number Referenced Summaries
ROUGE 1	Sample 1	1.0	0.29664	0.45732	1
ROUGE 1	Sample 2	1.0	0.09125	0.16841	1
ROUGE 1	Sample 3	1.0	0.33504	0.50192	1
ROUGE 1	Sample 4	1.0	0.44071	0.61180	1

6. Conclusion & Future Scope

The paper introduces the TextRank Algorithm which is a Extractive Summarization technique based on undirected graphs. We also talked about the way in which sentences are extracted and

the importance of a title in an article summarization. We, further devised a formula for the same, and illustrated how our implementation really worked. Engaging title in the process of summarization of the article (if available) ensures consistency and coherence and that the best suitable candidate/sentence is extracted in accordance to the sense of the article. The results of four sample articles were computed using ROGUE 2.0 evaluation toolkit based on several parameters, as depicted in Table No. 1.

While efforts have been made to extract a meaningful and coherent summary from the article, there is still a lot of scope of improvement as to how the sentences are extracted and whether they take the summary to its logical meaning. Considering various other factors like personal pronouns, lexemes [6] can further ensure a meaningful, logical and coherent summary of an article.

References

1. S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7).
2. 2002 DUC Document Understanding Conference 2000. http://wwwnlpir.nist.gov/projects/duc/.
3. V.Gupta ,G. S. Lehal, "A Survey of Text Summarization Extractive techniques", Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, August 2010.
4. D. Sureshrao, S. Subhash and P. Dashore, Analysis of Query Dependent Summarization Using Clustering Techniques, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2,Issue 1.
5. E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. 1999. Domain specific keyphrase extraction. In Proceedings of the 16th International Joint Conference on Artificial Intelligence.
6. M. Halliday and R. Hasan. 1976. Cohesion in English. Longman Bartłomiej Balcerzak , Wojciech Jaworski and Adam Wierzbicki. 2014. Application of TextRank algorithm for credibility assessment. Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland.
7. Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the 42nd Annual Meeting of the Association for Computational Lingusitics (ACL 2004) (companion volume), Barcelona, Spain.
8. Kavita Ganesan and ChengXiang Zhai and Jiawei Han. Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions
9. Dipti.D.Pawar, M.S.Bewoor, S.H.Patil. Text Rank: A Novel Concept for Extraction Based Text Summarization. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3301 – 3304
10. Rafael Ferreira, Luciano de Souza Cabral, Frederico Freitas, Rafael Dueire Lins, Gabriel de França Silva, Steven J. Simske, Luciano Favaro, A multi-document summarization system based on statistics and linguistic treatment, Expert Systems with Applications, Volume 41, Issue 13, 1 October 2014, Pages 5780-5787, ISSN 0957-4174, http://dx.doi.org/10.1016/j.eswa.2014.03.023.
11. Aker Ahmet, Kurtic Emina, Balamurali A. R., Paramita Monica, Barker Emma, Hepple Mark, Gaizauskas Rob,"A Graph-Based Approach to Topic Clustering for Online Comments to News", Advances in Information Retrieval: 38th European Conference on IR
12. Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Pages. 15-29, isbn-978-3-319-30671-1, doi-10.1007/978-3-319-30671-1_2
13. Vishwanath Bijalwan,Vinay Kumar, Pinki Kumari,Jordan Pascual," KNN based Machine Learning Approach for Text and Document Mining" International Journal of