Volume : 4 | Issue : 10 | Special Issue Oct- 2015 • ISSN No 2277 - 8179

Document Summarization using Ontology

Computer Science

KEYWORDS: Document Summarization, Lesk Algorithm,Natural language Processing(NLP),Ontology,Word Net

Lata Makhija Manav Rachna International University Faridabad, India

Poonam Nandal Department of CSE, FET, MRIU,Faridabad

ABSTRACT

Document Summarization aims to summarize multi- document of same domain or single document by selecting the most relevant information. With the increase amount of information there is need to retrieve relevant documents as Summarization plays a vital role in weather forecasting ,news headlines ,games statistics ,book reviews ,conference programs and many more\Users do not have much time to spend reading thousands of lines. Today users want maximum information described by minimum words and relationships that exists between words. Many algorithms have been developed for generating summary of a given document. This paper discusses an improved and efficient approach for document summarization by using Ontology which describes relationships and contains set of concepts. The proposed approach increases accuracy and performance of simplified lesk algorithm which is an unsupervised disambiguation approach.

INTRODUCTION

Document Summarization is the process of extracting important information from the dataset or original document in such a way that it covers maximum information related to the event described in the document.

Text summarization can be broadly classified into two types: Single document summarization and multi-document summarization. Summarization can be categorized as extractive and abstractive. In extractive summarization, the output is created using sentences related to original document. In abstractive summarization, summary is created with its own words after understanding the text [1]. This paper focuses on

single document summarization that generates summary of single document using extractive summarization approach [2].

The main problem faced in natural language processing (NLP) is word-sense disambiguation [11] that is difficulty in identifying correct sense or meaning of the word that has multiple meanings as the volume of the information is constantly increasing every day. For example the word "Bank" can have multiple meanings such as financial institution, river bank, blood bank and so on. Many summarization approaches have been developed so far to solve word-sense disambiguation problem faced in summarization but the summary obtained is still not clear and concise.

The proposed approach involves Ontology which is a formal specification of a shared conceptualization [12] which helps in adding meaning and finds hidden relationship between the words and gives relevant and accurate summarization.

Remainder of the paper covers: Section II discusses the related work in the area of document summarization. Section III describes the Proposed Approach with architecture and flow chart; Section IV depicts result analysis along with comparison with the existing approach for document summarization ; Section V represents Conclusion and future scope of the paper.

II. RELATED WORK

A. Pal et al. [1] presents an unsupervised approach ,simplified lesk algorithm that summarizes a single document text in which important sentences are selected with the help of WordNet, an online dictionary and weights of all the sentences which are evaluated with the help of proposed approach. The proposed approach is modified version of lesk algorithm. The proposed approach depends on the glosses (meaning) extracted from WordNet rather than on format of text and position of a sentence.

V. Patil et al. [8] proposed a statistical approach for document summarization. Approach used is an application that takes a query from the user in the form of keyword and then givessummary for that keyword from the two or more web documents retrieved for that keyword and described Summarization as the process that composed of various phases including searching for top results for user query, gathering documents, extraction of data, analysis, generation of summary, presentation of summary to user. Performance of summarization is done based on parameters like retention ratio, length of summary.

L. Hennig et al. [9] discussed an approach in which sentences are mapped to nodes of a flexible, wide-coverage ontology and described that the mapping provides a semantic representation of the information content of sentences that improves summarization quality. By considering ontology attributes the semantic representation of a sentence's information content is improved. The classifier that maps sentences to the taxonomy is trained using search engines and is therefore very flexible and not bound to a specific domain. In this paper, SVM classifiers are trained to identify summary sentences using ontology-based sentence features.

Miguel Angel R Gaona et al. [11] proposed a measure for sense number score assignment based on web statistics. Various amount of information is used to find the frequencies of co-occurrences between the gloss and the context. The frequency is measured by a query to a search engine and chooses the sense which maximises the weight.

Huge amount of work has been done in this area and many algorithms using statistical approaches, semantic approaches have been proposed. But there is still need of improvement to increase accuracy and efficiency of the summary obtained for the document. In this paper a novel approach for document summarization using ontology is presented.

III. PROPOSED APPROACH

Many algorithms have been used for document summarization, one of them is simplified lesk algorithm which is a variant of lesk algorithm [11]. Simplified lesk algorithm is simple and fast but it has low accuracy rate.

The 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. The word having more than one sense or meaning is called ambiguous word.

The proposed approach for document summarization is by using ontology which adds meaningful information to the summary by

extracting relevant words and relationship between these words.
The proposed architecture is shown in figure-1.The main components of ontology based document summarization are sentence extraction, keyword extraction, sense matching and ontology. WordNet is an online semantic dictionary, electronic lexical database of nouns, verbs, adverbs, adjectives. In proposed approach words from sentences of document are extracted. After extraction of the words senses are matched and then using simplified lesk algorithm summary is obtained according to matched senses .This is extended by using ontology for finding relevant words and relationship between words obtained using lesk algorithm. The proposed approach mainly works in three stages:-

Stage 1:-Sentence Selection
In this stage, the sentences are segmented from the document and then sentences are extracted one by one. Sentences extracted are simplified. Firstly the redundant words are removed from the sentences and then stop words like a, the, an ,about which have no meaning in the dictionary are removed from the sentences and only keywords which have meaning in the dictionary are left.

Stage 2:-Content Realization
At this stage, the words left in the sentences are used for sense matching using WordNet-an online semantic dictionary where overlap value is calculated .Overlap value returns the number of words in common between two sets, ignoring function words or other words on a stop list. Words having maximum overlap value are kept and the words having minimum overlap value are discarded. Document is summarized using Simplified Lesk algorithm which is an extended version of Lesk algorithm.

Stage 3:-Ontology
At this stage, Ontology helps in finding hidden relationship between the words or adding meaning to the data. Firstly, Ontology of the selected document from the dataset is generated. Words that have no relationship between them are discarded. Secondly, Ontology of summary obtained using simplified lesk algorithm is generated and this ontology is matched with main ontology generated earlier and relationship between keywords are checked and words having relationship are kept and other are discarded and summary is obtained that is more precise and covers maximum information.

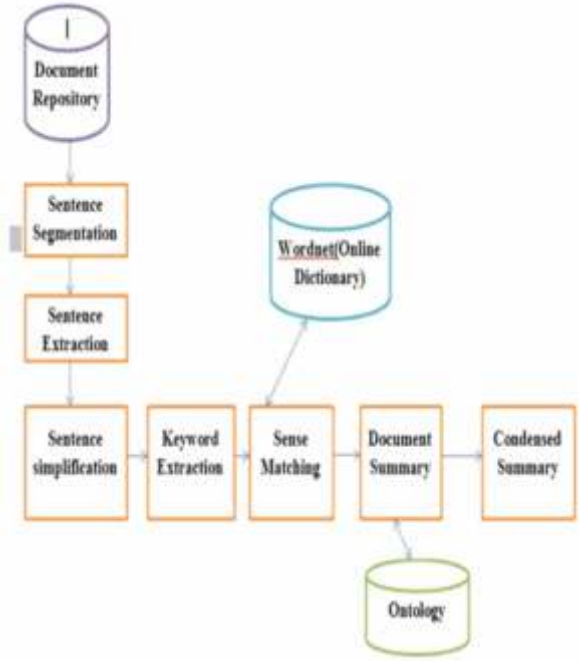


FIGURE-1 ONTOLOGY BASED DOCUMENT SUMMARIZATION ARHITECTURE

The step by step process of ontology based document summariza-



Figure-2 Step by step process of ontology based document summarization

A. Illustrative Example
We will explain our approach by taking example of a document D1 having content as:-

A book is written or printed sheet of information. The information may be related to various subjects and topics. The internet world has also introduced electronic form of books known as e-books.

According to our proposed approach sentences are extracted in first stage. Document D1 is break into sentences S1, S2, S3.

S1-A book is written or printed sheet of information

S2-The information may be related to various subjects and topics

S3-The internet world has also introduced electronic form of books known as e-books

A document is selected from the dataset of the same domain .Sentences are selected from the document and then redundant words and stop words are removed from the sentences and after removing them keywords are extracted from each sentence. Senses of each keyword are obtained using WordNet dictionary, an online semantic dictionary and overlap value is computed. Extracted keywords and their senses are shown in table-1.Keywords having maximum overlap value that is maximum common words matched in senses are used in the summary and the other words are discarded

Table 1-Keywods extracted for document D1

S.NO	Sentences	Keyword obtained	Senses
1.	S1	w1:-Book,w2:-written, w3:-printed, w4:-sheet ,w5:-information	W1:-written work, physical object, record, composition... W2:-set down in writing, systematically collected... W4:-used for writing or printing... W5:-message, collection of facts...
2.	S2	p1:-Related,p2:-various ,p3:-subjects,p4:-topics	P3:-situation or event, branch of knowledge... P4:-subject matter of conversation or discussion...
3.	S3	K1:-Internet,k2:-world,k3:-introduced ,k4:-electronic ,k5:-known ,k6:-e-books	K1:-computer consisting of network of network... K2:-involving the earth... K4:-using devices, concerned with electronics...

After calculating overlap value, summary is obtained using simplified lesk algorithm which is shown in Part-1

Document summary using simplified lesk algorithm:-

Part 1-A book is a written or printed sheet of information.

Lastly, Ontology is created to find relevant words on the basis of relationship exists between the words .Ontology of document D1 is generated and shown in figure-3



Figure-3-Main Ontology

After generating ontology of document D1.Ontology of summary obtained using simplified lesk algorithm is generated after discarding stop words and is shown in figure-4.

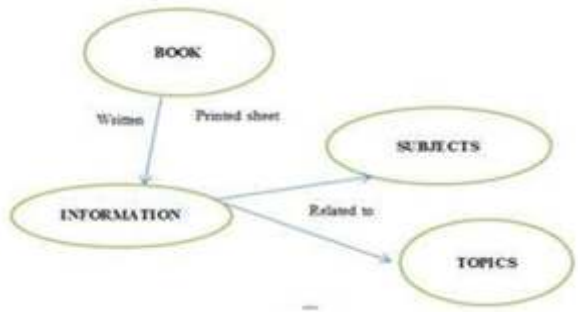


Figure-4-part of ontology from D1

Ontology created from the summary using simplified lesk algorithm is matched with the main ontology and only words having relationship are kept and others are discarded and a precise summary is obtained which is shown in part-2.

Document summary using ontology based summarization:-

Part 2-A book is written/printed sheet of information related to subjects.

IV. RESULT ANALYSIS

Performance of proposed approach for document summarization using ontology depends on words and the relationship that exits between the words in document and ontology. Our proposed algorithm is considered better than lesk algorithm because it give summary with words and the relationship that exits between them in more meaningful manner with respect to document with minimum number of words. We have found that in maximum number of cases we found our approach giving better results. The results obtained from the proposed approach and lesk algorithm presented in Table 2.

Table 2 Results of lesk algorithm and ontology based approach

S.NO	DOCUMENT	LESK ALGORITHM		ONTOLOGY BASED DOCUMENT SUMMARIZATION APPROACH	
		NO. OF WORDS	NO. OF RELATIONSHIP	NO. OF WORDS	NO. OF RELATIONSHIP
1.	D1	5	1	4	2
2.	D2	6	2	5	3
3.	D3	7	2	4	3
4.	D4	10	2	6	4

V. CONCLUSION AND FUTURE WORK

The proposed approach has given summary of document using ontology with meaningful words and relationship giving more information related to document. The proposed approach can be extended with automatic ontology based document summarization which will increase pre-processing speed.

REFERENCE

[[1] Alok Ranjan Pal, Projjwal Kumar Maiti and Diganta Saha,"An Approach To Automatic Text Summarization UsingSimplified Lesk Algorithm And Wordnet", InternationalComputer Modeling(IJCTCM)Journal of Control Theory andVol.3, No.4/5, September 2013|[2]A.R.Kulkarni, S.S.Apte, an automatic text summarizationusing lexical cohesion andcorrelation of sentences, International Journal of Researchin Engineering and Technology|[3]Fang Chen, Kesong Han and Guilin Chen, "An Approach to Sentence Selection Based Text Summarization", In the Proceedings of IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Volume: l,pp.489- 493,2002. |[4]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. |[5]I. Mani, M. Maybury (Eds.),"Advances in Automatic Text Summarization," MIT Press, Cambridge,MA, 1999. |[6]I. Mani," Automatic text summarization," John Benjamins, 2001. |[7]I. Mani, and M.T. Maybury, (eds), "Advances in Automatic Text Summarization," Cambridge, MA:MIT Press, 1999. |[8]Jayabharathy, Kanmani and Buvana, An Analytical Framework for Multi-Document Summarization, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011 |[9]K. Spärck Jones,"Automatic summarising: The state of the art," Information Processing & Management 4|(6),,pp 1449– 1481, 2007.|[10] Leonhard Hennig, Winfried Umbrath, Robert Wetzker,"An Ontology-based Approach to Text Summarization",IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,2008|[11] Miguel Ángel Ríos Gaona,Alexander Gelbukh, Sivaji Bandyopadhyay," Web-based Variant of the Lesk Approach toWord Sense Disambiguation", 2009 Eighth MexicanInternational Conference on Artificial Intelligence,IEEE|[14] Vishal Patil, Mahalakshmy Krishnamoorthy, Parag Oke,Prof. M. Kiruthika," A Statistical Approach for Document|[12] R. Barzilay, M . Elhadad, "Using lexical chains for text Summarization", International Journal of Advanced Computersummarization," In: Inderjeet Mani and Mark Maybury, Technology (IJACT),volume editors, Advances in Automatic Text Summarization, MITPress , , pp 111–122, 1999.|[13] ThomasR.Gruber," Toward Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal Human-Computer Studies 43, p.907-928, August 23,19