ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

A Semantic Based Approach for Abstractive Multi-Document Text Summarization

Dipti Bartakke [1], Dr S D Sawarkar[2], Archana Gulati [3]

P.G. Student, Department of Computer Engineering, Datta Meghe College of Engineering, Airoli, Mumbai, India[1]

Principal, Datta Meghe College of Engineering, Airoli, Mumbai, India[2]

Associate Professor, Department of Computer Engineering, Datta Meghe College of Engineering, Airoli, Mumbai, India[3]

ABSTRACT: One of the important Natural Language Processing applications is Text Summarization, which helps users to manage the vast amount of information available, by condensing documents content and extracting the most relevant facts or topics included. Text summarization is the process of extracting salient information from the source text and to present that information to the user in the form of summary. It is very difficult for human beings to manually summarize large documents of text. Text summarization approaches fall in two broad categories: extractive and abstractive. Extractive summarization produces summaries by choosing a subset of the sentences in the original document(s). This contrasts with abstractive summarization, where the information in the text is rephrased. In this paper, a novel approach is presented to create an abstractive summary for a multi-document using a rich semantic graph reducing technique. The approach summaries the input documents by creating a rich semantic graph for the original documents, reducing the generated graph, and then generating the abstractive summary from the reduced graph.

KEYWORDS: Text Summarization; Abstractive Summary; Semantic Representation; Rich Semantic Graph; Semantic Graph.

I. INTRODUCTION

Text summarization is the technique, where a computer summarizes a text. A text is entered into the computer and a summarized text is returned, which is a non-redundant extract from the original text. The intention of text summarization is to express the content of a document in a condensed form that meets the needs of the user. For more information that can realistically be digested is available on the World-Wide Web and in other electronic forms. News information, biographical information is so vast that it is not possible to read everything, hence it requires condensed information.

Text Summarization has become a very popular Natural Language Processing (NLP) task in recent years. Due to the vast amount of information, especially since the growth of the Internet, automatic summarization has been developed and improved in order to help users manage all the information available these days. The data on World Wide Web is growing at an exponential pace. Nowadays, people use the internet to find information through information retrieval (IR) tools such as Google, Yahoo, Bing and so on. However, with the growth of valid information on the internet, information abstraction or summary of the retrieved results has become necessary for users. In the current era of information overload, text summarization has become an important and timely tool for user to quickly understand the large volume of information. The goal of text summarization is to condense the documents into a shorter version and preserve important contents. This reduces user's time for finding the key information in the document.

Features: A summary can be defined as a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). Summarization systems can be characterised according to many features. There are three classes of context factors that influence summaries: input, purpose and output factors.

This allows summaries to be characterised by a wide range of properties. For instance, summarization has traditionally been focused on text, but the input to the summarization process can also be multimedia information, such as images, video or audio as well as on-line information or hypertexts. Furthermore, we can talk about summarizing only one

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

document (single-document summarization) or multiple ones (multi-document summarization). Regarding the output, a summary may be an extract (i.e. when a selection of "significant" sentences of a document is performed), abstract, when the summary can serve as a substitute to the original document or even a headline (or title). It is also possible to distinguish between generic summaries and user-focused summaries. The first type of summaries can serve as surrogate of the original text as they may try to represent all relevant features of a source text. The user-focused summaries rely on a specification of a user information need. Concerning also the style of the output, abroad distinction is normally made between two types of summaries. Indicative summaries are used to indicate what topics are addressed in the source text. As a result, they can give a brief idea of what the original text is about. The other types, the informative summaries, are intended to cover the topics in the source text.
Text summarization approaches can be broadly divided into two groups: extractive summarization and abstractive summarization. Extractive summarization extracts salient sentences or phrases from the source documents and groups them to produce a summary without changing the source text. Usually, sentences are in the same order as in the original document text. However, abstractive summarization consists of understanding the source text by using linguistic method to interpret and examine the text. The abstractive summarization aims to produce a generalized summary, conveying in information in a concise way, and usually requires advanced language generation and compression techniques. Summarization started with single document summarization which produces summary of one document. As research done, due to large amount of information on web, multi document summarization came into existence. Multi document summarization produces summaries from many source documents on the same topic or same event.

Abstractive methods need a deeper analysis of the text. These methods have the ability to generate new sentences, which improves the focus of a summary, reduce its redundancy and keeps a good compression rate. Generally, a summary should be much shorter than the source text. This characteristic is defined by the compression rate, which measures the ratio of length of summary to the length of original text.

II. RELATED WORK

In [2] Das focuses a survey on automatic text summarization which aims to investigate how empirical methods have been used to build summarization systems. It also describes single-document summarization, focusing on extractive techniques and discusses the area of multi-document summarization. Radev [3] define a summary as "a text that is produced from one or more texts, that convey important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". This simple definition captures three important aspects that characterize research on automatic summarization: Summaries may be produced from a single document or multiple documents, Summaries should preserve important information, Summaries should be short. In this paper [8] David presented a multilingual multi-document summarization system that uses text similarity to choose sentences from English documents based on the content of the machine translated documents. Summarization techniques can be classified according to many Summarization factors. For example, they can be classified according to the number of input documents (single-document versus multi-document), to the type of these documents (textual versus multimedia), to the output types (extractive versus abstractive), etc. [4]. Leskovec, et al. [9 and 11] provided a semantic graph-based approach to generate an extractive summary for a single input document. The purpose of their extractive summarization is to obtain the most important sentences from the original document by first generating the document semantic graph and then using the document and graph features to obtain the document summary. They created a semantic representation of the document, based on the logical form triplets (the basic form subject–verb–object of the document sentences) that have been retrieved from the text. For each generated triplet, they assign a set of features comprising linguistic, document, and graph attributes. They then train the linear Support Vector Machine classifier to determine those triplets that are useful for extracting sentences which are later compose the summary. In their approach, Leskovec, et al. aimed to create an extractive summary from the source document only, and hence they did not consider the abstractive summary. Besides, they use the semantic graph it its ordinary form to represent the input document, therefore the generated graph will be very huge, because the graph granularity level is high. Leskovec [17] presented a novel approach of document summarization by generating semantic representation of the document and applying machine learning to extract a sub-graph that corresponds to the semantic structure of a human extracted document summary. Stergos [22] Summarization techniques can be classified according to many Summarization factors. For example, they can be classified according to the number of input documents (single- document versus multi-document), to the type of these documents (textual versus multimedia), to the output types (extractive versus abstractive), etc. Lin

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

[23] proposed an approach; -gram sub-sequences of transitions per term in the discourse role matrix then constitute the more fine grained evidence used in our model to distinguish coherence from incoherence. Gangemi [24] have presented some preliminary results of OntoWordNet, a large scale project aiming at the "ontologization" of WordNet. We presented a twostep methodology: during the first, automatic phase, natural language word sense glosses in WordNet are parsed, generating a first, approximate definition of WN concepts (originally called synsets). In this definition, generic associations (A-links) are established between a concept and the concepts that co-occur in its gloss. In a second phase, the foundational top ontology DOLCE (in the DOLCE-Lite+ version), including few hundred formally defined concepts and conceptual relations, is used to interpret A-links in terms of axiomatised conceptual relations. This is a partly automatic technique that involves generating solutions on the basis of the available axioms, and then creating a specialized partition of the axioms (the set ∏ d + and its specializations) in order to capture more domain-specific phenomena.

III. PROPOSED SYSTEM

The aim of the project is to use the multiple documents in order to create abstractive summarization. At first, semantic graph is generated for every sentence in the documents by preprocessing each sentence. Thereafter, the generated graph is reduced to more reduced graph to generate abstractive summary. Heuristic rules have been used to generate abstractive summary. The goal of the system is to condense the documents into a shorter version and preserve important contents.

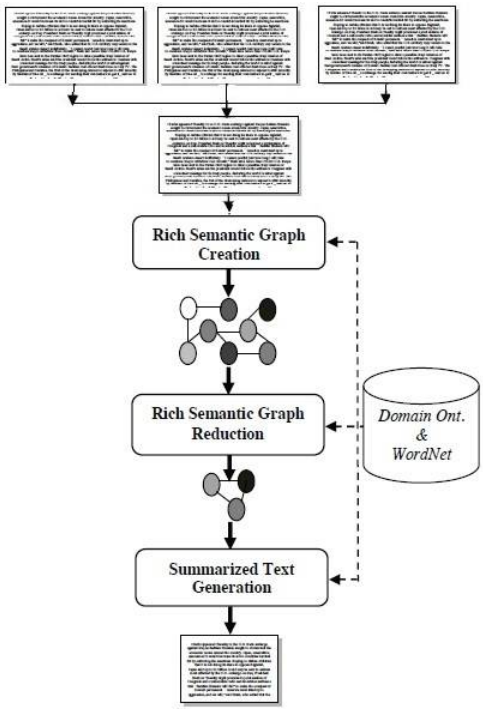


Figure 1: The Proposed System

The Rich Semantic Graph Creation Phase: The main objective of the Rich Semantic Graph Creation Phase is to represent the input documents semantically using Rich Semantic Graph (RSG). Unlike traditional semantic graph, the Rich Semantic Graph is able to capture the meaning of words, sentences, and paragraphs. The Rich Semantic Graph Creation phase is as shown in figure 2.

This phase starts with deep syntactic analysis of the input text, then generates typed dependency relations (grammatical relations), and syntactic and morphological tags for each word. After that, for each sentence, the model accesses the domain ontology to instantiate, interconnect and validate the sentence concepts to build rich semantic sub-graphs. Finally, the sentences rich semantic sub-graphs are merged together to represent the whole document semantically by

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

creating the final Rich Semantic Graph. It is composed of three modules: Preprocessing, Rich Semantic Sub-graphs Generation, and Rich Semantic Graph Generation modules.

1) The Preprocessing Module: Preprocessing module is responsible to accept the input text, and converts it to preprocessed sentences. It consists of four main processes: named entity recognition, morphological and syntactic analysis, cross-reference resolution, and pronominal resolution processes.

- The named entity recognition process locates atomic elements into predefined categories such as person names, organizations, etc.
- In morphological analysis, each word is divided into morphemes and figures out its grammatical categories, the syntactic analysis parses the whole sentence to describe each word syntactic function and build the parse tree, and typed dependencies expresses syntactic knowledge in terms of direct relationships between words.
- Co-reference and pronominal resolution reference resolution processes identify co-reference named entities and resolve pronominal references in the whole input text.

Co-reference is defined as the identification of surface terms (words within the document) that refer to the same entity.

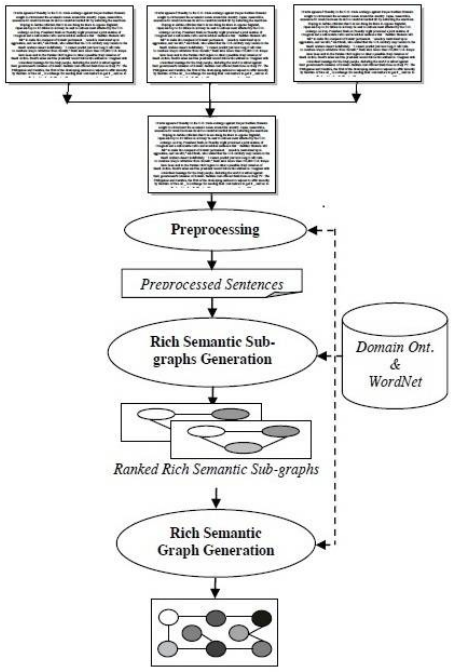


Figure 2: The Rich Semantic Graph Creation Phase

2) The Rich Semantic Sub-graphs Generation Module: The Rich Semantic Sub-graphs Generation module is responsible to transform each preprocessed sentence to a set of ranked rich semantic subgraphs. The main objective of the Rich Semantic Sub-graphs Generation module is to generate multiple rich semantic sub-graphs for each input preprocessed sentence. This module includes three processes: Word Senses Instantiation, Concepts Validation, and Semantic Sentences Ranking processes.

- Word Senses Instantiation process: For each input preprocessed sentence, this process instantiates a set of word concepts for both noun and verb senses based on the domain ontology.
- Concept Validation Process: In this process, for each preprocessed sentence, the sentence concepts instantiated are interconnected and validated to generate multiple rich semantic sub-graphs.
- Sentences Ranking Process: It aims to rank and to threshold the highest ranked rich semantic sub-graphs for each sentence. To generate single rich semantic graph and to keep the semantic consistency for the whole sentence, the process considers the first ranked rich semantic sub-graph only. The ranking method is based on deriving the average

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

weight of each concept (word sense). The weight of the word concept is derived according to its usage popularity (Wordnet usage popularity).

3) The Rich Semantic Graph Generation module: Finally, the Rich Semantic Graph Generation module is responsible to generate the final rich semantic graphs of the whole input document from the highest-ranked rich semantic sub-graphs of the document sentences. The semantic sub-graphs of the input document will be merged to form the final rich semantic graph.

The Rich Semantic Graph Reduction Phase: This phase aims to reduce the generated rich semantic graph of the original document to more reduced graph. In this phase, a set of heuristic rules are applied on the generated rich semantic graph to reduce it by merging, deleting, or consolidating the graph nodes.

	Rule 1	
IF	SN1 is instance of noun N	And
	SN2 is instance of noun N	And
	MV1 is similar to MV2	And
	ON1 is similar to ON2	
THEN	Merge both MV1 and MV2	And
	Merge both ON1 and ON2	
	Rule2	
IF	SN1 is instance of subclass of noun N	And
	SN2 is instance of subclass of noun N	And
	{[MV11, ON11],..[MV1n, ON1n]} is similar to {[MV21, ON21],..[MV2n, ON2n]}	
THEN	Replace SN1 by N1(instance N)	And
	Replace SN2 by N2(instance N)	And
	Merge both N1 and N2	
	Rule 3	
IF	SN1 and SN2 are instances of noun N	And
	MV1 is instance of subclass of verb V	And
	MV2 is instance of subclass of verb V	And
	ON1 is similar to ON2	
THEN	Replace MV1 by V1(instance V)	And
	Replace MV2 by V2(instance V)	And
	Merge both V1 and V2	And
	Merge both ON1 and ON2	
	Rule 4	
IF	SN1 and SN2 are instances of noun N	And
	MV1 is similar to MV2	And
	ON1 is instance of subclass of noun NN	And
	ON2 is instance of subclass of noun NN	
THEN	Merge both MV1and MV2	And
	Replace ON1 by NN1(instance NN)	And
	Replace ON2 by NN2(instance NN)	And
	Merge both NN1 and NN2	

Figure 3: Herustic Rules

The Text Generation Phase: The Rich Semantic Graph Generation module is responsible to generate set of ranked RSGs for the input ranked semantic sub-graphs. This phase aims to generate the abstractive summary from the reduced Rich Semantic Graph (RSG). There are four modules namely the Text planning, the Sentence Planning, the Surface

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Realization, and the Evaluation modules. These modules are performed by processes arranged as a pipeline, so the output of each process is the input of the next one as shown in figure 4.

1) The Text Planning module: It aims to select the appropriate content material to be expressed in the final text. This phase includes one process called "Content Determination", which decides what information should be included in the generated text.
2) The Sentence Planning module: It specifies the sentence boundaries, and generates and orders an intermediate paragraphs. The main objective of this phase is to improve the fluency or understandability of the text. The sentence planning consists of four main processes:

- Lexicalization Process: In this process, for each verb/noun object, its synonyms are selected by accessing the WordNet ontology to generate the target content.

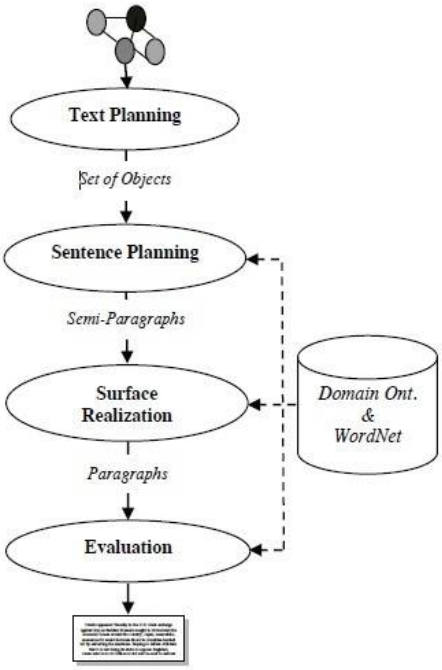


Figure 4: Text Generation Phase

- Discourse Structuring Process: The main aim of this process is to build a structure that contains the selected object synonyms in the form of pseudo-sentences
- Aggregation Process: The main aim of this process is to decide how pseudo-sentences should be combined into semi-paragraphs.
- Referring Expression Process: This process identifies and replaces the intended referent by its appropriate pronoun.

3) The Surface Realization module: This phase aims to transform the enhanced semi-paragraphs into paragraphs by correcting them grammatically (inflect words for tense, etc.) and adding the required punctuation (capitalization adding semicolon, etc). The Surface Realization module: This phase aims to transform the enhanced semi-paragraphs into paragraphs by correcting them grammatically (inflect words for tense, etc.) and adding the required punctuation (capitalization adding semicolon, etc).
4) The Evaluation module: The main objective of this phase is to evaluate and then rank the paragraphs according to two factors: coherence between paragraph sentences, and the most frequently used paragraph word synonyms.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

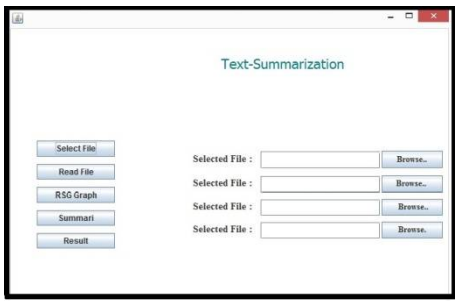
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

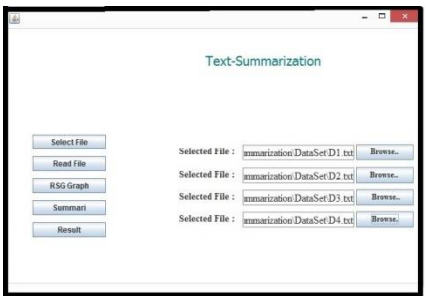
Vol. 4, Issue 7, July 2016

V. SYSTEM GUI

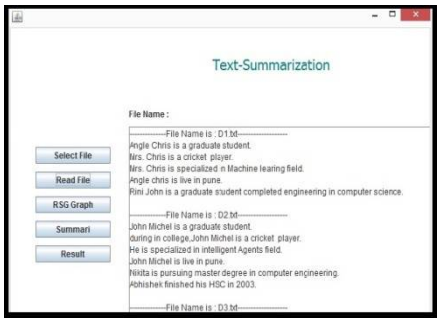
The Graphical User Interface (GUI) is designed to produce abstractive summary by using Semantic approach. The GUI is as illustrated in Figure (a). Figure (b) below shows files selection process, where the different documents are given as an input to the system. Figure (c) reads the selected files and will display on the screen. Figure (d) shows sentence wise graph with named entity recognizer and POS tagging. Figure (e) shows merging of sentences after heuristic rules are applied to the multiple documents. Figure (f) shows merging of sentences after heuristic rules are applied to the multiple documents. Figure (g) shows the summary of multiple documents



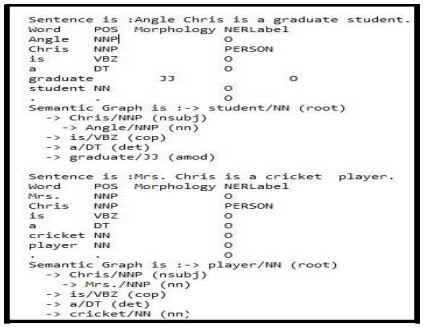
(a) **System GUI**



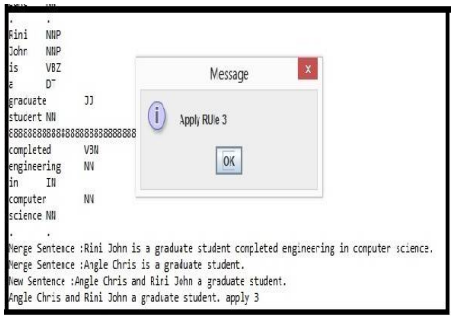
(b) **File Selection**



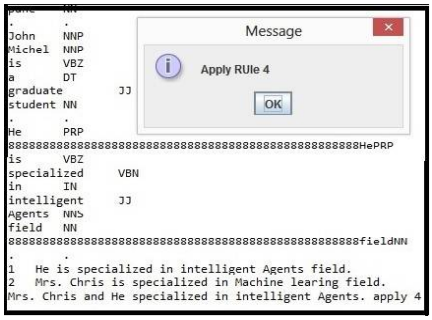
(c) **File Reading**



(d) **Sentence Wise Graphs**



(e) **Application of Heuristic rule 3**



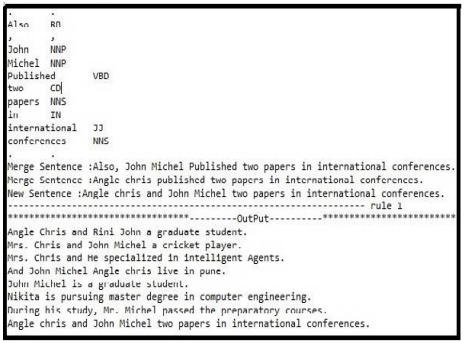
(f) **Application of Heuristic rule 4**

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016



(g) Abstractive Summary

VI. PERFORMANCE ANALYSIS

Our system provides evaluation by comparing the human generated summary with the system generated summary. The evaluation through precision and recall for the summary produced is shown in figure 6.

Original Documents:
File 1:
Angle Chris is a graduate student.
Mrs. Chris is a cricket player.
Mrs. Chris is specialized in Machine learing field.
Angle chris is live in pune.
Rini John is a graduate student completed engineering in computer science.
Albert is good in bating.
During his study, Mr. Michel passed the preparatory courses.

Human Generated Summary:
File 1:
Chris and Rini are graduate student.
He is a cricket player who live in Pune
And he is specialized in Machine learing field.
Mr. Michel passed the preparatory courses, during his study.

System Generated Summary:
File 1:
Angle Chris and Rini John a graduate student.
Mrs. Chris is a cricket player.
Mrs. Chris is specialized in Machine learing field.
Angle chris is live in pune.
Albert is good in bating.

Figure 6: Comparision of human generated summary with system generated summary

Precision & Recall

Precision
The precision of the summary can be defined as the measurement of the retrieved relevant sentences to the query of the total retrieved sentences. Precision measures the accuracy of the summary.
Precision=A/B
Where,
A = No of relevant retrieved sentences &

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

B = the total retrieved sentence.

Recall

The recall of the summary can be defined as the measurement of the retrieved relevant sentences to the total database sentences.

Recall=A/C

Where,

A = No of relevant retrieved sentences and

C = Total number of relevant sentences in the database.

On an average precision obtain for the summary generated by the system is 70 %.

VI. CONCLUSION AND FUTURE WORK

Information condensation is needed. Extractive summary leads usually for sentence extraction rather the summarization. So the need is to generate summary that captures the important text and relates the sentences semantically. The work is applicable in open domain. Abstractive summarization will serve as a tool for generating summary which is semantically correct and produced fewer amounts of sentences in summary. Extractive summarization leads to sentence extraction based on statistical methods which are not useful always. In this paper, a model is proposed to create an abstractive summary for multiple documents using a semantic graph reducing approach. The approach summaries the source documents by creating a semantic graph called Rich Semantic Graph for the original documents, reducing the generated semantic graph to more abstracted graph, and generating the abstractive summary from the reduced graph. It reduces the original document to almost fifty percent. The average precision for the above results is obtained as 70%. In the future work, we are going to develop a model for multilingual multi-document summarization. We can also try to summaries the multilingual multi- documents in the form of video, or in the form of audio.

REFERENCES

1. Ibrahim F. Moawad, Mostafa Aref," Semantic Graph Reduction Approach for Abstractive Text Summarization" 2012 IEEE
2. D. Das, A. Martins, "A Survey on automatic text summarization", Unpublished, Literature survey for Language and Statistics II, Carnegie Mellon University, 2007.
3. D. Radev, E. Hovy, K. McKeown, "Introduction to the Special Issue on Summarization", Computational Linguistics, Vol. 28, No. 4, pp. 399-408, 2002.
4. A. Khan, N. Salim, "A Review on Abstractive Summarization Methods", in JATIT, Vol. 59 No. 1, Jan 2005.
5. S. Ismail, I. Moawad, M. Aref, "Arabic Text Representation using Rich Semantic Graph".
6. I. Fathy, D. Fadl, M. Aref, "Rich Semantic Representation Based Approach for Text Generation", The 8th International conference on Informatics and systems (INFOS2012), Egypt, 2012.
7. K. Svore, L. Vanderwende, C. Burges, "Enhancing single-document summarization by combining RankNet and third-party sources", In Proceedings of the EMNLP-CoNLL, pp. 448-457, 2007.
8. D. Evans, K. McKeown, J. Klavans, "Similarity-based Multilingual Multi-Document Summarization", Technical Report CUCS-014-05,Department of Computer Science, Columbia University, Apr 2005.
9. J. Leskovec, M. Grobelnik, N. Milic-Frayling, "Learning Sub-structures of Document Semantic Graphs for Document Summarization", inKDD2004 Workshop on Link Analysis, 2004.
10. Stanford Parser, http://nlp.stanford.edu:8080/parser/index.jsp, June 15, 2012.
11. J. Leskovec, M. Grobelnik, N. Milic-Frayling, "Extracting Summary Sentences Based on the Document Semantic Graph", Microsoft Research, 2005.
12. D. Rusu, B. Fortuna, M. Grobelnik, D. Mladenić, "Semantic Graphs Derived From Triplets With Application In Document Summarization", International journal of Computing and Informatics, Vol.33, No.3, 2009.
13. B Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press, 1998.
14. R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber, "The Penn Discourse Treebank 2.0", Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Morocco.
15. https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.
16. http://nlp.stanford.edu/software/CRFNER.html.
17. J. Leskovec, M. Grobelnik, N. Milic-Frayling, "Learning Semantic Graph Mapping for Document Summarization", 2000.
18. Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh, "A Comprehensive Survey on Text Summarization Systems" 2009 In proceeding of: Computer Science and its Applications, 2nd International Conference.
19. Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacha Ibrahim F.Moawadryya, "Generic Text Summarization Using Wordnet", Language Resources Engineering Conference (LREC 2004), Barcelona, May, 2004.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

20. Silber G.H., Kathleen F. McCoy, "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization," Computational Linguistics 28(4): 487-496, 2002.
21. Barzilay, R., Elhadad, M, "Using Lexical Chains for Text Summarization," in Proc. ACL/EACL'97 Workshop on Intelligent Scalable Text summarization, Madrid, Spain,1997, pp.10–17.
22. A.Stergos, K. Vangelis, S. Panagiotis, "Summarization from medical documents: a survey", Artificial intelligence in medicine, Vol. 33, No. 2, pp. 157-77, 2005.
23. Z. Lin, H. Ng, M. Kan, "Automatically Evaluating Text Coherence Using Discourse Relations", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), Portland, Oregon, USA, 2011.
24. A. Gangemi, R. Navigli, P. Velardi, "The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet", In Proc. of International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003), Catania, Italy, pp. 820–838, 2003.

BIOGRAPHY

Dipti Abhishek Bartakke is a student in the Computer Department, Datta Meghe College of Engineering, Mumbai University. My research interests are Image Processing, Natural Language Processing, Data Mining, etc.