Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method.

Manjula Subramaniam[1], Prof. Vipul Dalal[2]

Computer Engineering, Vidyalankar Institute of Technology, Maharashtra, India

---***---

Abstract - *In this paper we present a method for summarizing Hindi Text document by creating rich semantic graph(RSG) of original document and identifying substructures of graph that can extract meaningful sentences for generating a document summary. This paper contributes the idea to summarize Hindi text document using abstractive method. We extract a set of features from each sentence that helps identify its importance in the document. It uses Hindi WordNet to identify appropriate position of word for checking SOV (Subject-Object-Verb) qualification. Therefore to optimize the summary, we find similarity among the sentences and merge the sentence which represented using Rich Semantic Sub graph which in turn produces a summarized text document.*

Key Words: Text Analysis, Text Summarization, Abstractive Summary and Rich Semantic Graph Representation.

1. INTRODUCTION

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, and Bing and so on.

However, with the exponential growth of 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. Therefore to achieve this goal of summarizing a text document is condense the document and preserve the important contents. Nowadays there is a wide range of technologies which focuses on areas like Human Language Technology (HLT). These include areas such as Natural Language Processing (NLP), Speech Recognition, Machine Translation, Text Generation and Text Mining. In this paper, we will focus on two of these areas: NLP and Text Mining which leads to summarizing text.

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. Currently, the need for text summarization has appeared in many areas such as news articles summary, email summary, short message news on mobile, and information summary for businessman, government officials, research, online search engines to receive the summary of pages found and so on[1].

Text summarization approach is broadly classified into two summary: extractive and abstractive.

Extractive summary is the procedure of identifying important sections of the text and producing them verbatim while Abstractive summary aims to produce important material in a new generalized form. [1]

In this paper, a novel approach is presented to generate an abstractive summary automatically for the Hindi input text document using a semantic graph reducing technique. This approach exploits a new semantic graph called Rich Semantic Graph (RSG) [3, 7].RSG is an ontology-based representation developed to be used as an intermediate representation for Natural Language Processing (NLP) applications. The new approach consists of three phases: creating a rich semantic graph for the source document, reducing the generated rich semantic graph to more abstracted graph, and finally generate the abstractive summary from the abstracted rich semantic graph.

2. BACKGROUND AND RELATED WORK

Text Summarization is shorter version of the original document while still preserving the main content available in the source documents. There are various definitions on text summary in the literature.

According to [8] "The aim of automatic text summarization is to condense the source text by extracting its most important content that meets a user's or application needs". According to [9],"A summary is a text that is produced from one or more texts that contains a significant portion of the information text(s), and is no longer than half of the original text(s)".

There are various effective techniques to generate extractive summary which helps to find relevant sentences to be added to the summary. This can be classified as : Statistical, Linguistic and Hybrid approach.

2.1 Statistical Method:

In Statistical Text summarization based on this approach relies on the statistical distribution of certain features and it is done without understanding whole document. It uses classification and information retrieval techniques. Classification methods classifies the sentences that can be part of the summary depending on the training of data. Information retrieval technique uses Position, length of sentence or word occurrences in the document . This method extracts sentences that occur in the source text, without taking into consideration the semantics of the words [10].

2.2 The Linguistics Method:

In this, method needs to be aware of and know deeply the linguistic knowledge, so that the computer will be able to analyze the sentences and then decide which sentence to be selected. Parameters can be cue words, Title feature or Noun and verbs in the sentences[11].Statistical approaches may be efficient in computation but Linguistic approaches look into term semantics, which may yield better summary results. In practice, linguistic approaches also adopt simple statistical computation (term-frequency-inverse-document-frequency (TF-IDF) weighting scheme) to filter terms.

3. THE PROPOSED APPROACH.

The proposed approach aims to summarize an input single text document by creating a semantic graph called Rich Semantic Graph (RSG) for the original document, reducing the generated semantic graph, and then generating the final abstractive summary from the reduced semantic graph. The approach consists of three phases: The Rich Semantic Graph Creation Phase, The Rich Semantic SubGraph Reduction Phase, and Summarized Text Generation Phase.

In RSG, the verbs and nouns of the input document are represented as graph nodes along with edge corresponding to semantic and topological relations between them.

The Rich Semantic Graph Reduction Phase aims to reduce the generated rich semantic graph of the source document to more reduced graph. A model of heuristic rules is applied to reduce the graph by replacing, deleting, or consolidating the graph nodes using the WordNet relations [12, 13]. Finally, the Summarized Text Generation Phase aims to generate the abstractive summary from the reduced rich semantic graph. This phase accepts a semantic representation in the form of RSG and generates the summarized text. The input graph contains the information needed to generate the final text.

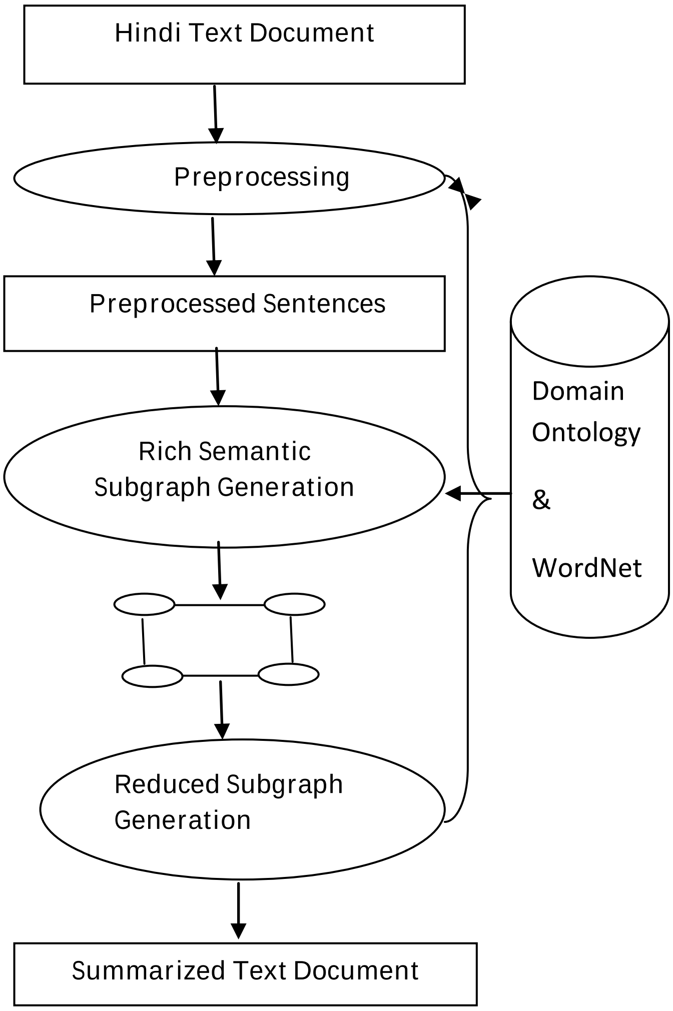

Hindi Text Document
Preprocessing
Preprocessed Sentences
Rich Semantic Subgraph Generation
Reduced Subgraph Generation
Summarized Text Document
Domain Ontology & WordNet

Fig 1 The Proposed Approach Architecture.

A. Rich Semantic Graph Creation Phase

This phase [2, 7] 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 creating the final Rich Semantic Graph. As shown in the figure 1 the Rich Semantic Graph Creation phase, where it is composed of three modules: Preprocessing, Rich Semantic Sub-graphs Generation, and Rich Semantic Graph Generation modules.

3.1 The Preprocessing module:

The preprocessing involves preparing text document for the analysis the text document based on Sentence Position Model, Word Position model, Stop word Removals and Stemming

3.2 Sentence Position Model:

Sentence position information is very vital in identifying the topic of the text. This is applied to entire text document. Given a sentence with its position in text and count the words in the given sentences. In Hindi, sentence is segmented by identifying the boundary of sentence which ends with a puram viram(|).

3.3 Word Position Model:

In this informative word position model. It split the sentence into words by identifying spaces, commas, special symbols between the words. So thereby identifying the word appearing at which position in the sentence.

3.4 Stop Word Removals:

In computing, stop words are words which are filtered out before or after processing of natural language data (text). Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. In the Hindi WordNet the words are grouped together according to their similarity of meanings. For each word there is a synonym set, or synset, in the Hindi WordNet, representing one lexical concept. Synsets are the basic building blocks of Word Net. The Hindi WordNet deals with the content words, or open class category of words. Thus, it contains the following category of words- Noun, Verb, Adjective and Adverb.

3.5 Stemming:

Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem needs not to be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Stemming is used for matching the words of sentence for checking similarity features.

3.6 Feature Extraction:

Actual analysis of the document for summarization begins in this phase. In this section very sentence is represented by a vector of feature terms .Every sentence is checked statically and linguistically. Each sentence has a score based on the weight of the feature terms which in turn is used for sentence ranking Feature term values ranges between 0 to 1.
Various parameters are calculated:
Sentence length, Average TF_ISF (Term Frequency-Inverse Sentence Frequency), Sentence position, Numerical Data, Title Feature, SOV Qualification, Subject Similarity.

B .Rich Semantic Sub Graph Generation

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. These rules exploit the WordNet semantic relations: hypernym, holonym, and entailment. There are many rules can be derived based on many factors: the semantic relation, the graph node type (noun or verb), the similarity or dissimilarity between graph nodes. The set of heuristic rule that can be applied on the graph nodes of two simple sentences: Sen1= [SN1, MV1, ON1] and Sen2= [SN2, MV2, ON2]. Each sentence is composed of three nodes: Subject Noun (SN) node, Main Verb (MV) node, and Object Noun (ON) node.

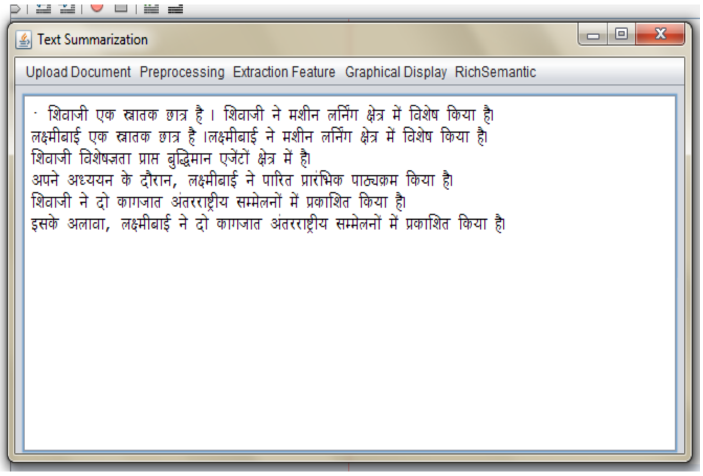
C. The Summarized Text Generation Phase:

This phase aims to generate the abstractive summary from the reduced Rich Semantic Graph (RSG) [18]. To achieve its task, the phase accesses the domain ontology, which contains the information needed in the same domain of RSG to generate the final texts. Besides, the WordNet ontology is accessed to generate multiple texts according to the word synonyms. The generated multiple texts are evaluated and ranked, where the most ranked text is considered.

CONCLUSION:

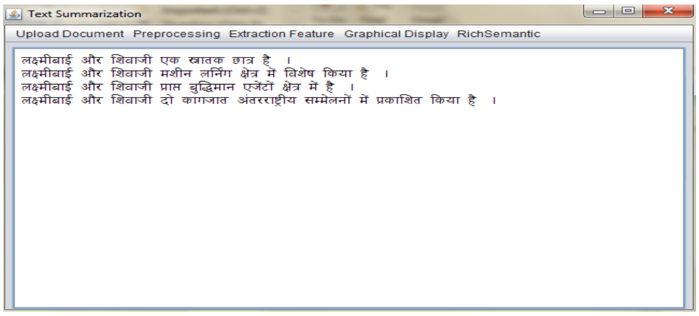
In conclusion, a novel approach to create an abstractive summary for a single document using a semantic graph reducing approach was presented in this paper. The approach summaries the source document by creating a semantic graph called Rich Semantic Graph for the original document, reducing the generated semantic graph to more abstracted graph, and generating the abstractive summary from the reduced graph. Therefore this model shows how Hindi text document is summarized using abstractive method.

Upload Text Document



· शिवाजी एक स्नातक छात्र है । शिवाजी ने मशीन लर्निंग क्षेत्र में विशेष किया है।
लक्ष्मीबाई एक स्नातक छात्र है ।लक्ष्मीबाई ने मशीन लर्निंग क्षेत्र में विशेष किया है।
शिवाजी विशेषज्ञता प्राप्त बुद्धिमान एजेंटों क्षेत्र में है।
अपने अध्ययन के दौरान, लक्ष्मीबाई ने पारित प्रारंभिक पाठ्यक्रम किया है।
शिवाजी ने दो कागजात अंतरराष्ट्रीय सम्मेलनों में प्रकाशित किया है।
इसके अलावा, लक्ष्मीबाई ने दो कागजात अंतरराष्ट्रीय सम्मेलनों में प्रकाशित किया है।

Final Summary



लक्ष्मीबाई और शिवाजी एक स्नातक छात्र है ।
लक्ष्मीबाई और शिवाजी मशीन लर्निंग क्षेत्र में विशेष किया है ।
लक्ष्मीबाई और शिवाजी प्राप्त बुद्धिमान एजेंटों क्षेत्र में है ।
लक्ष्मीबाई और शिवाजी दो कागजात अंतरराष्ट्रीय सम्मेलनों में प्रकाशित किया है ।

REFERENCES:

[1] Ibrahim F. Moawad, Mostafa Aref “Semantic Graph Reduction Approach for Abstractive Text Summarization” IEEE 2012.

[2] M. Aref, I. Moawad, S. Ibrahim., "Rich Semantic Graph Generation System Prototype", The tenth Conference on Language Engineering, Cairo, Egypt, 2010.

[3] Chetana Thaokar, Dr.Latesh Malik “Test Model for Summarizing Hindi Text using Extraction Method” Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).

[4] J. Leskovec, M. Grobelnik, N. Milic-Frayling, "Learning Sub-structures of Document Semantic Graphs for Document Summarization", in KDD2004 Workshop on Link Analysis, 2004.

[5] J. Leskovec, M. Grobelnik, N. Milic-Frayling, "Learning Semantic Graph Mapping for Document Summarization", 2000.

[6] Kedar Bellare, Anish Das Sarma, Atish Das Sarma, Navneet Loiwal,Vaibhav Mehta, Ganesh Ramakrishnan, Pushpak Bhattacharyya “Generic Text Summarization using WordNet”.

[7] I. Moawad, M. Aref, S. Ibrahim, "Ontology-based Model for Generating Text Semantic Representation", the International Journal of Intelligent Computing and Information Sciences “IJICIS”, Vol. 11, No. 1, pp. 117-128, January 2011.

[8]D. Evans, K. McKeon, J. Klavans, "Similarity-based Multilingual Multi-Document Summarization", Technical Report CUCS-014-05,Department of Computer Science, Columbia University, Apr 2005.

[9] A. Stergos, K. Vangelis, S. Panagiotis, "Summarization from medical documents: a survey", Artificial intelligence in medicine, Vol. 33, No. 2,pp. 157-77, 2005.

[10] Hovy, E. H. “Automated Text Summarization”. Oxford Handbook of Computational Linguistics, pages 583-598.Oxford University Press, 2005.

[11] Dehkordi, P.K., Khosravi, H., Kumarci. “Text Summarization Based on Genetic programming”. International Journal of Computing and ICT Research, 2009 - 3(1), 57-64.

[12]C. Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press, 1998.

[13] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, Five Papers on WordNet. Cognitive Science Laboratory, Princeton University, Princeton, 1990.