A Survey of Extractive and Abstractive Automatic Text Summarization Techniques

Vipul Dalal
Research Scholar, CSE Department
G.H.Raisoni College of Engineering,
Nagpur.
vipul.dalal@vit.edu.in

Dr. Latesh Malik
Professor & Head, CSE Department,
G.H.Raisoni College of Engineering,
Nagpur.
latesh.malik@raisoni.net

Abstract— **The existence of the World Wide Web has caused an information explosion. Readers are overloaded with lengthy text documents where a shorter version would suffice. All computer users, be it professionals or novice users, are particularly affected by this predicament. There exists an urgent need for the discovery of knowledge embedded in digital documents. This paper intends to investigate techniques and methods used by researchers for automatic text summarization. Special attention is paid to Bio-inspired methods for text summarization.**

Keywords—**bio-inspired algorithms, text mining, text summarization**

I. INTRODUCTION

To automate the process of abstracting, researchers generally rely on a two phase process. First, key textual elements, e.g., keywords, clauses, sentences, or paragraphs are extracted from text using linguistic and statistical analyses. In the second step, the extracted text may be used as a summary. Such summaries are referred to as "extracts". Another approach called "abstractive summarization" consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. Such abstracts may or may not contain the sentences from the original document.

The objective of this paper is to investigate both extractive and abstractive methods.

II. SURVEY OF EXTRACTIVE SUMMARIZATION

The extractive automatic text summarization work involving bio-inspired algorithms is as follows.

M. S. Binwahlan et al [1] introduced a work for feature selection. They exploited five features regarding to text summarization and the PSO was used to train the system to obtain the weights of each feature. These weights have been employed in their next work [2] to generate the best summary. The results shown that, the proposed PSO method generate summaries which are 43% similar to the manually generated summaries, while MS-Word summaries are 37% similar.

Albaraa Abuobieda M. Ali et al [3] presented a feature selection method using (pseudo) Genetic probabilistic-based Summarization (PGPSum) model for extractive single document summarization. The proposed method, working as features selection mechanism, was used to extract the weights of features from texts. Then, the weights were used to tune features' scores in order to optimize the summarization process. In this way, important sentences were selected for representing the document summary. The results showed that, their PGPSum model outperformed Ms-Word benchmarks by obtaining a similarity ratio closest to human benchmark summary.

IV SURVEY OF ABSTRACTIVE SUMMARIZATION

Barzilay et al [4] investigated a technique to produce a summary of an original text without requiring its full semantic interpretation, but instead relying on a model of the topic progression in the text derived from lexical chains. Summarization proceeds in four steps: the original text is segmented, lexical chains are constructed, strong chains are identified and significant sentences are extracted.

Kavita Ganesan et al [5] presented a novel graph-based summarization framework that generates concise abstractive summaries of highly redundant opinions. Evaluation results on summarizing user reviews show that their summaries have better agreement with human summaries compared to the baseline extractive method. The key idea of their approach is to first construct a textual graph that represents the text to be summarized. Then, three unique properties of this graph are used to explore and score various sub-paths that help in generating candidate abstractive summaries.

Eduard Hovy and Chin-Yew [6] Lin attempted to create a robust automated text summarization system, SUMMARIST, based on the equation:

summarization = topic identification + interpretation +generation.

Each of these stages contains several independent modules, many of them trained on large corpora of text. SUMMARIST provides both extracts and

978-1-4799-2561-2/13 $31.00 © 2013 IEEE
DOI 10.1109/ICETET.2013.31

abstracts for arbitrary English and other-language text. SUMMARIST combines robust NLP processing (using IR and statistical techniques) with symbolic world knowledge. To produce abstract-type summaries, the core process is a step of interpretation. In this step, two or more topics are fused together to form a third, more general one. This step must occur in the middle of the summarization procedure: First, an initial stage of topic identification and extraction is required to find the central topics in the input text; finally, to produce the summary, a concluding stage of sentence generation is needed.

Jurij Leskovec et al [7] presented a method for extracting sentences from an individual document to serve as a document summary. They applied syntactic analysis of the text that produces a logical form analysis for each sentence. They used subject–predicate–object (SPO) triples from individual sentences to create a semantic graph of the original document and the corresponding human extracted summary. Using the Support Vector Machines learning algorithm, they trained a classifier to identify SPO triples from the document semantic graph that belong to the summary. The classifier is then used for automatic extraction of summaries from test documents.

IV CONCLUSION AND FUTURE SCOPE

The abstractive approaches perform detailed linguistic analysis of the text and generate summary similar to a summary generated by human. So they outperform extractive approaches but are more expensive computationally. Bio-inspired algorithms are well known for their optimization capabilities. Combining bio-inspired techniques with abstractive approach should optimize the computation cost and still generate summaries similar to human extracted summaries. After this detailed survey, our future work is to propose a novel technique based on bio-inspired abstractive approach for automatic text summarization.

REFERENCES

[1] M. S. Binwahlan, Salim, N., & Suanmali, L., "Swarm based features selection for text summarization," *International Journal of Computer Science and Network Security IJCSNS,* vol. 9, pp. 175-179, 2009b.

[2] M. S. Binwahlan*, et al.*, "Swarm Based Text Summarization," in *Computer Science and Information Technology – Spring Conference, 2009. IACSITSC '09. International Association of*, 2009, pp. 145-150.

[3] Albaraa Abuobieda M. Ali, Naomie Salim, Rihab Eltayeb Ahmed, Mohammed Salem Binwahlan, Ladda Sunamali, Ahmed Hamza, "Pseudo Genetic And Probabilistic-Based Feature Selection Method For Extractive Single Document Summarization", *Journal of Theoretical and Applied Information Technology, 15th October 2011. Vol. 32 No.1, ISSN: 1992-8645, E-ISSN: 1817-3195.*

[4] Regina Barzilay, Michael Elhadad. "Using Lexical Chains for Text Summarization".
In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97). Madrid: ACL, 1997. 10-17.

[5] Kavita Ganesan, ChengXiang Zhai, Jiawei Han, "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions".

[6] Eduard Hovy and Chin-Yew Lin. "Automated Text Summarization in SUMMARIST". *In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. MIT Press..*

[7] Jurij Leskovec, Natasa Milic-Frayling, Marko Grobelnik, "Extracting Summary Sentences Based on the Document Semantic Graph" *Microsoft Research, Microsoft Corporation*