STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720
Volume 6, Issue 3
www.stmjournals.com

Semantic Summarizer

Basit Farooq**, Amjad Hussain
Department of Information Technology, Central University of Kashmir, Jammu and Kashmir, India

Abstract

This work presents the survey of the existing approaches used for automatic text summarization. Automatic text summarization technique belongs to the natural language processing area, and is applied on the source document to produce its compact version that preserves its aggregate meaning and key concepts. On a broader scale, approaches for text summarization task are classified into two categories: (1) abstractive and (1) extractive. In abstractive summarization, main contents of the input text are paraphrased possibly using vocabulary that is not present in the source document, while as in extractive summarization, output summary is a subset of the input text and is generated by using sentence ranking technique. In this paper, the main ideas behind the existing methods used for abstractive and extractive summarization are discussed broadly. Comparative study on these methods is also highlighted.

Keywords: *Text summarization, extractive summarization, sentence ranking methods, abstractive summarization, structured-based approach, semantic-based approach*

****Author for Correspondence*** E-mail: basit_ahmad@gmail.com

INTRODUCTION

The exponentially increasing digital data which is accessible world-wide makes utilization of an automatic text summarization tool inevitable as manual text summarization entails considerable number of impartial and knowledgeable experts. The sole objective of automatic text summarization is to express all information in the input text in vivid, concise and comprehensive manner enabling users to save efforts and time. Initially automatic text summarization techniques were applied to one input document called *single document text summarization.* The enormous amount of redundant data present on the web provoked the use of *Multi-document text summarization*, where set of multiple documents served as an input to the system.

According to Aliguliyev [1] process of automatic summarization can be divided into three steps: (a) preprocessing of the input text, (b) intermediate representation, and (c) generating an output as summary. SUMMARIST text summarization system introduced by Hovy and Lin [2] implements three phases: (a) topic identification, (b) interpretation, and (c) generation.

Text summarization task is broadly classified into two categories: (1) abstractive and (2) extractive [3]. Extractive summaries are formed by concatenating main sentences or phrases of the source document. Identification of the main sentences in the input document is a challenging task, which is accomplished using sentence scoring or ranking methods. On the other hand, Abstractive summaries are the compressed paraphrased version of the input text and thus are not a mere concatenation of the main sentences or phrases present in the input document.

On the basis of the content summaries are classified into two classes: (1) *indicative summary* and (2) *informative summary*. Indicative summary points to the main concepts of the input document while as informative summary includes all the relevant information that is reported in the input document [4]. Table 1 briefly described the summarization types.

This paper explains the different methods used for extractive and abstractive summarization. Section below presents the related work; subsequent sections explain the different extractive summarization methods; and present

Table 1: *The Summarization Types.*

Summarization types	Definition
Extractive	Concatenation of important sentences or phrases of input text.
Abstractive	The compressed paraphrased version of the input text.
Single Document	Single document serves as an input to the system.
Multi-document	Multiple documents serve as an input to the system.
Indicative	Points to the main concepts of the input document.
Informative	Includes all the relevant information that is reported in the input document.
Keyword	Consists of set keywords or phrases present in the input text.
Headline	Summarizes input document by a single important sentence.
Generic	Does not make any assumptions regarding domain or genre of the input; determines importance w.r.t. the contents of the input document.
Query Focused	Determines important parts of the input document w.r.t. the query provided by the user.

the different abstractive summarization methods; conclusion is given in a separate section and last section contains references.

RELATED WORK

The task of automatic text summarization began in 1950s [5]. It is now over half a century old and is still progressing because of the increased use of digital data. Luhn [5] unfolded the concept that how frequently occurring words can help in determining the important sentences. Then Edmundson [4] broadened Luhn's approach by imparting several other features for indicating salient sentences: (a) frequency/count of the word in input text, (b) frequency of the title words in the sentence, (3) position of the sentence w.r.t. the entire input document, (4) count of cue-phrases like "significantly", "concluding" [4]. Researchers mostly focused on the single and multi-document summarization using extractive approach.

At that time, Paice was the one who focused on the techniques for language generation. He pinpointed the main problem that sentence extraction algorithms suffered from (that was unintended inclusion those sentences in the summary that contained references to the sentences absent in the summary), which resulted in inconsistent summaries [6]. This was the very early work done; subsequent section discusses the recent work done in the field of automatic text summarization.

Methods like lexical aggregation used in also helped in condensing the input text by replacing two related concepts by another concept, for example, *selling and buying* are related to each other and therefore, were replaced by *business* [7]. For redundancy removal, syntactic aggregation method was used, for example, *Sam plays and Lin plays* becomes *Sam and Lin plays* [7].

Summaries generated on the basis of the keywords are called keyword summaries. For keyword summarization, we need to identify the keywords present in the input document, Lin and Hovy [8] describes the methods for keyword identification. Query focused summarization determines important parts of the input document w.r.t. the query provided by the user. According to Ouyang et al. [9], the similarity of the sentences to the user query is calculated using SVR (support vector regression). Carbonell and Goldstein [10] also summarize multiple documents on the basis of user query.

In order to generate quality summary, researchers in recent work have focused on employing neural networks and fuzzy logic [11]. It was also demonstrated that neural networks and fuzzy logic based summarizers outperform statistical method based summarizers. Neural networks and fuzzy logic were even used for improving and addressing in sentence scoring techniques [12].

Recently, Chengcheng [13] used neural network in order to summarize the news articles. In the training phase neural network learnt how to check important features of sentences. On the basis of these features input text was filtered by the neural network and in the end summary of the news article was generated.

Deep learning approaches for text summarization have also shown considerable results. According Yousefi-Azar and Hamey

[14], *deep auto-encoder* is used to generate extractive query-focused summary for a single document. *Ensemble noisy auto-encoder*, an extension of *deep auto-encoder*, creates noisy input by adding random noise to input representation, in order to choose sentences from a cluster of noisy inputs. Experiments were performed on two separate email corpora which were publically available and the system was evaluated using ROUGE.

Lately Nallapati [15] used Attentional Encoder-Decoder Recurrent Neural Networks to frame their abstractive automatic text summarizer. This work also attempted to address serious problems occurring in basic model by proposing few novel models. This paper claimed that these models contributed to boost the system performance further.

EXTRACTIVE SUMMARIZATION METHODS

To perceive different methods used for extractive summarization, it is preferable to understand the primary stages of extractive summarization approach. Predominantly extractive approach is divided into three primary stages. These stages are interdependent that is output of previous stage serves as an input to next stage.

1. Intermediate input text representation.
2. Calculating sentence ranks/scores.
3. Generating summary.

Intermediate Input Text Representation

Input text document is considered *raw* until it is preprocessed and then transformed into particular format. In order to apply scoring algorithms raw input needs to be transformed into scoring algorithm specific representation. Frequency based algorithms consider frequent words as keywords; therefore, text segmentation takes place at word level. And segmented input text is transformed into a table representation, containing words and their corresponding frequencies. Likewise sentence length, sentence position, etc. based algorithms segment input text at sentence level therefore, represents each sentence as an indicator. Graph based algorithms may represent entire input document as set of interconnected sentences.

Calculating Sentence Ranks/Scores

After transforming input text into certain intermediate representation, sentence ranking algorithms are applied to it in order to assign scores to the sentences. TF-IDF or tf-idf (term frequency-inverse document frequency), Sentence position, Sentence length, Proper noun, Word co-occurrence and Lexical similarity are some of the existing sentence scoring algorithms. The sentence scores assigned determine the importance of the sentences; highly scored sentences have greater chances of being selected for the summary.

Generating Summary

In the last phase, linear combination of highly ranked sentences forms a summary. Size of the summary is apparently less than the size of the input text document. In this phase similarity check algorithms can be employed in order to remove redundancy in the summary.

The subsequent section presents the main sentence ranking methods employed for extractive summarization approach. The various sentence ranking methods are widely categorized as statistical methods and semantical methods [16].

Statistical Methods

The methods most widely used in literature for extractive summarization approach are statistical methods. Statistical methods operate by observing statistics (like number of words, probability of a particular word, tf-idf, etc.) of the input text document, in order to identify salient sentences. These methods do not consider the meaning or sense of the words, phrases or sentences contained by the input text document. The methods are described below.

Word Frequency Method

The concept of word frequency is quite old that was unfolded by *Luhn*. According to this method, frequency of each word is recorded and the sentences are ranked on the basis of recorded frequencies. Sentence rank is incremented for every frequent word that occurs in the sentence. Thus, sentences containing most frequent words are said to be salient.

Term Frequency-Inverse Document Frequency Method

The problem with simple frequency method is that preposition, determiners and domain specific words always acquire highest frequency counts. These words do not play any role in determining the importance of the sentence instead they could affect the consistency of the summary. TF-IDF method eliminates the effect of these words by comparing the frequency of each word (*f (w)*) in the input document with its frequency in all the background documents (*bg (w)*).

$$TF_i * IDF_i = f(w) * \log\left(\frac{bg}{bg(w)}\right)$$

TF_i, is the term frequency, IDF_i is the inverse document frequency (*where i indicates the* i^{th} *word in the input*) and *bg* is the total number of background documents taken.

Sentence Length Method

Length of a sentence is the total number of words in the sentence [17]. For the optimal selection of sentence, this method constrains the shortened and lengthy sentences.

Upper Case Method

This method tries to identify the important words by assigning higher scores to the words containing upper case alphabets [17]. This method accounts the importance of acronyms, initials and proper names.

Sentence Position Method

Position of the sentence w.r.t. the entire input document is used as a criterion in this method to indicate sentence importance [4]. In the paper of Abuobieda et al. [18], leading sentence in the document is considered as important sentence and therefore, should be the candidate of the final summary.

Cue-Phrase Method

This method identifies summary sentences on the basis of cue-phrases (in particular, salient, the best, hardly, the most important, according to the literature, etc.) present in the input sentences [19].

Proper Noun Method

Proper noun method assigns higher scores to those sentences that contain one or more than one proper nouns [20].

Numerical Data Method

Important information like bank transactions, amount, balance, event date, time, etc. is always numerical. This method treats those sentences important ones that incorporate numerical data [21].

Sentence–Title Similarity Method

This method checks the similarity of the sentence with the title of the document. Thus sentences similar to the title become summary candidates.

Semantical Methods

The summarizers that use statistical methods for extraction of salient information, to some extent fail to generate coherent summaries as they do not explore the meaning of the input text. Semantical methods like *emotion* used by Khurshid et al. [16] generate rational summaries by understanding the sentiment or emotion of every sentence in the input document (Table 2) [22, 23].

ABSTRACTIVE SUMMARIZATION METHODS

Methods for abstractive text summarization are broadly categorized as structure-based methods and semantic-based methods (Table 3).

Structure-based Methods

Structure-based methods represent input document using structures like trees, templates, cognitive schemas etc. Important information is then encoded from these structures. Various structure-based methods are as follows.

Tree-based Method

In this method, input text document is represented as a dependency tree. Important information is identified by applying different algorithms like theme intersection, etc. Finally, for summary generation language generators are used (Figure 1) [24].

Template-based Method

In this method, input text document is represented as a template. For mapping text snippets into template slots, extraction rules or linguistic patterns are used. Important data are indicated by text snippets (Figure 2).

Table 2: *Presents a Comparative Study on Various Extractive Text Summarization Methods.*

Authors	Year	Input	Methods	Results
Lloret and Palomar	2009	Single document	Word frequency	The system performance was improved by 10% over DUC 2002
Gupta et al.	2011	Single document	Word frequency, cue-phrase and sentence position	The deficiently connected sentences were removed from the summary which resulted in more coherent summary
Kulkarni and Prasad	2010	Single document	Word frequency, cue-phrase, numerical data and sentence–title similarity	This system performed semantically better than the MS-word Summarizer
Abuobieda, Salim, Albaham, Osman, and Kumar	2012	Single document	Word frequency, sentence length, sentence position, numerical data and sentence–title similarity	Results in optimal features selection for the summarization process
Satoshi et al.	2001	Single document	TF-IDF, sentence position and sentence–title similarity	Having compression ratio equal to 10% the system obtained better results compared to lead-based and TF-based systems
Murdock	2006	Single document	TF-IDF	Shows that approaches for language modeling that employ Statistical Translation Models are ineffective
Fattah and Ren	2009	Single document	Proper noun, sentence length, sentence position, numerical data and sentence–title similarity	Promising results were achieved when the system was tested at different compression rates
Barrera and Verma	2012		Sentence Position	System outperformed baseline and was evaluated on DUC 2002 and set of articles from scientific magazine
Bhat et al.	2018	Single document	TF-IDF, cue-phrase, proper noun, sentence length, sentence position and semantical method "*Emotion"*	Having compression ratio equal to 35% the system obtained better results compared to MS-word generated summaries and used DUC 2007 for evaluation

Table 3: *Comparative Analysis on Various Abstractive Text Summarization Methods.*

Authors	Year	Input	Methods	Results
Barzilay and McKeown	1999	Multiple documents	Tree based	The system was able to correctly identify 74%, 69%, 74% and 56% of predicate-argument structures, the subjects, the main verbs, and the other constituents in the list, respectively
Barzilay and McKeown	2005	Multiple documents	Tree based	Summary that is grammatically strict
Harabagiu and Lacatusu	2002	Single and Multiple Documents	Template based	Evaluated GISTEXTER using DUC 2002 and resulted in coherent and organized summary
Lee and Jian	2005	Single document	Ontology based	Showed that for summarization of news articles news agent operates effectively
Tanaka and Kinoshita	2009	Single document	Lead and body phrase	Operations such as insertion and substitution were performed on phrases
Genest and Lapalme	2012	Multiple documents	Rule based	Results in high density information summary
Greenbacker	2011	Multimodal Document (both text and images)	Multimodal semantic model	Abstract summaries that incorporate concepts obtained by graphical data
Genest and Lapalme	2011	Multiple documents	INIT based	Evaluated system using TAC 2010; average performance was satisfactory
Moawad and Aref	2012	Single document	Semantic graph based	Reduced input text document to almost 50%

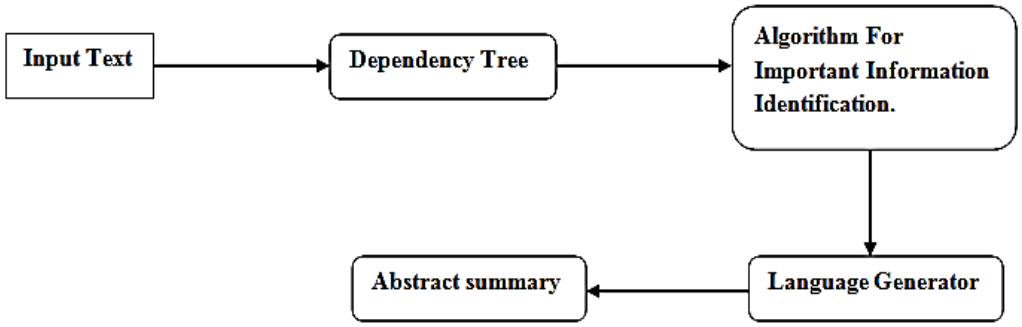

Fig. 1: *Block Diagram of Tree-based Method.*

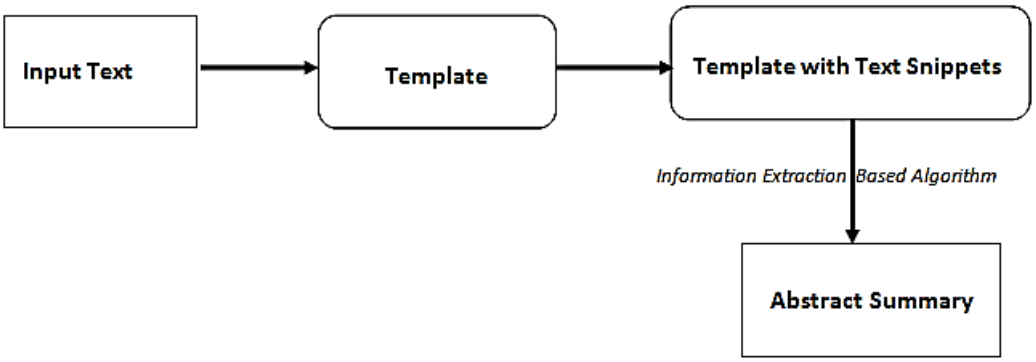

Fig. 2: *Block Diagram of Template-based Method.*

Ontology-based Method

The domain related documents can be coherently summarized by Ontology-based methods because ontology can better represent a domain as each domain possesses knowledge structure [25].

Lead and Body Phrase Method

This method tries to rewrite the lead sentence by substitution or insertion of phrases that have similar triggers in the lead and body sentences (Figure 3).

Rule-based Method

In this method, input text document is represented as list of aspects and categories. Information extraction rules are used to generate candidates. Then the best candidates are selected by content selection module. Finally, summary is generated using generation patterns.

Semantic-based Methods

Semantic-based methods transform input document into semantic representation. This intermediate representation is then supplied to the natural language generation (NLG) system. NLG system processes the linguistic data in order to identify noun phrases and verb phrases.

Multimodal Semantic Method

This method takes as input in the form of both text as well as images. This multimodal input document is represented by a semantic model, which clearly apprehends the concepts and the relationship among them. Some measures are used to score the important concepts. Finally, concepts opted for summary are framed as sentences [26].

Information Item-based Method

In this method, the information in the input text is transformed as abstract representation. From this abstract, representation contents of summary are selected.

Semantic Graph Based Method

In this method, the input text document is converted into rich semantic graph (RSG). Then reduction of the RSG takes place. Finally, this reduced RSG acts as the basis for final abstract summary generation (Figure 4) [27].

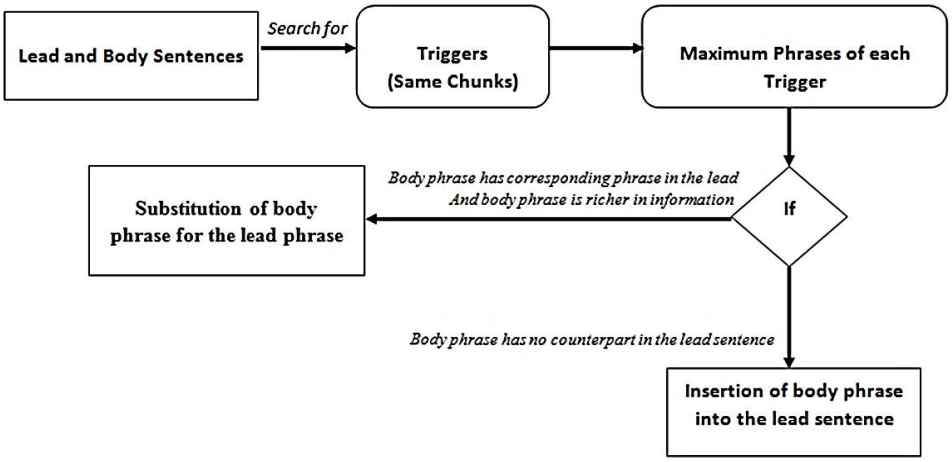

Fig. 3: *Block Diagram of Lead and Body Phrase Method.*

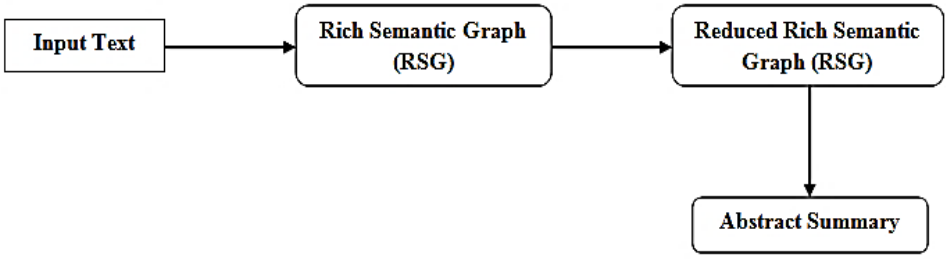

Fig. 4: *Block Diagram of Semantic Graph Based Method.*

CONCLUSION

Manual text summarization entails considerable number of impartial and knowledgeable experts, and lot of time. However, digital data which is accessible world-wide is increasing exponentially thus makes utilization of an automatic text summarization tool inevitable in order to achieve coherent summaries in less time. Automatic text summarization approaches are widely categorized as extractive approaches and abstractive approaches. This paper presents a review on both extractive as well as abstractive approaches. Different extractive and abstractive methods are explored and a comparative analysis between different methods implemented in literature is presented.

REFERENCES

1. Aliguliyev RM. Automatic document summarization by sentence extraction. *Computational Technologies.* 2007;12(5): 5–15p.
2. Hovy E, Lin CY. Automated text summarization and the SUMMARIST system. *Proceedings of a workshop on held at Baltimore*, Maryland: October 13–15, 1998, 197–214p.
3. Hovy E, Lin CY. Automated Text Summarization in SUMMARIST. In *Proceedings of the Workshop of Intelligent Scalable Text Summarization*, July 1997.
4. Edmundson HP. New methods in automatic extracting, *J ACM*. 1969;16(2):264–285p.
5. Luhn HP. The automatic creation of literature abstracts. *IBM J Res Dev*. 1958;2(2): 159–165p.
6. Nenkova A, McKeown K. Automatic summarization. *Found Trends® Inf. Retr*. 2011;5(3):235–422p.
7. Dalianis H, Hovy E. Aggregation in Natural Language Generation. In Adorni, G. & Zock, M. (Eds.), Trends in Natural Language Generation: an Artificial Intelligence Perspective, *EWNLG'93, Fourth European Workshop, Lecture Notes in Artificial Intelligence, No. 1036*, 1996, 88–105p.
8. Lin CY, Hovy E. Identify Topics by Position, *Proceedings of the 5th Conference on Applied Natural Language Processing*. March 1997.

9. Ouyang Y, Li W, Li S, Lu Q. Applying regression models to query-focused multi-document summarization. *Info Process Manag*. 2011;47(2):227–237p. doi:10.1016/j.ipm.2010.03.005.
10. Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, August, 335–336p, ACM. doi:10.1145/290941.291025.
11. Vishal Gupta, Gurpreet Singh Lehal, A survey of Text summarization techniques. *J Emerg Technol Web Intell*. August 2010;2(3):258–268p.
12. Chang T, Hsiao W. A hybrid approach to automatic text summarization, *8th IEEE International Conference on Computer and Information Technology (CIT 2008)*, Sydney, Australia, 2008.
13. Chengcheng L. Automatic Text Summarization Based on Rhetorical Structure Theory. *IEEE. 2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, 2010.
14. Yousefi-Azar Mahmood, Len Hamey. Text summarization using unsupervised deep learning. *Expert Syst Appl.* 2017;68:93–105p.
15. Nallapati Ramesh, *et al.* Abstractive text summarization using sequence-to-sequence RNNs and beyond. 2019. *Available from* https://arxiv.org/abs/1602.06023.
16. Bhat Iram Khurshid, Mudasir Mohd, Rana Hashmy. SumItUp: A Hybrid Single-Document Text Summarizer. *Soft Comp Theor Appl*. Springer, Singapore, 2018, 619–634p.
17. Kupiec J, Pedersen J, Chen F. A trainable document summarizer, *in Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1995; 68–73p.
18. Abuobieda A, Salim N, Albaham AT, Osman AH, Kumar YJ. Text summarization features selection method using pseudo genetic-based model. *In International Conference on Information Retrieval Knowledge Management*, 2012, 193–197p.
19. Lee CS. *et al.* A fuzzy ontology and its application to news summarization, *IEEE T Syst Man Cy Part B.* 35, 2005, 859–880p.
20. Barzilay R. *et al.* Information fusion in the context of multi-document summarization. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics.* 1999; 550–557p.
21. Harabagiu SM, Lacatusu F. Generating single and multi-document summaries with GISTEXTER. *Document Understanding Conferences.* 2002.
22. Barzilay R, McKeown KR. Sentence fusion for multidocument news summarization. *Comput Linguist.* 2005;31:297–328p.
23. Genest PE, Lapalme G. Fully abstractive approach to guided summarization. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers.* 2012;2: 354–358p.
24. Greenbacker CF. Towards a framework for abstractive summarization of multimodal documents. *Proceedings of the ACL 2011 Student Session*. 2011;75–80.
25. Tanaka H. *et al.* Syntax-driven sentence revision for broadcast news summarization. *Proceedings of the 2009 Workshop on Language Generation and Summarisation.* 2009; 39–47p.
26. Genest PE, Lapalme G. Framework for abstractive summarization using text-to-text generation, in *Proceedings of the Workshop on Monolingual Text-To-Text Generation.* 2011, 64–73p.
27. Moawad IF, Aref M. Semantic graph reduction approach for abstractive Text Summarization. *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on*, 2012; 132–138p.

Cite this Article

Basit Farooq, Amjad Hussain, Semantic Summarizer. *Journal of Artificial Intelligence Research & Advances.* 2019; 6(3): 1–8p.