Literature Study on Text Summarization

Anubhav Mishra[1] Mrinalini Sawhney[2] Akshita Sharma[3]
[1,2,3]Student
[1,2,3]Department of Computer Engineering
[1,2,3]Bharati Vidyapeeth's College of Engineering A-4, Paschim Vihar New Delhi

Abstract— a summary is a concise representation of the detailed text, which conveys the general idea of the original text with least redundancy. The manifold increase in information on the Web demands intelligent summarization tools to present an abridged version of the source text. The goal of automatic text summarization is to represent the source text in a shorter version, rich in information and with correct semantics. This paper discusses various summarization techniques based on two distinctions: extractive and abstractive. Furthermore, the merits of graph-based text summarization is explored with the detailed analysis of Opinosis.

Key words: Extractive Summary, Abstractive Summary, Opinosis

I. INTRODUCTION

Data is ever increasing and with today's scenario the abundance of this information is great inconvenience to the readers. It becomes a tedious task to remove the redundancy in the text and extract the relevant content. The gist of the matter often gets subdued in the vastness of the information it carries. This arise the need to express the information in a concise and optimised format with summarised representation of the text. Popular approaches for the task focus on finding keywords and ranking them on the specified metric. This achieved through various methods comprised within extractive and abstractive techniques.

Section 1 discusses various extractive approaches of summarisation and their limitations which paved a way for a better method with abstractive techniques. Section 2 discusses these in detail. Section 3 focuses on Graph-based summarisation framework-Opinosis, which imbibes the advantages of both extractive and abstractive techniques minimising their shortcomings.

II. EXTRACTIVE TEXT SUMMARISATION

Extractive summarization is a statistical approach for text summarisation which focuses on selecting a subset of the keywords, important sentences, phrases or paragraphs from the original text and concatenating them together to form the summary. The selection of text fragments is based on statistical and linguistic features of sentences (such as frequency of words or phrases, location and cue words). This approach however results in sentences that are longer than the average length and does not provide in-depth understanding of the text.

Compression rate is a vital factor in the extraction method of building summaries which defines the ratio of the length of the summary and the original text. Summaries with high compression rates tend to be longer and hence contain insignificant content, while those with lower compression rate are shorter but fail to incorporate vital information. The text resulting from a compression rate of 5-30% is acceptable.

A. Term frequency- inverse document frequency based approach

This model makes use of weighted term frequency and inverse sentence frequency standard for the retrieval of keywords. Weighted term frequency is the measure of the occurrence of a word w in the document, while the inverse sentence frequency corresponds to the number of sentences in the document that contain that word. Tf-idf determines statistically the importance of a word in a document. Higher weights are assigned to the most descriptive words which occur very frequently in a sentence but rarely appear in the whole document, e.g. Punctuations. These scores are then compiled with reference to the query and the sentences with the highest scores are concatenated together to form the summary.

B. Cluster based Approach

Clustering is a method of grouping similar content, built around a common theme. This incorporates the variety of ideas which might appear in the document. It takes the clustered sentences as input and makes use of key phrase extraction techniques like term frequency inverse document frequency (TF-IDF) to assign ranking to sentences on the basis of the important keywords contained in them. The overall score of a sentence is based on similarity of the sentences to the theme of a particular cluster, the location where the sentence takes place in the document, and the similarity of the given sentence to the first sentence in the corresponding document.

C. Centroid-based approach

MEAD is a publicly available platform for multi-lingual summarization and evaluation. It performs summarisations based on multiple algorithms such as position-based, centroid-based, largest common subsequence and keywords. One of the popular methods defined in MEAD, based on sentence extraction, is the centroid based technique. MEAD uses cluster centroids to group together sentences/phrases that are recurring in multiple articles. Furthermore, the phrases are scored by linearly combining the three default features of MEAD: centroid score (measures the centrality of the sentence to the theme of a cluster), position score (depends on the position of a sentence from the beginning of the document) and first sentence overlap score (product of TFIDF score of a given sentence and the first sentence). The quality of the summaries can be evaluated by intrinsic (such as percent agreement, cosine similarity, and relative utility) and extrinsic (document rank for information retrieval) methods.

D. Graph-based approach

It moulds the document in an undirected graph and the sentences as nodes of the graph. An edge exists between two sentences if they contain common phrases. Every isolated sub

graph is a partition of the distinct themes covered in the text. Summary can be built around a specific theme or chosen from each of the sub graphs to present a general view. Furthermore, the important sentences within a sub-graph can be identified by their cardinality (number of edges connected). Higher the cardinality, higher would be the preference of that sentence to be included in the summary. Graph based approaches such as TextRank model any document into a graph and rank the elements in the graph based on the similarities in the content of the sentences. These are inspired by Google's PageRank, an algorithm to rank web pages on the web.

E. Fuzzy logic based approach

Fuzzy logic system proposes extraction of sentences using fuzzy rules and creation of a fuzzy set based on features such as sentence length, sentence location, similarity to keyword, sentence to sentence similarity, word frequency, number of proper nouns and so on. The performance of the fuzzy system is largely affected by the choice of fuzzy rules and membership functions. The knowledge base of the fuzzy system is built using these rules, providing an input to the fuzzy system along with the membership function which divides the features into sets of unimportant, median and important values. The inference engine works on the fuzzy IF-THEN rules, extracting the important sentences for inclusion in the final summary. Output from the inference engine is defuzzified into the final score of the sentence.

Extracted summaries are seen to be longer than average, due to which unimportant segments of the test get included. If the summary isn't verbose, it will fail to capture the relevant information spread across the document. Extraction methods focus on extraction of keywords, sentences, or phrases as they are, without any modification in the linguistic construction of the sentence. Sentences containing pronouns may be extracted out of context. This results in incoherent summaries.

Extractive Summarization	Abstractive Summarization
In this method, the summary of the reviews is a collection of tangible sentences or phrases.	The method of abstractive summarization discards original sentences and creates summary in coherent, non-redundant manner.
The frequency and location of the words in sentence decides its presence in the final output.	Important phrases and word chains from the document are identified using natural language processing tools and are bound together into meaningful sentences.

Table 1: Difference between Extractive and Abstractive Text Summarisation

III. ABSTRACTIVE TEXT SUMMARISATION

Most of the work in text summarization has focused on extractive summarization, which forms summary by selection of important sentences from the documents. Abstractive summarization methods extract phrases and lexical chains from the documents by using language understanding tools to generate a summary. The process of abstraction is complex with refining and reforming summary with heuristic approach.

These techniques are broadly classified into two categories: Structured based approach and Semantic based approach.

A. Structured based approach

Structured based approach encodes most important information from the document through cognitive schemas such as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure.

1) Tree based method

The tree-based method uses a dependency tree for a document to visualise its contents It makes use of a language generator combined with an algorithm to summarise the text. Initially, the similar sentences are pre-processed with the use of a shallow parser, and eventually, sentences are mapped to a predicate- argument structure. Theme-intersection algorithm is used next, to determine the common link judging the predicate- argument structure. This is performed using a content planner. Common links of informative phrases are selected and finally a language generator is used to combine them to form new summary sentences using a sentence generator. This method is advantageous in the use of language generation of improved resultant output consequently reducing redundancy and maintaining fluency of the content. Importance of context supersedes the intersection of phrases in this approach.

Other method in this approach aims to find the centroid of the dependency tree which is then augmented with sub-tress and finally prunes the predefined constituents.

2) Template based method

A template of the whole document is presented in this technique where patterns based on linguistics and rules of extraction are put in line to identify snippets of text and are subsequently mapped to the defined slots of template. The snippets are identified as the part of the summary content. Extraction of information from documents in the template format gives a set of related concepts on the topic. The frame of the template containing slots, are filled with information extracted by the systems. Informative and articulated summaries are obtained using this technique as it relies only on the relevant extracted information. However, the summarisation of multi-documents cannot be handled with this approach as similarity between various documents may not always be found.

3) Ontology based method

Researchers have preferred ontology (knowledge base) process for summarisation. The knowledge structure present in various documents is exhausted to be represented with this approach with fuzzy concepts implementation and pre-processing of meaningful terms are performed in this corpus. A classifier then sifts the basis of news and presents a summary based on fuzzy ontology. This approach may, however, be restricted to just Chinese news.

4) Lead and body phrase method

Insertion and substitution forms the basis of this method where same syntactic head chunk are looked for in head and body sentences. Triggers are searched in the sentence and maximum phrases of each trigger are recognised using the similarity metric. Substitution and insertion into the body phrase has information rich context in the summary.

5) Rule based method

With this method, summary of the documents are presented in terms of sub-categories. Selection of module-content gives the best possible candidate with the defined extraction rules. Further, generation patterns are used to determine the complete abstract from the cluster based on same event.

B. Semantic based Approach

Using Semantic based method, documents are represented for semantic analysis using the Natural Language generator (NLG). The noun-phrases and verb-phrases are the linked linguistic nodes of a domain-specific ontology. Resultant information is used to convert regions into semantic representation. Few methods of semantic approaches are as follows.

1) Multimodal Semantic Model

In the method of multimodal semantic, relationships with the concepts are searched and the important ones are rated based on some metric to form the summary. A multimodal document is a combination of text and images. A three step process is followed in this approach: construction of semantic model with knowledge representation of objects; rating informational content determining the relevance of and occurrence in current document; final summarising sentences. The final summary produced using this approach has excellent coverage including all salient and graphical information.

2) Information item based method

Information item in this approach is the abstract representation of the document and not the original text. Information item retrieval is done through syntactic analysis of text and verb's subjects are extracted. Followed by sentence generation, the method uses NLG realiser and ranks the sentences based on Average Document Frequency score. Finally, a summary generation step includes all the aspects of dates and location to generate a coherent and information-rich summary.

3) Semantic Graph based Method

This approach uses the Rich Semantic Graph of the original document. RSG updates the verbs and nouns as graph nodes and edges represent semantic and topological relationships between them. In the subsequent phases RSG is reduced using heuristic rules providing a path to spawn abstractive summary. The scalability and less-redundancy of the grammatically-correct sentences is the major strength of this technique of summarisation.

Author	Type	Technique	Description
Zhang, Milios, Zincir-Heywood, 2007	Extractive	Term frequency- inverse document frequency based	This method uses weighted tem frequency for summary retrieval through keywords.
Alguliev and Aliguliyev, 2009	Extractive	Cluster based	This method uses clustering of similar content using key-phrase extraction.
Radev et al, 2004	Extractive	Centroid-based	The cluster of centroid phrases are grouped together with their score ensuring quality summaries.
Mihalcea and Tarau, 2004	Extractive	Graph-based	An undirected graph is formed between phrases and summary is generated based on cardinality between edges.
Suanmali and Binwahlan, 2009	Extractive	Fuzzy logic based	Fuzzy rules and set determine the important median in sentences used in summary generation for this method.
Barzilay and McKeown, 1999	Abstractive	Tree based	It uses a dependency based representation with an algorithm to summarise the text.
Harabagiu and Lacatusu ,2002	Abstractive	Template based	The template frames use linguistic patterns for content selection.
Lee and Jian, 2005	Abstractive	Ontology based	This method uses the fuzzy concept representation.
Tanaka and Kinoshita, 2009	Abstractive	Lead and body-phrase	The lead body structure searches for revision candidates for content selection.
Genest and Lapalme, 2012	Abstractive	Rule based	The technique presents documents in sub-categories with extraction rules for generating patterns.
Moawad and Aref, 2012	Abstractive	Semantic graph based	This method uses natural language generator with noun-phrases and verb phrases linked in domain specific manner.

Table 2: Various Methods of Text Summarisation

IV. OPINOSIS

Opinosis is a graph based method to generate flat abstractive summaries of highly verbose data-sets like user reviews, opinions or text. Opinosis summaries have better correlation with human generated summaries than the baseline extractive methods. The produced summaries are comprehensible, concise and reasonably well formed.

Opinosis works by first creating a graphical representation (called the Opinosis graph) of the textual data to be summarized, which is then used to identify common paths and subpaths to generate possible summaries. A scoring measure is then used to rank and select the best summary from the list of various possible summaries.

A. Opinosis Graph

Opinosis graph is a graph structure which is used to represent the verbose data-set text in a graphical format to convert the problem to abstractive summarization to that of appropriate path finding in a graph.

Opinosis graph consists of nodes that represent word-units and directed edges representing structure of the sentences. Each node also contains the positional reference

information (PRI) related to each word-unit that outlines a word-unit's membership in a sentence as well its position within it.

B. Summarisation Scheme

The generation of summary is done through constant searching of the opinosis graph to find paths that represent valid sentences and that have high redundancy scores. The identified sentences are the used to generate the flat abstractive summary.

The summarization scheme has following parts to it:

1) Valid Paths

Paths that correspond to a relevant sentence are called valid paths. Valid Paths are identified using following:

- Valid Start Node - VSN: Nodes that correspond to words that are natural starting points in a sentence.
- Valid End Node - VEN: Nodes that correspond to words that complete or conclude a sentence.
- Valid Path: A path is called a valid path if it is connected by a set of directed edges and starts at a VSN, ends at a VEN and follows a set of predefined POS (Parts of Speech) rules.

2) Path Scoring

Path scoring is done to identify the path with the highest redundancy score. A Path that can represent majority of the verbose opinions or text is selected to generate the abstractive summary.

3) Collapsed Paths

Certain paths in the Opinosis graph are collapsible. In such cases, the scoring is done after the collapsing of paths.

- Collapsible Node: Any node is a possible collapsible node if it represents a verb.
- Collapsed Candidates: Paths continued after the collapsed node are referred to as its collapsed candidates.
- Anchor: The sub-path formed by the nodes just before the collapsed node is called the anchor.
- Stitched Sentence: A proper sentence formed by combining the Anchor and the Collapsed candidates.
- Collapsed Path Score: The average of the path score of all individual sentences of Collapsed candidates is called the Collapsed Path Score.

4) Summary Generation

The Generation of summary is done in three steps:

- All the Paths are ranked in the decreasing order of their path scores.
- All the Duplicate paths are removed.
- The top few paths are then taken as the generated summary.

C. Algorithm: Opinosis

1) Input: Sentences to be summarized S = $\{s\}^{n}_{i=1}$
2) Output: O = {Opinosis summaries}
3) For i = 1 to n do
4) w ← Split(s_i)
5) word_count ← SizeOf(w)
6) For j = 1 to word_count do
7) LABEL ← w_j
8) PID ← j
9) SID ← i
10) FindOrCreate(G, LABEL)
11) If NewCreated() then
12) PRI_{vj} ← (SID, PID)
13) Else if Matched() then
14) PRI_{vj} ← PRI_{vj} U (SID, PID)
15) End if
16) FindOrCreate(NODE,$v_{j-1} \rightarrow v_j$, G)
17) If NewCreated() then
18) NODE.count = 1
19) Else if Matched() then
20) NODE.count = NODE.count + 1
21) End if
22) End for
23) End for
24) g← Graph(G)
25) node_count← SizeOf(g)
26) For j = 1 to node_count do
27) If VSN(v_j) then
28) newLen← 1
29) score← 0
30) newList← CreateNewList()
31) Traverse_dfs(newList, v_j, score, PRI_{vj}, $label_{vj}$, newLen)
32) candidates← {candidates U newList}
33) End if
34) End for
35) C← RemoveDuplicates(candidates) && SortBy(C, PathScore)
36) For i = 1 to e_{ss} do
37) = { O U PickNextBestCandidate(C) }
38) End for
39) Routine Traverse_dfs(list, v_k, score, $PRI_{overlap}$, sentence, len):
40) redundancy← SizeOf($PRI_{overlap}$)
41) If redundancy >= e_r then
42) If VEN(v_k) then
43) If ValidSentence(sentence) then
44) finalScore← score/len
45) AddCandidate(list, sentence, finalScore)
46) End if
47) End if
48) For v_n in Neighbours v_k do
49) PRI_{new} ← $PRI_{overlap}$ n PRI_{vn}
50) redundancy← SizeOf(PRI_{new})
51) newSent← Concat(sentence, $label_{vn}$)
52) L← len + 1
53) newScore← score + PathScore(redundancy, L)
54) If Collapsible(v_n) then
55) C_{anchor}← newSent
56) tmp← CreateNewList()
57) For v_x in Neighbours v_n do
58) Traverse_dfs(tmp, v_x, 0, PRI_{new}, $label_{vx}$, L)
59) CC← EliminateDuplicates(tmp)
60) CCPathScore← AveragePathScore(CC)
61) finalScore← newScore + CCPathScore
62) stitchedSent← Stitch(C_{anchor}, CC)
63) AddCandidate(list, stitchedSent, finalScore)
64) End for
65) Else
66) Traverse_dfs(list, v_n, newScore, PRI_{new}, newSent, L)
67) End if
68) End for
69) End if

V. CONCLUSION

Text summarisation offers a simple comprised gist of heavily redundant and complex data. With focus shifting from extractive summarisation approaches to abstractive summarisation approaches, a cohesive, coherent and definitive summary is generated as a result output. With all its benefits, abstractive summarisation is still a difficult arena with continuous research trying to make it more adaptive and less complex. The natural language processing is being tested and modified to get better results. The various methods are explored and compared to define their technique and if future enhancements to the same can be identified. The study aims to highlight the opinosis method which utilises the best of both the approaches with simplicity of extraction and coherence of abstraction in single retrieval. Overall, this study allows one to have a better understanding of the techniques described and be utilised to the best of their functionality.

REFERENCES

[1] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions" Proceedings of the 23rd International Conference on computational Linguistics, 2010.
[2] Y. Zhang, E. Milios, N. Zincir-Heywood, "A Comparative Study on Key Phrase Extraction Methods in Automatic Web Site Summarization", Journal of Digital Information Management, vol. 5, no. 5, 2007.
[3] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. C¸elebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel and Z. Zhu, "MEAD - a platform for multi document multilingual text summarization", LREC 2004.
[4] G. Erkan, D. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, vol. 22, pp 457-479, December 2004.
[5] D. Radev, E. Hovy and K. McKeown, "Introduction to the Special Issue on Summarization", Computational Linguistics, vol. 28, no. 4, pp. 399-408, 2002.
[6] J. Cheung, "Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection", Undergraduate, Department of Computer Science, University of British Columbia, pp. 5-8, 2008.
[7] U. Hahn and I. Mani, "The Challenges of Automatic Summarization", IEEE Computer Society Press Los Alamitos, CA, USA, vol. 33, no. 11, pp. 29-36, 2000.
[8] L. Suanmali, N. Salim, M. Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security (IJCSIS), vol. 2, no. 1, 2009
[9] S. Hariharan, "Extraction Based Multi Document Summarization using Single Document Summary Cluster", International Journal of Advances in Soft Computing and its Applications, vol. 1, no. 1, 2009.
[10] R. Sharma, P. Sharma, "A Survey on Extractive Text Summarization", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, no. 4, 2016.
[11] R. ALGULIEV, R. ALIGULIYEV, "Evolutionary Algorithm for Extractive Text Summarization", Intelligent Information Management, vol. 1, pp 128-138, 2009.
[12] N. Munot, S. Govilkar, "Comparative Study of Text Summarization Methods", International Journal of Computer Applications, vol. 102, no. 12, September 2014.
[13] S. Saziyabegum, P. Sajja, "Literature Review on Extractive Text Summarization Approaches", International Journal of Computer Applications, vol. 156, no. 12, December 2016.
[14] A. Khan and N. Salim, "A Review on Abstractive Summarisation Methods", Journal of Theoretical and Applied Information Technology, vol. 59, no. 1, January 2014.
[15] V. Gupta and G. Lehal, "A Survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, vol. 2, no. 3, August 2010.
[16] D. Vidyadharan and A. CR, "A survey on Various Summarisation Techniques", International Journal of Engineering and Computer Science, vol. 3, no. 12, pp. 9528-9532, December 2014.
[17] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.
[18] R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," Computational Linguistics, vol. 31, pp. 297-328, 2005.
[19] S. M. Harabagiu and F. Lacatusu, "Generating single and multi-document summaries with gistexter," in Document Understanding Conferences, 2002.
[20] C.-S. Lee, et al., "A fuzzy ontology and its application to news summarization," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 35, pp. 859-880, 2005.
[21] H. Tanaka, et al., "Syntax-driven sentence revision for broadcast news summarization," in Proceedings of the 2009 Workshop on Language Generation and Summarisation, pp. 39-47, 2009.
[22] P.-E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, vol.22, pp. 354-358, 2012.
[23] I. F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive Text Summarization", Computer Engineering & Systems (ICCES), pp. 132-138, 2012.
[24] Y. Zhang, N. Zincir-Heywood, and E. Milios, "Narrative text classification for automatic key phrase extraction in web document corpora" In Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 51-58, November 2005.
[25] R. Mihalcea, and P. Tarau, "TextRank: Bringing order into texts" In Proceedings of the EMNLP-04 and the

2004 conference on empirical methods in natural language processing, July 2004.

