International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Survey of Relevant Text Content Summarization Techniques

Priya Ganguly[1], Dr. Prachi Joshi[2]

MIT College of Engineering, Survey No. 124, MIT College Campus, Ex-Serviceman Colony, Kothrud, Pune 411038 India

Abstract: ***With lots of information getting generated every day, document summarization is becoming essential. Instead of having to go through the entire text, it is convenient to understand the text fast and easily by a relevant summery. [1] Text summarization is the procedure of automatically making a shorter version of one or more text documents. It is a significant method of detection related material in huge text libraries or in the Internet [2]. It is also important to extract the information is such a manner that the content would be of interest of the user. This paper covers various techniques that are used for relevant content summarization***.

Keywords: Text Summarization, extractive summary, information extraction, relevant content.

1. Introduction

Text summarization is gaining much significance currently. One reason for this is, recently due to the enormous growth in material, requirement for involuntary text summarization has enlarged. It is very difficult for human beings to manually précisbig documents of text. There is an profusion of text material available on the internet. However, usually the Internet offers more material than is required. Therefore, a twofold problem is encountered: examining for relevant documents through a crushing amount of documents existing, and absorbing a large quantity of relevant information [3]. The aim of involuntary text summarization is reducing the source text into a shorter form preserving its information content and overall meaning. With a big volume of documents, giving the operator with a summary of each document greatly facilitates the job of finding the relevant and desired documents. The main goal of a summary is to present the key concepts in a document in less space. If altogether sentences in a text document were of equivalent significance, creating a summary would not be very effective, as any decrease in the dimension of a document would carry a relational decrement in its informativeness. Luckily, information content in a document looks in gusts, and single can therefore differentiate between more and less informative segments. Identifying the helpful and relevant segments at the cost of the rest is the main challenge in summarization. A variety of document summarization methods have been established newly. The paper [4] reviews research on automatic summarizing over the last decade. This paper appraisals striking notions and developments, and pursues to measure the state-of-the-art for this interesting natural language processing (NLP) task. The appraisal displays that some useful summarizing for various dedications can already be done but also, not amazingly, that there is a enormous amount more to do. Sentence based extractive summarization methods are usually used in involuntary summarization to yield extractive summaries. Systems for extractive summarization are usually based on method for sentence extraction, and attempt to identify the set of sentences that have greatest meaning for the general understanding of a given document.

In this study, we have encountered various techniques for sentence based extractive summarization, various similarity measures and their comparisons.

2. Literature Survey

2.1 Feature for Extractive Text Summarization

2.1.1 Content word / Keyword feature:

Keywords or content words [8][9][10] are principally nouns and determined using Tf-Idf measure. Sentences having keywords have a lot of probabilities to be concerned in outline. Another keyword extraction technique is given below, having 3 categories:

1) Morphological Analysis
2) phrase Extraction and rating
3) phrase bunch and rating

2.1.2. Title word feature:

Sentences having words that seem within the title also are suggestive for understanding the essence of the document. These sentences are having a lot of probabilities for obtaining enclosed in outline.

2.1.3. Sentence location feature

Generally, 1st and last sentence of first and last paragraph of a text document are a lot of vital and have high likelihood to be enclosed in outline.

2.1.4. Sentence Length feature

Very massive and really little sentences are typically not concerned in outline.

2.1.5. Proper Noun Feature

Proper noun is name of someone, place and construct etc. Sentences having relevant nouns have larger probabilities for obtaining enclosed in outline.

2.1.6. Upper-case Word Feature

Sentences having acronyms or correct names ar enclosed.

2.1.7. Cue-Phrase Feature

Sentences that contain phrases that signifies some signals as an example ~~–in~~ conclusion", ~~–th~~is letter", ~~–th~~is report",

–summary",–discussed", etc. have the very best likelihood to be contained in summaries.

2.1.8. Biased Word Feature
If a word that's gift in a very sentence is from biased glossary, then that sentence is of larger significance and vital. Biased glossary is antecedently outlined and will have domain specific words.

2.1.9. Font based feature
Sentences having words trying in majuscule, bold, italics or Underlined fonts are typically a lot of vital.

2.1.10. Pronouns
Words like –he, she, they" don't seem to be enclosed within the document outline unless they're reworked into individual nouns.

2.1.11. Sentence-to-Sentence Cohesion:
For every sentence s the similarity between s and every alternative sentence s' of the document is calculated. Then those likeness values are escalated, getting the raw worth of this feature for s. the method is frequent for each sentences.

2.1.12. Sentence-to-Centroid Cohesion:
For each sentences vector is computed that signifies the center of mass or centroid of the document. It is the arithmetic average over the consistent coordinate worth of all the sentences of the document; then the similarity between the centroid and every sentence is calculated to realize the raw value of the options for every sentence.

2.1.13.Occurrence of non-essential information:
Some words indicate non-essential data. These words ar speech pointers like –because", –furthermore", and –nevertheless" etcThese words typically occur at the beginning of a sentence. This is often likewise a binary feature, assigning worth –true" if the sentence contains a minimum of one in every of these markers, and –false" otherwise.

2.2 Extractive Summarization Methods

Extractive summarizer aims at choosing out the foremost relevant sentences within the document whereas conjointly maintaining a reduced redundancy within the outline.

2.2.1. Term Frequency-Inverse Document Frequency (TFIDF) approach
Bag-of-words model is made at sentence level, with the traditional biased term-frequency and opposite sentence frequency paradigm, wherever sentence-frequency is that the range of sentences within the document that have that term. These sentence vectors area unit then scored by similarity to the question and therefore the most marking sentences area unit designated to be a part of the outline. to get a generic outline, continuous-words that occur most frequently within the document(s) is also taken as the question words. Since these words represent the theme of the document, they manufacture generic summaries. Term frequency is typically zero or one for sentences—since unremarkably the equal content-word doesn't appear again and again in a very given sentence. If users produce question words the approach they produce for data retrieval, rest the question primarily based outline generation would become generic summarization.

2.2.2. Clustering Based Approach
Documents area unit usually written specified they address totally different topics one when the opposite in associate degree organized vogue. they're sometimes uneven expressly or implicitly into sections. Documents area unit portrayed mistreatment term frequency inverse document frequency (TF-IDF) of various words. Term frequency employed in this context is that the average range of existences (per document) over the cluster. IDF price is computed supported the whole corpus. The summarizer takes antecedently clustered documents as input. every cluster is measured an issue. The theme is portrayed by words with high ranking term frequency, inverse document frequency (TF-IDF) scores in this cluster. Sentence choice is predicated on similarity of the sentences to the theme of the cluster Ci .The subsequent issue that's thought-about for sentence choice is that the location of the sentence within the document (Li).

2.2.3. Machine Learning Approach
Given a group of documents and their extractive summaries, the summarization method is displayed as a classification problem: sentences area unit classified as outline sentences and non-summary sentences supported the options that they maintain. The classification likelihood is that learnt statistically from the obtained information, using Bayes' rule:

P (s∈<S | F1, F2, ..., FN) = P (F1, F2, ..., FN | s∈S) *P (s∈S) / P (F1, F2,..., FN)

where s may be a sentence when the document assortment, F1,F2…FN area unit options employed in classification. S is that the outline to be generated, and P (s∈< S | F1, F2, ..., FN) is that the chance that sentence s are chosen to create the outline on condition that it possesses options F1,F2…FN.

2.2.4. Graphical Approach:
As seen within the previous strategies, the primary step concerned within the method of summarizing one or a lot of documents is recognizing the problems or topics addressed within the document. Graphical illustration of passages delivers a technique of identification of those themes. When the common preprocessing steps, namely, stop word elimination and stemming, sentences within the documents area unit portrayed as nodes in associate degree less graph. There's a node for every sentence. 2 sentences area unit joined with a footing if the 2 sentences share some common words, or in more words, their (cosine, or such) likeness is higher than some threshold. This illustration yields 2 results: The partitions restricted within the graph (that is those sub-graphs that area unit unconnected to the opposite sub graphs), type distinct topics encircled within the documents. this allows a alternative of coverage within the outline. For query-specific summaries, sentences is also appointive solely from the relevant sub graph, whereas for generic summaries, representative sentences is also chosen from every of the sub-graphs. The second result made by the graph-theoretic technique is that the identification of the numerous sentences within the document. The nodes with nice cardinality (number of edges connected to it node), area unit the

necessary sentences within the divider, and thus bring higher preference to be enclosed within the outline.

2.3 Sentence Evaluation Methods

The first mention to text summarization sentence evaluation dates back to 1958 [26][27]. As antecedently given, the main focus of those analysis area units are self-addressed by the subsequent question: however will a system fix that sentences are symbolic of the content of a given text? This approach analyzes the options of the sentence itself and was used for the primary time in 1968 [25] examining the presence of cue words in sentences. The most ways that monitor this idea area unit outlined below:

2.3.1. Cue-phrases

In general, the sentences started by _in summary‘‘, _in conclusion‘‘, _our examination‘‘, _the paper describes‘‘ and emphasizes like _the best‘‘, _the most important‘‘, _conferring to the study‘ ‘meaningfully‘‘, _important‘‘, _in particular‘‘, _hardly‘‘, _impossible‘ ‘as well as domain-precise bonus expressions terms may be sensible pointers of great content of a text document. a better score is appointed to sentences that contain cue words/phrases, mistreatment the formula:

CP =CPS/CPD

where,

CP = Cue-phrase score,
CPS = variety of cue-phrases within the sentence,
CPD = Total variety of cue-phrases within the document.

2.3.2. Sentence position

There area unit several approaches that use the sentence position as a score criterion. In reference [28], the primary sentence within the paragraph is taken into account a very important sentence and a troublesome candidate to be comprised within the summary; [29] says that the primary sentences of paragraphs and words in titles and headings area unit additional applicable to summarization; the tactic projected in reference [17] allots score one to the primary N sentences and zero to the others, wherever N may be a given threshold for the amount of sentences. Reference [15] follows constant principle as reference [17] and assume that the primary sentences of a paragraph area unit the foremost important ones. The sentence rankings area unit as follows: the primary sentence in an exceedingly paragraph contains a mark price of 5/5; the second sentence contains a mark 4/5, and so on. Sentences any embedded within the paragraph aren't vital. The most recent approach within the literature [12] exploits 3 position models. The primary assumes that sentences earlier to the beginning and finish of a document area unit additional doubtless to be additional content representative. The second orders solely the highest elements of the text. The last one uses sentences near topic headings to form the outline.

2.3.3. Sentence Similitude to the Title

Sentence likeness to the title is that the vocabulary overlap between this sentence and therefore the document title [15][16][17] [28]. during this case, sentences just like the title and sentences that embody the words within the title area unit thought of vital. A simple way to calculate this score is:

Score =Ntw/T
where,
Ntw = variety of title words in sentence, and
T = variety of words within the title.

3. Conclusion

This survey paper explains about various accounts of extractive summarization. An extractive summery is choice of main sentences from the corresponding documents. The importance of sentences is based on applied statistical and linguistic features of sentences. Many dissimilarities of the extractive approach are tried within the last 10 years. However, it's difficult to mention as to extent is the quantity of instructive sophistication, at sentence or text level, contributing to the performance. Without using NLP, the produced outline could bear from deficiency of cohesion and linguistics. If texts containing multiple topics, the created outline won't be balanced. Assigning correct weights of individual options is incredibly crucial and vital as excellence of ultimate outline is dependent there on. We should always devote longer decide feature weights.

References

[1] Ferreira, Rafael, et al. "Assessing sentence scoring techniques for extractive text summarization." *Expert systems with applications* 40.14 (2013): 5755-5764.
[2] Gupta, Vishal, and Gurpreet Singh Lehal. "A survey of text summarization extractive techniques."*Journal of Emerging Technologies in Web Intelligence* 2.3 (2010): 258-268.
[3] Amini, Massih-Reza, and Patrick Gallinari. "Self-supervised learning for automatic text summarization by text-span extraction." *Proceedings of the 23rd BCS European Annual Colloquium on Information Retrieval (ECIR'01)*. 2001.
[4] Yue Hu and Xiaojun Wan, *―PPSGen: Learning-Based Presentation Slides Generation for Academic Papers*", Knowledge and Data Engineering, IEEE Transactions on Volume: 27, Issue: 4, pp 1085 - 1097 April 2015.
[5] K.Gokul Prasad, Harish Mathivanan, Madan Jayaprakasam, T. V. Geetha, *"Document Summarization and Information Extraction for Generation of Presentation Slides*", Advances in Recent Technologies in Communication and Computing, 2009. ARTCom '09. International Conference, pp.126 – 128, Oct. 2009.
[6] NingZhong, Yuefeng Li, and Sheng-Tang Wu, *―Effective Pattern Discovery for Text Mining"* Knowledge and Data Engineering, IEEE Transactions Volume: 24, Issue: 1, pp. 30–44, Jan. 2012.
[7] Mohsen Pourvali and Mohammad Saniee Abadeh, *―Automated Text Summarization Base on Lexicales Chain and graph Using of WordNet and Wikipedia Knowledge Base*," IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.
[8] 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.

[9] Fang Chen, Kesong Han and Guilin Chen, *"An Approach to sentence selection based text summarization"*, Proceedings of IEEE TENCON02, 489-493, 2002.

[10] Rasim M. Alguliev and Ramiz M. Aliguliyev, "*Effective Summarization Method of Text Documents*", in Proceedings of IEEE/WIC/ACM international conference on Web Intelligence (WI'05), 1-8, 2005.

[11] Nenkova, Ani, & Mc Keown, Kathleen (2012*)."A survey of text summarization Techniques"*. In Mining text data (pp. 43–76). Springer.

[12] Barrera, Araly, & Verma, Rakesh (2012*). "Combining syntax and semantics for automatic extractive single-document summarization"*. In Proceedings of the 13th international conference on computational linguistics and intelligent text processing (pp. 366–377).Springer-Verlag.

[13] Madhavi K. Ganapathiraju, *"Overview of summarization methods*", 11-742: Self-paced lab in Information Retrieval, November 26, 2002.

[14] Klaus Zechner, *"A Literature Survey on Information Extraction and Text Summarization*", Computational Linguistics Program, Carnegie Mellon University, April14, 1997.

[15] Fattah, Mohamed Abdel, &Ren, Fuji (2009)."*Ga, mr, ffnn, pnn and gmm basedmodels for automatic text summarization*". Computer Speech and Language, 23(1),126–144.

[16] Kulkarni, U. V., & Prasad, Rajesh S. (2010*)."Implementation and evaluation of evolutionary connectionist approaches to automated text summarization*". In Journal of Computer Science (pp. 1366–1376).Science Publications.

[17] Satoshi, Chikashi Nobata., Satoshi, Sekine., Murata, Masaki., Uchimoto, Kiyotaka., Utiyama, Masao., & Isahara, Hitoshi. (2001). Keihanna human info communication."*Sentence extraction system assembling multiple evidence*."In Proceedings 2nd NTCIR workshop (pp. 319–324).

[18] Yongzheng, Nur and Evangelos, *"Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora*", WIDM'5, 51-57, BremenGermany,2005.

[19] Canasai Kruengkari and Chuleerat Jaruskulchai, "*Generic Text Summarization Using Local and Global Properties of Sentences*", Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI'03), 2003.

[20] Junlin Zhanq, Le Sun and Quan Zhou, *"A Cue-based Hub- Authority Approach for Multi-Document Text Summarization*", in Proceeding of NLP-KE'05, IEEE, 642-645, 2005.

[21] Joel larocca Neto, Alex A. Freitas and Celso A. A. Kaestner, "*Automatic Text Summarization using a Machine Learning Approach*", Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215, 2002.

[22] Rene Arnulfo Garcia-Herandez and Yulia Ledeneva, *"Word Sequence Models for Single Text Summarization*", IEEE,44-48, 2009.

[23] Gupta, P., Pendluri, V. S., & Vats. I. (2011).*"Summarizing text by ranking text units according to shallow linguistic features*". In 13th International conference on advanced communication technology (pp. 1620–1625).

[24] Van Britsom, Daan, Antoon Bronselaer, and Guy De Tre."*Using data merging techniques for generating multi-document summarizations.*" (2014).

[25] Edmundson, H. P. (1969). *"New methods in automatic extracting.*" Journal ACM, 16(2), 264–285.

[26] Lloret, Elena, & Palomar, Manuel (2009). *"A gradual combination of features for building automatic summarization systems*". In Proceedings of the 12th international conference on text. Speech and dialogue (pp. 16–23). Berlin, Heidelberg: Springer-Verlag

[27] Luhn, H. P. (1958)."*The automatic creation of literature abstracts*." IBM Journal of Research and Development, 2(2), 159–165.

[28] Abuobieda, A., Salim, N., Albaham, A. T., Osman, A. H., & Kumar, Y. J. (2012). *"Text summarization features selection method using pseudo genetic-based model*."In International conference on information retrieval knowledge management (pp.193–197).

[29] Gupta, P., Pendluri, V. S., & Vats. I. (2011).*"Summarizing text by ranking text units according to shallow linguistic features*". In 13th International conference on advanced communication technology (pp. 1620–1625).

Author Profile

Priya Ganguly received the B.E. degree in Information Technology from G.H. Raisoni College of Engineering in 2013 .Currently she is pursuing her M.E. Degree from MIT College of Engineering.

Dr Prachi M. Joshi is an Associate Professor in the Department of Computer Engineering, MITCoE Pune. She has received her B, Tech, M. Tech and Ph.D degrees from College of Engineering Pune. She has successfully supervised a plethora of projects at graduate and post-graduate level encompassing the domains of Artificial Intelligence, Data Mining and Machine Learning. Her research interest includes Information Retrieval and Incremental Machine Learning and has multiple research publications to her credit. She also has co-authored book on Artificial Intelligence by PHI