International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey of Generating Multi-Document Summarizations

Patil Ajita S.[1], P. M. Mane[2]

Pune University, Computer Department, DCOER, Pune, Maharashtra, India

Abstract: ***Summarization is a Process of filtering the most important information from source/sources for a particular user and task. Summarization is a very useful task which gives support to many other tasks. It takes advantage of the techniques developed for Natural Language Processing tasks. Multidocument summarization is a technique of summarize the multiple document into one paragraph. Multi-document summarization is different from single-document summarization. Single-document summarization can be considered as one of the sub-tasks of multi-document summarization. There exist several other important sub-tasks including identification of important common ideas in the documents, selecting representative summaries for each of these ideas and organizing the indicative summaries for the final summary. In this paper we describe a system for multi-document summarization. This survey plan to analyze some of the most relevant approaches in the areas of multiple-document summarization, giving special importance to empirical methods and extractive techniques***.

Keywords: Multi-document summarization, Text Summarization, Text mining, Information Extraction

1. Introduction

Text Summarization, as the method of determining the most salient information in a document or group of documents (for multi document summarization) and transferring it in less space, turned an effective area of study in both Information Retrieval (IR) and Natural Language Processing (NLP) communities. Summarization shares some fundamental methods with indexing as both are focused on recognition of the quality of a document. Also, very high quality summarization needs advanced NLP practices to be able to handle various Parts Of Speech (POS) taxonomy and inherent subjectivity. Generally, it's possible to differentiate different forms of summarizers.

Multi document summarization involves developing a short summary from a couple of documents which focuses on a single topic. Often yet another query can also be given to establish the data require of the summary. Typically, an effective summary should be appropriate, brief and fluent. It indicates that the summary should protect the most crucial methods in the original document set, include less repetitive data and should be well-organized.

Since the time that humankind was capable of writing down its thought and its findings in Ancient Egypt, he had the need of bundling it into libraries and get a review of this data. It's a well known fact that with the presentation of the World Wide Web this data generation has taken exponential extents. With an expected number of around 50 billion web sites indexed on Google today it is truly difficult to get an organized review of the information on offer or discover the craved information without the utilization of any kind of tool. Keeping in mind the end goal to help in this assignment there are search engines like Google and Yahoo! that always slither and index the web so one can seek through this massive supply. Both private and public institutes assemble data in databases, and obviously news offices and different news suppliers keep everybody, apparently whilst being target, posted on each and every occasion happening everywhere throughout the world. With this information presently accessible, despite the fact that it is indexed and bundled, one hardly discovers the time to peruse everything one would need to. Whilst more data gets to be quickly accessible each nanosecond, the time one has the capacity spend to absorb this data continues as before. That is the reason it steadily got to be more basic to automatically compress summarize textual data into briefer forms that still reflect the significant information. The way that there is a genuine requirement for summarizing information is not under debate any more [1]. Where one is intended to consider summarizing a document when one hears “summarize”, in the course of recent decades more research has been done for summarizing multiple documents related to the same topic, which is known as Multi-Document Summarization (MDS) problem [2]. At present, there are a number of agencies which are investing in building information frameworks that have the capacity of comprehending the Multi-Document Summarization problem among different problems. For example: The Defense Advanced Research Projects Agency (DARPA) all the more particularly their Translingual Information Detection Extraction and Summarization system (TIDES). A group inside TIDES is rising with help of external researchers which are mainly focusing on summarization and its evaluation. Because of this they held a workshop in 2000. In their introductory workshop with the plan to search for a long term evaluation of these sorts of frameworks developed an evaluation related to text summarization called the Document Understanding Conferences (DUC) series. Document Understanding Conferences (DUC) is run by National Institute of Standards and Technology (NIST). There are obviously numerous different institutes that do research on this topic and normally they assess their frameworks utilizing the difficulties and reference material gave by the DUC.

2. Literature Review

In the study by Kathleen McKeown, Rebecca J. Passonneau, David K. Elson [1], they had demonstrated that it is achievable to direct a task-based, or extrinsic, assessment of

summarization that yields huge conclusions. In their study, they also answered positively to the question Do Summaries Help? They showed that subjects deliver better quality reports utilizing a news interface with Newsblaster summaries than with no summaries. User satisfaction increases, because summary quality rose from none to human. Specifically, full multi- document summaries help clients perform better at fact-gathering than they do with no summaries. Users are more pleased by multi-document summaries than with negligible one-sentence summaries, for example, those utilized by business online news frameworks. Results of their studies assert the profit of research in multi-document summarization. Nonetheless, they had exhibited that numerous factors impact the degree to which summaries help. The answer to the question is clearly complex and a single study can just give halfway knowledge. They also identified possible effects on task completion.

C.-Y. Lin and E. Hovy [2] had proposed multi document summarization framework called as NeATS. They also evaluated it in DUC-2001. As a prototype framework, NeATS purposely utilized basic strategies guided by some principles: 1. Extracting vital concepts focused around dependable facts. 2. Filtering sentences by their positions and disgrace words. 3. Reducing repetition utilizing MMR. 4. Presenting summary sentences in their sequential request with time annotations. These principles worked successfully. On the other hand, the simplicity of the framework fits easily into advance improvements. They also expected use linguistic units smaller than sentences to enhance our maintenance score. The fact is NeATS executed and the human in pseudo accuracy. But it did not perform well in retention shows its summaries may contain good yet duplicated data. For improving ability of NeATS, they used to parse sentences containing key unigram, bigram, and trigram concepts. By using this they were able to identify relations of sentences in clusters. They investigated discourse processing techniques [3] or summary operators [4] for improving cohesion and coherence. They were also examining the DUC evaluation scores in the trust of recommending enhanced and steadier metrics.

J. Bleiholder and F. Naumann [5] had presented problem of data fusion in large data integration context. In this context data fusion is final step data integration procedure. In the field of information integration, merging these repeated records into a single representation and in the meantime determining existing information conflicts is still out of the center of standard research. In last few years, this problem had been attended by a number of researchers. They had compared common relational methodologies of data fusion. They indicated how they adapt to data conflicts and notice the qualities of the outcomes about that they deliver. They displayed and remarked on a list of information integration frameworks that are equipped for fusing data in different ways. They ordered the frameworks as per their capacities of dealing with conflicts. Conflict handling can be defined by the methodology the frameworks utilize to handle conflicts. Most of existing information integration systems does not permit to accomplish data fusion easily. The reason behind this is only small number of systems actually handles the problem.

In the study by K. S. Jones [6] had mentioned some important issues and topics related to automatic summarization. The status and state of automated summarizing had drastically changed in the last decade. There is an expansive research community, and there are operational frameworks working with open- domain sources in differed conditions. Compressing has profited from work with neighboring tasks, outstandingly retrieval and also question answering. Above all, it has profited from the evaluation projects of last decade. These have been noteworthy both for the system work they have fortified also the results got, and for the improvement of evaluation strategies also a growing consciousness of the requirement for legitimate specification and performance appraisal. In connection to summarization methods themselves, this wave of work has been valuable in investigating the possibilities and potential utilities of extractive summarizing, and particularly factual and shallow typical techniques that do not oblige substantial model instantiation, for instance in domain ontologies. There is some evidence such procedures can provide valuable goods where the summary prerequisites are unobtrusive, and hybrid systems somewhat more than simply factual ones. There is no reason, thusly, to assume that compressing exploration and improvement won't proceed.

On the other hand, against this, the work and assessments done so far have been constrained and miscellaneous when evaluated with the perspectives of the summarizing space explored at the Dagstuhl Seminar in 1993 [7]. The work on deep methodologies seems to be required if source-to-summary buildup obliges radical change of content and expression. This is not astounding: that studies [8] [9] experiments recommend, we do not know idea, aside from with huge application-particular direction, how to automate such methods.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad [10] had demonstrated implemented algorithm for multi-document summarization which overcomes the sentence extraction model. Accepting a set of comparable sentences as input obtained from multiple documents on the same event [11], their framework detects normal expressions over sentences and uses language generation to reformulate them as an intelligible summary. The utilization of generation to consolidate comparable information is another approach that significantly enhances the quality of the ensuing summaries, lessening repetition and expanding fluency.

E. Canhasi and I. Kononenko [12] had formalized the problem of the query-focused document summarization as the weighted archetypal analysis problem. Also they had exhibited how to incorporate query information in the own nature of archetypal analysis and how to use weighted version of archetypal analysis for simultaneous sentence clustering and ranking. They had inspected the proposed technique on a number of input matrix modeling designs, where the paper reports the best comes results on the multi-element graph model. They discovered that weighted archetypal analysis summary is a powerful summarization technique. Experimental results on the Duc2005 and Duc2006 datasets show the adequacy of their proposed

methodology, which contrasts well with the vast majority of the current matrix factorization strategies in the literature.

R. M. Aliguliyev [13] had proposed the methodology to automatic document summarization focused around clustering and extraction of sentences. Their methodology comprised of two steps. Initially sentences arcs clustered, and afterward on each one cluster representative sentences are characterized.

Daan Van Britsom, Antoon Bronselaer, Guy De Tr´e had proposed how to utilize data merging methods to summarize a set of coreferent archives that has been grouped whilst utilizing delicate computing strategies. The primary focus of this paper lies with the f_β-optimal merge function, a function recently presented here, that uses the weighted harmonic mean to discover a harmony in the middle of precision and recall. The worldwide precision and recall measures stated are characterized by means of a triangular norm receiving local precision and recall values as a data, to produce a multiset of key concepts that we can use to create summarizations. The f_β-optimal merge function is contrasted with a distance based merge function and a few pointwise merge functions from both a hypothetical and additionally an exploratory perspective. It will be demonstrated that the f_β-optimal merge function has advantages over the others, particularly if one takes the practical usage in the perspective of data merging and summarizing various documents concerning the same subject.

3. Conclusion

In this paper we have surveyed Multi-document Summarizations techniques. Summarization is the process that reduces the amount of text in a document while preserving its original meaning. In this paper, we have given a general overview of multidocument text summarization. It has specially benefits to tasks such as information retrieval, information extraction or text categorization. Research on this field will continue due to the fact that text summarization task has not been finished yet and there is still much to investigate and to improve. The main applications of this work are Web search Engines, text compression and word processor.

4. Acknowledgment

This article is supported by College authority and guide as well as University of Pune.

Reference

[1] K. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg, “Do summaries help?” in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR ’05. New York, NY, USA: ACM, 2005, pp. 210–217.

[2] C.-Y. Lin and E. Hovy, “From single to multi-document summarization: a prototype system and its evaluation,” in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ser. ACL ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 457–464.

[3] Marcu, D. 1999. Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds), *Advances in Automatic Text Summarization*, 123–136. MIT Press.

[4] Radev, D.R. and K.R. McKeown. 1998. Generating Natural Language Summaries from Multiple On-line Sources. *Computational Linguistics*, 24(3):469–500.

[5] J. Bleiholder and F. Naumann, “Data fusion,” ACM Comput. Surv., vol. 41, no. 1, pp. 1:1–1:41, Jan. 2009.

[6] K. S. Jones, “Automatic summarizing : The state of the art,” Inf. Process. Manage., vol. 43, no. 6, pp. 1449–1481, 2007.

[7] Endres-Niggemeyer, B., Hobbs, J. and Sparck Jones, K. (Eds.) (1995) Summarising text for intelligent communication, Dagstuhl-Seminar-Report; 79 (Full version), IBFI GmbH Schloss Dagstuhl, Germany, 1995.

[8] Marcu, D. (1999) ‘Discourse trees are good indicators of importance in text’, in Mani and Maybury (1999), 123-136. (1999a)

[9] Marcu, D. (2000) The theory and practice of discourse parsing and summarisation, Cambridge MA: MIT Press, 2000.

[10] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad, “Information fusion in the context of multi-document summarization,” in ACL ’99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 550–557.

[11] Kathleen R McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. submitted.

[12] E. Canhasi and I. Kononenko, “Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization,” Expert Systems with Applications, vol. 41, no. 2, pp. 535 – 543, 2014.

[13] R. M. Aliguliyev, “A new sentence similarity measure and sentence based extractive technique for automatic text summarization,” Expert Syst. Appl., vol. 36, no. 4, pp. 7764–7772, May 2009.

[14] Daan Van Britsom, Antoon Bronselaer, Guy De Tr´e, “Using data merging techniques for generating multi-document summarizations” IEEE TRANSACTIONS ON FUZZY SYSTEMS, 2014.

Author Profile

Ajita Patil received the B.E(I.T).from Rajarambapu Institute of technology, Sakharale, Islampur and appearing for M.E in computer Engineering from Zeal Education Scociety, Dyanganga College of Engineering and research, Narhe, Pune, India.