A review of recent progress in multi document summarization

Shazia Tabassum[1], Eugenio Oliveira[2]

[1]Ph.D. Student, Faculdade de Engenharia, Universidade does Porto
Porto, Portugal
[2]Professor, Faculdade de Engenharia, Universidade do porto
Porto, Portugal

Abstract. The increase of information available in the form of text, led to the need of extensive research in the area of text summarization. Early the researches in this area started with single document summarization and drove towards multi document summarization**.** We present here a comparative review of the recent progress in the field of multi document summarization. The strengths and weaknesses of the techniques used in the recent researches are highlighted. The state of the art including methods and algorithms on multi document summarization is outlined and discussed. Finally some open research issues are identified.

Keywords: Multi Document summarization, Extraction, Abstraction, Approaches**.**

1 Introduction

One of the major problems being addressed in computer science and informatics from past few years is Big data. A considerable part of the Big data is text oriented. For example Social media, blogs, emails, comments, reviews, news wires etc. The major concerns associated with this information overload are the unlimited text for humans to read or analyze and the limited storage capacity for machines. The present era with huge amounts of unstructured data raises the need for extensive research in the area of text summarization. With the advent of Web 2.0, the Big data would get bigger which implies a strong need for text summarization.

Text Summarization can be seen as an automatic system that makes a précis of text from a single or multiple documents while maintaining the information, meaning, significance and the order of events in the original text. [39] States that, "Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user or task."

Multi document summarization (MDS) is the task of producing a concise and fluent summary to deliver the major information for a given document set. Multi-document summaries can be used for users to quickly browse document collections, and it has been shown that multi-document summaries can be helpful in information retrieval systems [1].

The summarization task is mainly divided into two categories, extractive summarization and abstractive summarization. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary assuming that these sentences convey the meaning of the whole text. Extraction based summaries produce much less accuracy compared to human made summaries. These methods are easier to apply compared to abstraction based summaries. Abstraction based methods create a compressed version of text conveying the summarized meaning of the original text. Abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original text [2].

The main goal of this paper is then to overview text summarization different approaches and extract some useful conclusions about them. The rest of the paper is organized as follows: Section 2 outlines the state of the art in text summarization. Section 3 classifies text summaries into different types. Section 4 identifies different approaches followed in previous researches. Section 5 compares the most recent researches in the field. Section 6 presents concluding remarks and scope for future work.

2 State of the Art

Early approaches to text summarization began with the work of [3] in 1950`s. Many approaches have been addressed and many methods have been evaluated since then. Recent approaches used statistical methods such as word frequency, TF-IDF weighting like Sum-Basic [4] [5], sentence position, title relation, and cue-phrases etc. Other approaches take account of semantic associations between words and combine them with those shallow features in the process of sentence similarity. Examples of such approaches are, among others, latent semantic analysis [6], topic signatures [7] and sentence clustering [8].

In recent years, multi-document summarization research has shown increased interest in graph-based approaches [9] [10] [11] [12] [13] [14] and Bayesian topic model based approaches [15] using two-tiered topic model (TTM) in [16] with topic segmentation [17] and topic sum [18]. Others identify the relevance of a sentence by using bigram pseudo sentences for implementing hybrid statistical sentence-extraction [19], rhetoric-based MDS [20] and semantic document concept technique [21] for analyzing grammatical structures in discourses. Recently developed Bayesian collection models incorporated the concept of latent topics into n-gram language models, such as the LDA-HMM model integrating topics and syntax [22], structured topic model [23], and topical n-grams [24]. [25] Uses the centroids to identify sentences in each cluster that are central to the topic of the entire cluster.

On the contrary only few works concentrated on abstractive summarization. Like [37] the author deals with identifying and synthesizing similar elements across related text from a set of multiple documents using natural language text to text generation techniques like content selection, paraphrasing rules, temporal ordering. A fully Abstractive Approach to Guided Summarization is presented by [38] using Information extraction, content selection and generation. Sentence compression [39]. Sentence fusion [40] or sentence revision [41].

3 Types of Summaries

Based on the research work in the field of text summarization, we discuss here the following types of summaries that have been generated.

Generic summaries
Generic summarization tries to extract the most general idea from the original document set without any specified preference in terms of content. For generic summarization, a saliency score is usually assigned to each sentence, the sentences are ranked according to the saliency score, and then the top ranked sentences are selected as the summary based on the ranking result. Recently, both unsupervised and supervised methods have been proposed to analyze the information contained in a document set, and extract highly salient sentences into the summary based on syntactic or statistical features.

Query-Focused Summaries
Query-focused summarization aims at generating a short summary based on a given document set and a given query. The generated summary reflects the condensed information related to the given query within the specified summary length. In query-focused summarization, the information related to a given topic or query should be incorporated into summaries, and the sentences suiting the user's declared information need should be extracted. Many methods for generic summarization can be extended to incorporate the query information.

Update Summaries or incremental summaries
Update summarization is automatically updating summaries as new documents are added to the existing batch of documents. Generating updated summaries as the new documents arrive. Most of existing summarization methods work on a batch of documents and do not consider that documents may arrive in a sequence and the corresponding summaries need to be updated in real time.

Topic Focused Summaries
Topic focused summaries are generated using topic or event based models. A topic model is a type of top-down approach. It considers the same problem based on semantic associations behind the content. In the Bayesian topic model based approaches,

similarity is analyzed using advanced methods with respect to probabilistic distributions of topics [26].

4 Approaches

In order to generate the above types of summaries the approaches followed in the recent works are listed below.

Feature based approach
One of the most common methods used in text summarization field is the feature based method. In the process of identifying important sentences, features influencing the relevance of sentences are determined. Some features that are often considered for sentence selection are word frequency, title words, cue words, sentence location and sentence length [27].

Domain-Specific/Ontology based approach
Generally speaking, ontology is often provided by domain experts [28]. Such ontology provides answers for the questions concerning what entities exist in the domain and how such entities can be related within a hierarchy and subdivided according to similarities and differences among them.

Cognitive based approach
Cognitive psychology is the study of mental processes such as "attention, language use, memory, perception, problem solving, creativity and thinking." [31] Cognitive based approach uses human cognitive factors in reading process. The previous researches in this area have mainly used three cognitive processes, i.e. forget process, recall process and association process for generation of summaries.

Event based approach
Event-based MDS was first proposed by [30], the authors selected sentences based on relevance for one or more sub-events of the topic at hand. Human judges manually determined the sub-events of a topic and assigned to each sentence a relevance score for each sub-event. They show that the algorithm that selects sentences with the highest sum of scores over all sub-events produces the most informative summaries.

Discourse based approach
Discourse is an organic structure. Different parts of discourse bear different functions, and have complex relationships among them. Automatic summarization based on discourse attempts to analyze the structural features of discourse to identify the main content of the article. Currently, automatic summarization based on discourse has five main research topics: rhetorical structure analysis, pragmatic analysis, lexical chain, relationship map and latent semantic analysis [31].

Table. I The comparison of most recent researches on MDS.

NAME	*CATEGORY*	*TYPE*	*APPROACH*	*METHOD*	*HIGHLIGHTS*	*LIMITATIONS/IMPROVEMENTS*	*EVALUATION ROUGE 2*
1.A MDS system based on statistics and linguistic treatment (2014) [11]	*Extractive*	*Generic*	*Statistics, Linguistics and machine learning*	*Graph based clustering algorithm*	*To deal with multi document issues such as redundancy and problem diversity*	*The limitation of the approach is the problem of sentence ordering, as the system tries to find relevant sentences in groups of different topics.*	*200 word sumary: Recall- 62% .400 word summary: Recall 53%*
2.A novel contextual topic model for multi-document summarization (2015) [26]	*Extractive*	*Topic focused*	*Contextual topic model based approach*	*Bayesian topic model*	*A model that can capture both the hierarchies and the word dependencies over latent topics*	*The model has to take longer time to be trained under a larger data set in order to keep certain level of accuracy in prediction. It is poor to cover co-occurrence among multiple words. Other limitations include the problem of sentence coherence and a lack of online settings for stream text because the model has been trained using a large data corpus before summarizing documents..*	*DUC(2006) Recall 0.0986*

5. Event graphs for information retrieval and multi-document summarization (2014) [12]	*4.FoDoSu:Multi-document summarization exploiting semantic analysis based on social Folksonomy (2015) [32]*	*3.Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization(2014) [9]*
Extractive	*Extractive*	*Extractive*
Event focused	*Generic*	*Query based*
Graphs, machine learning and rule based extraction methods	*Using statistical and linguistic methods*	*Graph model using matrix factorization*
Event graphs, Logistic regression classifier, argument extraction, Temporal relation extraction	*HITS algorithm and semantic analysis.*	*Weighted archetypal analysis*
The contribution of this article is a novel event-centered document representation that accounts for the semantics of events. The other contribution of this article is a novel event-centered mds.	*To analyze the relationship between the semantics of the words in Web documents.*	*1.To incorporate query information in its own nature of an archetypal analysis.* *2. To increase variability and diversity of the produced query based summary.*
To improve the extraction of temporal relations between events which would yield further performance improvements in event-centered retrieval and summarization. The second improvement would be to enrich event graphs with other relations that can hold between events,such as causality, entailment,and atiotemporal contai nment..	*To improve methods for analyzing the semantics of words that is difficult to analyze, such as proper nouns and newly-coined words.*	*In future work WordNet could be used to calculate the semantic similarity between sentences by using the synonyms sets of their component terms;Another possible enhancement can be reached by introducing the multi-layered graph model that emphasizes not only the sentence to sentence and sentence to terms relations but also the influence of the under sentence and above term level relations, such as n-grams, phrases and semantic role arguments levels.*
DUC 2002,2004 Recall 0.116 ,0.107	*Tac 2008, tac 2009 Recall 0.06853*	*DUC(2006) Recall - 0.0917*

8.Cognitive based MDS (2014) [34]	*7.An empirical study on ontology based multi document summarization in disaster management (2014) [28]*	*6. Multi document summarization based on news components using fuzzy cross – document relations (2014) [33]*
Extractive	*Extractive*	*Extractive*
Topic focused	*Generic and Query focused*	*Generic*
Cognitive based	*Ontology based using statistics and machine learning*	*Feature based approach using evolutionary algorithm, Fuzzy logic, machine learning*
IRatio, GWI(global word impression)LWI(local word impression)	*TF, TF-IDF, TF-ICF, Concept hierarchy and clustering*	*Genetic algorithm, Case based reasoning, classification and fuzzy scoring*
Proposed inter-document recall process and forget process in the scanning mechanism.	*Introduction of domain specific ontology in disaster management for generic and query focused summaries.*	*Introduction of a multi document summarization model by taking into account the generic components of news story. The study further investigates the utility of cross-document relations (CST relations) to identify highly relevant sentences to be included in the summary.*
To employ sophisticated methods on understanding the semantics and redundancy of sentences to improve the quality of summary.	*To utilize the hierarchical correlations in the ontology to further improve the quality of the summary. To employ information extraction techniques to further improve summarization results.*	*To explore how natural language processing techniques can be employed to connect semantic concepts with news components. To study the utility of cross-document relations identified from un-annotated text documents to generate better summaries by treating issues related to multi document such as contradictions and historical information.*
UCI 2007 *Recall 0.12288*	*Hurricane dataset* *Recall 0.30160*	*DUC 2002* *Recall - 0.1280*

11.Exploring actor–object relationships for query-focused MDS(2014) [36]	10. SRRank: Leveraging Semantic Roles for Extractive MDS(2014) [1]	9.Incremental MDS: An Incremental Clustering Based Approach (2014) [35]
Extractive	Extractive	-
User Focused	Topic Focused	Incremental/ Update
User based, Feature based , Machine learning, Actor Object relationship (AOR)	Graph based ranking algorithm	Machine learning, Clustering
Back propagation on neural network,	Semantic parsing SRRank	Clustering
Combines ensemble summarizing system and AOR to generate summaries	Proposes a novel graph-based sentence ranking algorithm SRRank to incorporate the semantic role information into the graph-based ranking algorithm.	Study of order dependency on clustered documents
Plan to examine further the effectiveness of exploiting other types of patterns resulting from using dependency parsers in a user-based summarization system and investigate their effectiveness.	making use of deep semantic information instead of shallow semantic information for further improving multi-document summarization.	For maintaining the summaries stable, there must be some mechanism that keeps the extracted important sentences in the order irrespective of the order in which the input documents are handled by the program..
DUC 2006 & 2007 Recall 0.933, 0.1219	UCI 2006 & 2007 Recall 0.9044 & 0.956	-

5 Comparison of most recent Researches

In this section we present a comparative study of the researches on multi document text summarization techniques from previous year. Recent researches have concentrated on different approaches discussed in the previous section. Table I highlights the comparative points between those techniques. We have pointed out different types of summaries generated by using different approaches and methods. We have also brought out the limitations or improvements suggested for those approaches. Last column reports on the evaluation results using ROUGE set of metrics.

6 Conclusion and future work

In this paper we have clearly classified text summarization into different types, following different approaches and using different methods. We have also compared the most recent researches in the area of MDS. Referring to the work above, we would say that there are very few works using abstraction while most of the summaries use extractive approaches. The research work in text summarization is expanding with the implementation of methods and methodologies from various fields like cognitive psychology, evolutionary algorithms, discourses etc. By analyzing the results of the recent research, we found that the works using these fields have outperformed the previous methods, but are still far from human generated summaries.

Some limitations in previous research works help us to identify few research issues in MDS like, problem of sentence ordering, redundancy of sentences, enriching graphs with semantic relationship between sentences and documents, improving feature extraction methods, understanding the human way of summarizing, improving coherence in summaries.

These challenges can be seen as relevant motivation and possible guidelines for future research topics in the area of MDS.

References

1. Yan S, Wan X (2014) SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization. IEEE/ACM Transactions on audio, speech, and language processing, Vol.22, No.12.
2. Rafeeq Al-Hashemi (June 2010) Text Summarization Extraction System (TSES) Using Extracted Keywords. International Arab Journal of e-Technology, Vol. 1, No. 4
3. Luhn HP (1958) The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.
4. Nenkova A & McKeown K (2012) A survey of text summarization techniques. In Mining text data , US: Springer, pp 43–76.

5. Vanderwende L, Suzuki H, Brockett, C & Nenkova A (2007) Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing & Management, 43(6), 1606–1618.
6. Gong Y & Liu X (2001) Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 19–25.
7. Lin CY & Hovy E (2000) The automated acquisition of topic signatures for text summarization. In Proceedings of the international conference on computational linguistic. pp 495–501.
8. He R, Qin B & Liu T (2012) A novel approach to update summarization using evolutionary manifold-ranking and spectral clustering. Expert Systems with Applications, 39(3), 2375–2384.
9. Canhasi E & Kononenko I (2014) Weighted archetypal analysis of the multi element graph for query-focused multi-document summarization. Expert Systems with Applications, 41(2), 535–543.
10. Ferreira R, de Souza Cabral L, Lins RD, Pereira e Silva G, Freitas F, Cavalcanti GD et al (2013) Assessing sentence scoring techniques for extractive text summarization. Expert Systems with Applications, 40(14), 5755–57
11. Ferreira R, de Souza Cabral L, Freitas F, Lins R D, de França Silva G et al (2014) A multi-document summarization system based on statistics and linguistic treatment. Expert Systems with Applications, 41(13), 5780–5787.
12. Glavaš G & Šnajder J (2014) Event graphs for information retrieval and multi-document summarization. Expert Systems with Applications, 41(15), 6904–6916.
13. Mendoza M, Bonilla S, Noguera C, Cobos C & León E (2014) Extractive single document summarization based on genetic operators and guided local search. Expert Systems with Applications, 41(9), 4158–4169.
14. Zhao L, Wu L & Huang X (2009) Using query expansion in graph-based approach for query-focused multi-document summarization. Information Processing & Management, 45(1), 35–41.
15. Daumé III H & Marcu, D (2006) Bayesian query-focused summarization. In Proceedings of the conference of the association for computational linguistics (ACL) and 44th annual meeting of the ACL, Sydney, pp 305–312.
16. Celikyilmaz A & Hakkani-Tur D (2011) Discovery of topically coherent sentences for extractive summarization. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, Vol. 1, pp. 491–499.
17. Eisenstein J & Barzilay R (2008) Bayesian unsupervised topic segmentation. In Proceedings of the conference on empirical methods in natural language processing, Oct 25–27, Honolulu, Hawaii.
18. Haghighi A & Vanderwende L (2009) Exploring content models for multi document summarization. In Proceedings of human language technologies: The annual conference of the North American chapter of the association for computational linguistics, Boulder, Colorado, pp. 362–370.
19. Ko Y & Seo J (2008) An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. Pattern Recognition Letters, 29(9), 1366–1371.
20. Atkinson J & Munoz R (2013) Rhetorics-based multi-document summarization. Expert Systems with Applications, 40(11), 4346–4352.
21. Ye S, Chua T S, Kan M Y & Qiu L (2007) Document concept lattice for text understanding and summarization. Information Processing & Management, 43(6), 1643–1662.

22. Griffiths T, Steyvers M, Blei D & Tenenbaum J (2005) Integrating topics and syntax. Advances in Neural Information Processing Systems, 17.
23. Wallach H M (2006) Topic modeling: Beyond bag-of-words. In Proceedings of the 23rd international conference on machine learning , ACM, pp. 977–984.
24. Wang X, McCallum & Wei X (2007) Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Proceedings of the 7th IEEE international conference on data mining , pp 697–70.
25. Radev DR, Jing H, Stys M and Tam D (2004) Centroid-based summarization of multiple documents. Information Processing and Management, 40, 16-17.
26. Yang G, Wenb D, Kinshuk, Chen N, Sutinen E (2014) A novel contextual topic model for multi-document summarization. Expert Systems with Applications, 42, 1340–1352.
27. Gupta V, Lehal GS (2010) A survey of text summarization extractive techniques, J. Emerg. Technol. Web Intell. 2 258–268.
28. Wu K, Li L, Li J, Li T (2013) Ontology-enriched multi-document summarization in disaster management using sub modular function. Information Sciences 224, 118–129.
29. American Psychological Association (2013). Glossary of psychological terms. Apa.org. Retrieved 2014-08-13.
30. Daniel N, Radev D & Allison T (2003) Sub-event based multi-document summarization. In Proceedings of the HLT-NAACL 03 Workshop on Text Summarization, Association for Computational Linguistics, Vol. 5, pp 9–16.
31. Wang S, Li W, Wang F, Deng H (2010) A Survey on Automatic Summarization. International Forum on Information Technology and Applications.
32. Heu J, Qasim I, Lee D (2015) FoDoSu: Multi-document summarization exploiting semantic analysis based on social Folksonomy. Information Processing and Management 51, 212–225.
33. JayaKumar Y, Salim N, Abuobied A, Albaham A T (2014) Multi document summarization based on news components using fuzzy cross-document relations. Applied Soft Computing 21, 265–279.
34. Chen J, Li W (2013) Cognitive-based Multi-Document Summarization Approach. Ninth International Conference on Semantics, Knowledge and Grids.
35. Johney John, Asharaf S (2014) Incremental Multi-Document Summarization: An Incremental Clustering Based Approach. International Conference on Data Science & Engineering (ICDSE).
36. Valizadeh M & Brazdil P (2014) Exploring actor–object relationships for query-focused multi-document summarization. Soft Computing. doi:10.1007/s00500-014-1471-x.
37. Regina Barzilay and Kathleen R. McKeown (2005) Sentence fusion for multidocument news summarization. Computational Linguistics, 31(3):297-328.
38. Genest PE and Lapalme G (2012) Fully abstractive approach to guided summarization. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2. Jeju Island, Korea, Association for Computational Linguistics: 354-358.
39. Cohn T and Lapata M (2009) Sentence compression as tree transduction. J. Artif. Int. Res. 34(1): 637-674.
40. Barzilay R et al (1999) Information fusion in the context of multi-document summarization. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, Maryland, Association for Computational Linguistics, pp 550-557.

41. Tanaka H et al (2009) Syntax-driven sentence revision for broadcast news summarization. Proceedings of the 2009 Workshop on Language Generation and Summarization. Suntec, Singapore, Association for Computational Linguistics, pp 39-47.