Literature Review on Automatic Text Summarization: Single and Multiple Summarizations

Neelima Bhatia
Amity School of Engineering and Technology (ASET)
Amity University Noida, India

Arunima Jaiswal
Amity School of Engineering and Technology (ASET)
Amity University Noida, India

ABSTRACT
The online information available on world wide web is in enormous amount. Search engines like Google, Yahoo were developed to retrieve information from the databases. But actual results were not obtained as the electronic information is increasing day by day. Thus automatic summarization came into demand. Automatic summarization gathers several documents as input and provides the shorter summarized version as output which is informative, unambiguous, save valuable time. Research was done on a single document and moved towards multiple documents. This review categorizes single and multiple summarization methods.

KEYWORDS
Automatic Text summarization

1. INTRODUCTION
Internet is a wide source of electronic information. But the outcome of information retrieval becomes a tedious task for humans. Thus automatic summarization came into demand which automatically retrieves the data from documents by utilizing our precious time. H.P. Luhn was the first one who invented automatic summarization of text in 1958 [24].

NLP community invented the subfield of summarization. Radev et al [28] says that one or more documents are processed and a short summary is produced which is less than the size of original documents. He portrayed his definition as:

- One or more documents can produce summary.
- Important information is not lost.
- Short length is maintained.

There are approaches which are helpful to generate a summary – extraction and abstraction. Extraction is domain independent and picks up the important sentences and gives a summary while on the other hand, abstraction is domain dependent and takes the human knowledge by understanding the whole text and prepares a goal and produce a summary [25].

Summarization is of two types [1]:

- Single document text summarization
- Multi-document text summarization

The idea of single document summarization dropped after 2002 [26]. The major focus was on multi-document summarization because it believes in size reduction, gathering ideas from several documents and compare them, maintaining the syntax and semantic relationship [27].

The paper organization is as follows. Section 2 describes the related work done by the pioneers in single and multi-document summarization. Section 3 provides the classification among the methods used by single and multi-document summarization. Section 4 concludes this paper.

2. RELATED WORK
Single Document Summarization: Various technical documents were focused in single-document summarization. Luhn in 1958 shows the significance of words based on frequency measures. He deleted the stop words and rest words are given a hierarchy starting from root and index describes the significance of each word. This is calculated on the number of occurrences in a document called as significant factor and are ranked. Based on ranking top sentences are selected to form a summary [17].

Baxendale in 1958 focused on sentence position to find the salient features. He took 200 paragraphs and examined that in 85% of paragraphs topic sentences are placed in the beginning while in rest 7% he found, it occurred in the last [18].

Edmundson in 1969 proposed a typical structure that produces extracts. In the beginning he took around 400 technical documents and build a protocol producing manual extracts. He addressed the above two features (word frequency, word position) and gave the two new features named cue words and skeleton (title or heading). Also the weights were attached with these. He evaluated and found that 44% machine extracts matched with manual extracts [19].

Various other pioneers were there who applied different techniques in single document summarization:

- In 1961 G.J. Rath [29] used lexical indicators to determine the relevant information from documents.
- In 1995 Julian Kupiec [30] used algebraic method to determine different features like uppercase words, length, position of words by using naïve-bayes classifier.
- In 1997 ChinYew Lin [31] determine the position of sentences by using algebraic methods.
- In 1999 Eduard Hovy [32] used symbolic word knowledge with strong NLP processing to show the concepts relevancy.
- In 2005 S.P Yong [33] used neural netwok. He showed Summarization = Text pre-processing sub-system + Keywords Extraction sub-system + Summary production sub-system.
- In 1976 M.A. K. Halliday [34] used lexical semantic relationships to build lexical cohesion blocks and their patterns.
- In 1984 Ruqaiya Hasan [35] used lexical cohesion to identify similarity chains.

- In 1988 William C.Mann [36] used RST (rhetorical structure theory) to encode the terminal nodes of a tree.
- In 1991 Jane Morris [37] used cohesion chains to determine the sequence of associated words.
- In 1997 Branimir Boguraev [38] used saliency-based content characterization to rank the important sentences in unstructured document.
- In 2010 Li Chengcheng [39] used RST to analyze candidate sentence, discover rhetoric relations and give the construction.
- In 2000 Hongyan Jing [40] used human abstraction concept by taking the closely realated sentences and eliminating the extra ones.

Multi-Document Summarization: The major contribution was done by McKeown and Raedev in 1995 (NLP group) at Columbia University and SUMMONS was build [20]. Similarity measures were used and extractive techniques were applied. McKeown et al. in 1999 and Radev et al. [20] in 2000 identified common themes using clustering while Barzilay et al. [21] in 1999 produced composite sentences from cluster whereas Carbonell and Goldstein [22] in 1998 used maximal marginal relevance (MMR). A major contribution where multi-document summarization was concatenated to multilingual environment by Evans in 2005 [23].

Various other pioneers has worked in this field using different techniques.

- G.Salton in 1989 [41] used TFI X IDFI techniques to evaluate the frequency.
- Jun'ichi Fukumoto in 2004 [42] generate abstract by using TF/IDF for single and multiple documents.
- You Ouyang in 2009 [43] used word hierarchial technique for most frequent terms at the top.
- Vikrant Gupta in 2012 [44] used kernel which serve as a guideline to choose other sentences for summary by using statistical measures.
- Inderjeet Mani in 1997 [45] used graph based method to discover the nodes by applying a spreading activation technique.
- Rada Mihalcea in 2004 [46] used graph based method by adding a vertex for every sentence by creating links for similar sentences.
- Xiaojun Wan in 2008 [47] used graph based method by introducing two-link graph for both sentences and documents.
- Kathleen McKeown in 1995 [48] used time based technique which focuses on how the trends of events change with respect to time.
- Shanmugasundaram Hariharan in 2012 [49] used sentence co-relation method where sentences are extracted on the basis of vote casting, scores and positions to get extracts.
- Tiedan Zhu in 2012 [50] emphasized on logical-closeness rather than topical-closeness using sentence co-relation method.
- Jade Goldstein in 2000 [51] used clustering, coverage, anti redundancy and summary cohesion for minimizing redundancy and maximizing both relevance and diversity,
- Judith D.Schlesinger in 2008 [52] combines clustering, linguistics, statistics for summarization by using clustering based method.
- Nitin Agarwal in 2011 [53] used query-oriented approach with unsupervised approach with the help of clustering based method.

3. CLASSIFICATION OF AUTOMATIC TEXT SUMMARIZATION

Automatic Text Summarization can be characterized into single document text summarization and multi document summarization.

Single-Document Summarization: The biggest challenge in summarization is to identify or generalize the most important and informative sentences from a document because the information in the document is non-uniform usually [1].

There are certain ways for single document summarization:

Naïve-Bayes [2]: Here a classification function namely naïve-bayes is used to distinguish whether sentences are likely to be extracted or not.

Rich Features and Decision Trees [3]: Generally the text is portrayed in a predictable discourse structure and the important sentences occur at specific locations. This method is known as "position method" which shows the position of sentences.

Hidden Markov Model [4]: Conroy et al used hidden markov model (HMM) and identified the problem of sentence extraction from a document.

Log Linear Model [5]: Osborne used log-linear models and showed that existing approaches used feature independence and these models produce better extracts than naïve-bayes model.

Neural Networks [6]: Due to its outperforming statistical significance, neural network overcome the problem of extractive summarization.

Deep Natural Language Analysis Method [7]: Here a set of heuristics are used to make document extracts. Also they model the discourse structure of texts.

Multi-Document Text Summarization: Since 1990's, single document extraction has moved to multiple document extraction in the domain of news articles. Various news articles like Google News [8], Columbia News Blaster [9] and News In Essence [10] were inspired from multi-document summarization. Though single document puts contradictory results by overlapping the information because of multiple documents availability [1]. So the major focus on summary is that summary should follow the completeness, correctness, erroneous property.

There are certain ways for multi-document summarization:

Abstraction and Information Fusion [11,12]: Here a summary is built by fusing multiple documents by giving input to process the text and then extracting the important information to produce a well structured summary.

Topic-driven Summarization and MMR [13]: Here the main focus is on the query and the information retrieved from text retrieval to topic-driven summarization. In maximal marginal relevance (MMR), the redundant sentences are less rewarded by some similarity measures.

Graph Spreading Activation [14]: In this a document is treated as a graph and each node represents the word with its position. Also a node can have various links like adjacency links (ADJ) which shows the adjacent words, Same links which shows the number of occurrences of a word, Alpha links encodes the meanings. Also Phrase links binds the sequence of adjacent nodes in a phrase whereas Name and Coref links checks the occurrence of co-referential name.

Centroid-based Summarization [15]: Here articles are grouped together which describes the same event. Every cluster constitutes of 2-10 articles from different sources and are arranged in chronological order. This step is called as topic detection. An agglomerative clustering algorithm adds documents to clusters by using TF-IDF vector and recomputes the centroids. Thus centroids are known as pseudo-documents because a cluster formation occurs with the help of TFIDF scores. After this sentences are identified from each cluster which describes the topic by using centroids.

Multilingual Multi-document Summarization [16]: Here multiple documents are there in multiple languages. First, a translation system is applied for translation of document in a single preferable language. Then similar sentences are searched in the documents. If found relevant then they are included in summary directly rather than translating. This is useful for news applications that take information from other agencies of different language.

4. CONCLUSION

This literature review mainly focused on the pioneered work by great personalities who contributed in the field of automatic summarization. Also a brief classification is explained by various methods. In this era of abundant online information for a single topic, multi-document summarization is necessary as due to the abundant electronic information which is a known problem in terms of big data.

5. REFERENCES

[1] Suneetha S., "Automatic Text Summarization: The Current State of the art", "International Journal of Science and Advanced Technology (ISSN 2221-8386) Volume 1 No 9 November 2011".

[2] Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In Proceedings SIGIR '95, pages 68{73, New York, NY, USA.

[3] Lin, C.-Y. and Hovy, E. (1997). Identifying topics by position. In Proceedings of the Fifth conference on Applied natural language processing, pages 283{290, San Francisco, CA, USA.

[4] Conroy, J. M. and O'leary, D. P. (2001). Text summarization via hidden markov models. In Proceedings of SIGIR '01, pages 406{407, New York, NY, USA.

[5] Osborne, M. (2002). Using maximum entropy for sentence extraction. In Proceedings of the ACL'02 Workshop on Automatic Summarization, pages 1{8, Morristown, NJ, USA.

[6] Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In Proceedings of AAAI 2005, Pittsburgh, USA.

[7] Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization. In Proceedings ISTS'97.

[8] http://news.google.com.

[9] http://newsblaster.cs.columbia.edu.

[10] http://NewsInEssence.com.

[11] McKeown, K. R. and Radev, D. R. (1995). Generating summaries of multiple news articles. In Proceedings of SIGIR '95, pages 74{82, Seattle, Washington.

[12] Radev, D. R. and McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. Computational Linguistics, 24(3):469{500.

[13] Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR '98, pages 335{336, New York, NY, USA.

[14] Mani, I. and Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In AAAI/IAAI, pages 622-628.

[15] Radev, D. R., Jing, H., Stys, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing and Management 40 (2004), 40:919-938.

[16] Evans, D. K. (2005). Similarity-based multilingual multi-document summarization. Technical Report CUCS-014-05, Columbia University.

[17] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159{165.

[18] Baxendale, P. (1958). Machine-made index for technical literature - an experiment. IBM Journal of Research Development, 2(4):354{361.

[19] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM, 16(2):264{285.

[20] McKeown, K. R. and Radev, D. R. (1995). Generating summaries of multiple news articles. In Proceedings of SIGIR '95, pages 74{82, Seattle, Washington.

[21] Barzilay, R., McKeown, K., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In Proceedings of ACL '99.

[22] Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR '98, pages 335{336, New York, NY, USA.

[23] Evans, D. K. (2005). Similarity-based multilingual multi-document summarization. Technical Report CUCS-014-05, Columbia University.

[24] Hans P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, 1958.

[25] B. Cretu, Z. Chen, T. Uchimoto and K. Miya, "Automatic Summarization Based on Sentence Extraction: A Statistical Approach," International Journal

of Applied Electromagnetics and Mechanics, IOS Press, vol. 13, no. 1-4, pp. 19-23, 2002.

[26] Krysta M. Svore, Lucy Vanderwende and Christopher J.C. Burges, "Enhancing Single document Summarization by Combining RankNet and Third-party Sources," Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 448–457, 2007.

[27] Pervin S., Haque M., "Literature Review of Automatic Multiple Documents Text Summarization", International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 3 No. 1 May 2013, pp. 121-129 © 2013 Innovative Space of Scientific Research Journals.

[28] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. Computational Linguistics., 28(4):399-408.

[29] G. J. Rath, A. Resnick and T. R. Savage, "Comparisons of four types of lexical indicators of content," Journal of the American Society for Information Science and Technology, vol. 12, no. 2, pp. 126-130, April 1961.

[30] Julian Kupiec, Jan Pedersen and Francine Chen, "A Trainable Document Summarizer," Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 68-73, 1995.

[31] Chin-Yew Lin and Eduard Hovy, "Identifying Topics by Position," In Proceedings of the Fifth conference on Applied natural language processing, San Francisco, pp. 283-290, 1997.

[32] Eduard Hovy and Chin-Yew Lin, Automated Text Summarization in SUMMARIST, In: Inderjeet Mani and Mark T. Maybury (Eds.), Advances in Automatic Text Summarization, MIT Press, chapter 8, pp. 18-24, 1999.

[33] S. P. Yong, A. I. Z. Abidin and Y. Y. Chen, "A Neural Based Text Summarization System," 6th International Conference of Data Mining, pp. 45-50, 2005.

[34] M. A. K. Halliday and Ruqaiya Hasan, Cohesion in English, Longman, London, 1976.

[35] Ruqaiya Hasan, Coherence and Cohesive Harmony, In: Flood James (Ed.), Understanding Reading Comprehension: Cognition, Language and the Structure of Prose. Newark, Delaware: International Reading Association, pp. 181-219, 1984.

[36] William C. Mann and Sandra A. Thompson, Relational Propositions in Discourse, Defense Technical Information Center, Information Sciences Institute, Marina del Rey, 1983. [Online] Available: http://www.tandfonline.com/doi/abs/10.1080/01638538609544632?journalCode=hdsp20#preview

[37] Jane Morris and Graeme Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text," Journal of Computational Linguistics, vol. 17, no. 1, pp. 21-48, March 1991.

[38] Branimir Boguraev and Christopher Kennedy, "Salience-based Content Characterization of Text Documents," In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, 1997.

[39] Li Chengcheng, "Automatic Text Summarization Based On Rhetorical Structure Theory," International Conference on Computer Application and System Modeling (ICCASM), vol. 13, pp. 595-598, October 2010.

[40] Hongyan Jing, "Sentence Reduction for Automatic Text Summarization," In Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, USA, pp. 310-315, 2000.

[41] G. Salton, "Automatic Text Processing: the transformation, analysis, and retrieval of information by computer," Addison- Wesley Publishing Company, USA, 1989.

[42] Jun'ichi Fukumoto, "Multi-Document Summarization Using Document Set Type Classification," Proceedings of NTCIR- 4, Tokyo, pp. 412-416, 2004.

[43] You Ouyang, Wenji Li and Qin Lu, "An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation," Proceedings of the ACL-IJCNLP Conference Short Papers, Suntec, Singapore, pp. 113–116, 2009.

[44] Mr.Vikrant Gupta, Ms Priya Chauhan, Dr. Sohan Garg, Mrs. Anita Borude and Prof. Shobha Krishnan, "An Statistical Tool for Multi-Document Summarization," International Journal of Scientific and Research Publications (ISSN 2250-3153), vol. 2, issue 5, 2012.

[45] Inderjeet Mani and Eric Bloedorn, "Multi-document summarization by graph search and matching," AAAI/IAAI, vol. cmplg/ 9712004, pp. 622-628, 1997.

[46] Rada Mihalcea and Paul Tarau, "Text-rank: Bringing Order into Texts," Proceeding of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004.

[47] Xiaojun Wan, "An Exploration of Document Impact on Graph-Based Multi-Document Summarization," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,

[48] Kathleen McKeown and Dragomir R. Radev, "Generating Summaries of Multiple News Articles," Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, pp. 74–82, 1995.

[49] Shanmugasundaram Hariharan, Thirunavukarasu Ramkumar and Rengaramanujam Srinivasan, "Enhanced Graph Based Approach for Multi Document Summarization," The International Arab Journal of Information Technology, 2012. [Online] Available:www.ccis2k.org/iajit/PDF/vol.10,no.4/4460-11.pdf (March 11, 2013)

[50] Tiedan Zhu and Xinxin Zhao, "An Improved Approach to Sentence Ordering For Multi-document Summarization," IACSIT Hong Kong Conferences, IACSIT Press, Singapore, vol. 25, pp. 29-33, 2012.

[51] Jade Goldstein, Vibhu Mittal, Mark Kantrowitz and Jaime Carbonell, "Multi-Document Summarization by Sentence Extraction," ANLP/NAACL Workshops. Association for Computational Linguistics, Morristown, New Jersey, pp. 40-48, 2000.

[52] Judith D. Schlesinger, Dianne P. O'Leary and John M. Conroy, "Arabic/English Multi-document Summarization with CLASSY - The Past and the Future," Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, Haifa, Israel, pp. 568–581, 2008.

[53] Nitin Agarwal, Gvr Kiran, Ravi Shankar Reddy and Carolyn Penstein Ros´e, "Towards Multi-Document Summarization of Scientific Articles:Making Interesting Comparisons with SciSumm," Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Portland, Oregon, pp. 8–15, 2011.