DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4788

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

QUERY-BASED SUMMARIZATION METHODS FOR CONVERSATIONAL AGENTS: AN OVERVIEW

Ketakee Nimavat
UG student , Computer Engineering,
L.D. College Of Engineering,
Ahmedabad, India

Prof. Hetal A. Joshiara
Computer Engineering Department
L.D. College Of Engineering,
Ahmedabad, India.

Abstract: Summarization is a topic that will be of a great important in the coming age since intelligent assistants especially the ones in the form of conversational agents will have to sift through the abundance of raw unstructured text data to provide relevant information. The data will be in the form of Social media posts, content websites and other user generated text content from which the user shall require tailored information from and about the data. The paper hence explores various methods for summarization and focuses particularly on extracting the gist from the perspective of a given keyword i.e. query based summarization from raw unstructured text data sources available at scale. Along with that, the need for a proper framework to mine relevant knowledge from the said data is acknowledged and the challenges that a conversational agent would hence face are identified. Various approaches that contribute to building a framework and solve the identified challenges are explored as well. It is hoped that the approaches discussed in the paper will be of use to researchers building algorithms in areas of knowledge mining and understanding, such as summarization, that deal with the challenges that are expected to arise.

Keywords: Query based summarization; Conversational agents; Raw unstructured text data; Text cubes;

I. INTRODUCTION

Today, we're living in the age of data deluge and intelligent systems which range from the algorithms that power our searches to conversational agents that are omnipresent in our surroundings like Alexa and Siri. All these systems are required to be able to answer our questions and make relevant data more accessible for us. Without these systems, mere data would turn out to be useless. These systems are hence the bridge between humans and data expected and are required to have the ability to understand and work with the knowledge to make it more accessible. The better they understand the context, the better they will be able to provide insights instead of mere answers. For example, Alexa must be able to read the news content from a news website and answer our questions about a certain event. It should be able to answer questions such as "How many spoons of sugar does the recipe require?" or "What does the chapter say about Osmosis?". Taking it a step further, it should give us the key points like a human reading the article would. Similarly, search results require extraction of content relevant to the query from the website and present the information as a short gist. Therefore, when asked, these assistants are perceived to understand the data and return the correct answers. Existing systems, such as the Google and Alexa's search mechanisms, focus on making structured information more accessible and they have successfully done so. But with the penetration of internet and particularly social media, unstructured content especially in text form is increasing in amount. Along with that, their volume which is visible to a certain user is huge as well. In these cases it is required that the agents that interact with the system can sift through this deluge and gives us either the required content or an overview of the content. This will be of increasing importance because one might not have the time, mental capacity or the attention span to go through every piece of information. We want our agents to be able to reduce this content to the essential parts of it and filter it for relevancy. Here, in this paper the methods that are seen fit to address the characteristics of data in the internet age and extract valuable information from text medium are discussed. Text is chosen because the techniques discussed here will be useful for conversational agents such as chatbots or virtual assistants and for systems such as search systems to extract relevant knowledge from unstructured and structured information. The technique being addressed is the summarization technique which minimizes the information, maintains the essence and mentions the important factors. Particularly, query based summarization is discussed which involves summarizing the given content from the perspective of a query. This is because a user might want to know about a specific subject the text talks about or has set up a filter to find emails that talk about a specific topic and give the gist of what the emails say. Query based summarization hence is quite useful in terms of knowledge mining where the subject of the knowledge to be mined is known.

In this paper, first types of summarizations are discussed in section II, followed by the need for query based summarization along with the features the systems of tomorrow will require in order to provide satisfactory services in section III. Section IV looks at methods that solves these queries and finally in section V the paper is concluded with observations and future directions.

II. TYPES OF SUMMARIZATIONS

Summarization involves understanding which are the important parts in a document and presenting them to the user. This can be achieved in various ways in terms of the process used, presented in various ways in terms of the kind of information returned, and offer different things to the user in terms of the scope of the information returned. Below are some classifications of approaches to summarization that have been already identified. Following which, approaches based on

usecases are identified. Below, various categories of summarization are discussed (as shown in Figure 1.) along with approaches that have been proposed to accomplish the given classification in [1] [2]. Along with that, new approaches to classifying summarization have been added as well.

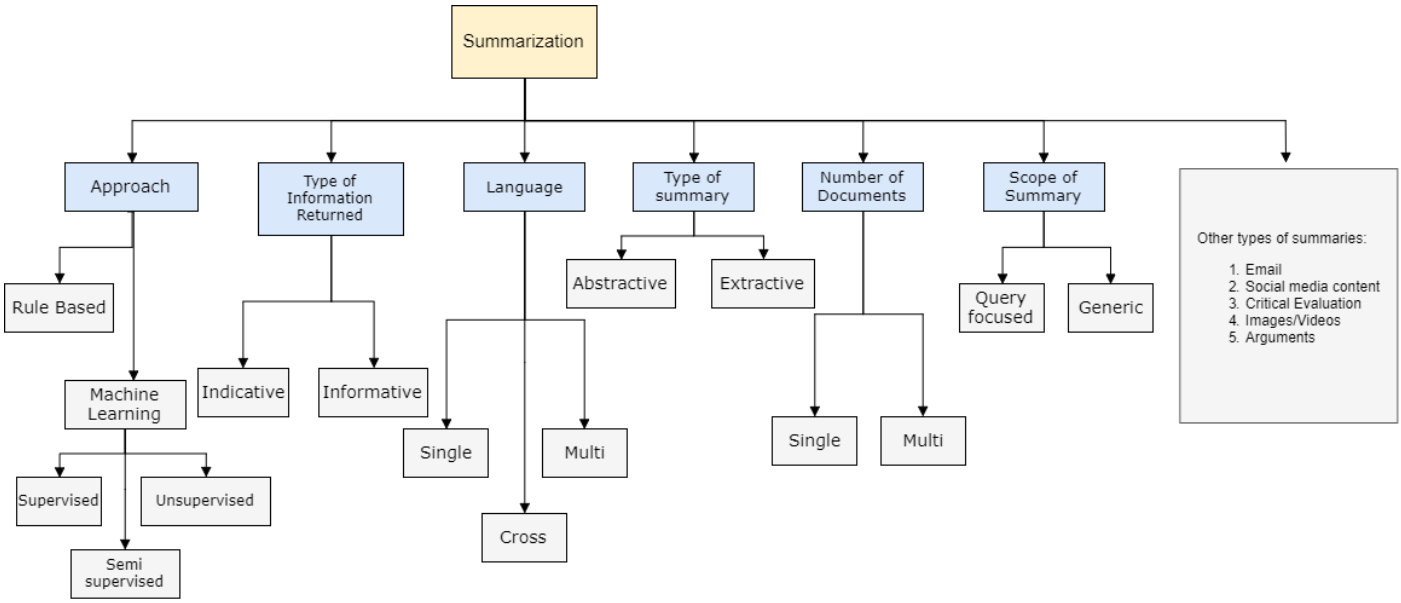

Figure 1: Classification of summarization based on various factors.

A rule based approach involves relevance rules and extraction rules that help decide which sentences to pick for adding to the summary. Machine learning approaches on the other hand either use a dataset to learn or learn dynamically.

- *Based on method of machine learning approach: supervised, semi-supervised, unsupervised*
  This classification is based on the way summaries are built using machine learning.
  Summarization can be achieved by supervised means and unsupervised means.[3][4] [5]
  Supervised summarization includes approaches such as text ranking using supervised methods. A dataset of text and its respective human generated summary is used to train a machine learning model such as classifiers that classify whether a sentence belongs in the summary or not. This type of summarization is hard to evaluate since each person would mention different summaries. Secondly, coming across labelled datasets for supervised summarization is difficult as well. Unsupervised on the other hand is more suitable since it requires less training data.[4][3]
- *Based on type of content returned: generic, query based*
  This classification is based on the type of content returned. If the content is generic and covers the entire text, it falls under generic summarization. On the other hand, query based summarization includes a summary relevant to the given query or keywords. [1]
- *Based on documents: multi document, single document*
  When the number of documents being summarized are more than one, it is a multi document summary. It includes gist of various documents altogether and is a multistep process which involves finding relevant documents and then finding relevant sentences and reducing redundancy. [6][7] Whereas a single document summary includes summary of only the given document.
- *Based on level of linguistic process: abstractive, extractive*
  Abstractive summaries consist of generating summaries using natural language generation techniques. The gist is interpreted and rephrased.[8]
  Extractive summarization consists of picking sentences that would describe the given text the most accurately. In this, the sentences are used as they are. [9]
- *Based on type of information returned: indicative, informative*
  Indicative summaries provide the metadata about the document, they give a bird's eye view of the document.
  Informative summaries provide an elaborate summary of the information contained in the text.
  [1][10]
- *Based on languages: multi, single, cross*
  When the source language and the language of the summary are the same, it is a single language summarization. Majority of the research focuses on single language summarization.
  When the document is available in multiple languages and so is the generated summary, it is multi document summary.[7]
- *Based on approach of building summarization: Rule based and Machine learning*
  This classification is based on how the system approaches the task of summarization.

Finally, cross document summary is when the source document and the summary are in different languages. One approach involves translation of the documents.[11]

Apart from these various other forms of summarization exist. They are:
: Summarizations in the form of reviews but they are a step ahead of summarization since they also consist of evaluation of code content.[1]
Web based summarizations: These involve summarization of content sourced from websites. An example of such summarizations would be summarizing search results or summarizing user reviews[12].[13]
Email based summarizations involve summarizing contents of an email along with attachments and summarizing the conversations that took place in a chain of email along with the conclusion reached. [14]–[16]
Summarizing tweets/social media posts is also an upcoming field since with the proliferation of social media and the amount of information available, it will become necessary to be able to summarize the texts into a concise representation of what went down.[17]–[19] Especially in times of disasters when social media turns out to be the only way possible to contact people.[20], [21]
Image groups/video summarizations consist of summarizing a string of images or a video. Stories about the subject can be generated from a string of images as well.[22]
Another possible mode of summarization is *summarizing arguments or negotiations* where the tool summarizes the stance taken by each party throughout the course of the argument by mentioning the salient features one claims for one's own stance. [23]
All these categories are not exclusive of each other, a summarization problem would require cross-lingual query based extractive summarization using supervised learning. Summarization is a method for extraction of core information the paragraph is trying to express. Methods that can build the metadata of the system by extracting the important factors that replicate human intelligence will help

III. NEED FOR QUERY BASED SUMMARIZATION

As seen above, a variety of approaches to summarization exist, here we focus on query based summarization. With the advent of conversational agents, query based summarization would see a wide usage. Conversational agents would require awareness of the context of the information they fetch. FAQ bots when asked a question would have to provide the specific information regarding the topic. Even general bots when asked questions similar to "What does this article say about health benefits?" would be required to return the content relevant to health benefits only.
With conversational interfaces such as Alexa being omnipresent along with chatbots such as those on messenger or Google Allo, the amount of space available for content display along with the attention span of the user will decrease. On the other hand, the expectations for the bots to provide understand the query and return the exact information will increase. The summarization and information extraction systems will have to keep up with them.

The way query based summarization differs from generic summarization is that generic summarization focuses on picking the important sentences from a text base where there are no other restrictions other than importance. In generic summarization, the task consists of detecting the relevance of the sentence to the query. Intuitively, one would first have to find out the sentences that are relevant to the query followed by picking the ones that would contribute to the summary.

As discussed above, the structure and scale of data available now are different. Due to the abundance of data, the way human brains deal with information has also changed. Search engines are very important since they pick the data we see. With voice search being more and more common, the way information will be presented has to change too since there's only so much a person can hear and remember. The information available has increased but so have the constraints associated with presenting and accessing the information. Below are the limitations and expectations that the systems will be required to consider while fetching and presenting the results.

***Attention span*:** In the coming times, the attention span will be reduced and also the expectations for tailored and specific answers will increase.

***Big and Raw data*:** Apart from that, the data will be massive and most of it will be unlabeled and unstructured. Training data hence will be sparse in spite of abundance of data.

Context and Length: Context will be of extreme importance since a conversational agent cannot give long irrelevant answers. The answers have to be highly semantically relevant to the user's query and concise.

Conversational Features: As text based algorithms pervade the space of human computer interaction, the systems will have to be more considerate of the expected features from a human companion such as spontaneity, speed and context awareness.

Hence, new methods to deal with data must include considerations for context, scale and dynamic nature of learning. In the coming section we discuss approaches that solve the specified challenges and might prove to be a stepping stone for the coming algorithms.

Quite expansive reviews of techniques of summarization have been provided in[2] and [1]. For query based summarization, in the early 2000, two noteworthy approaches involved query based summarization of web based documents [13] and centroid based summarization [24], topics that are of great relevance today. Other often used methods include relevance detection, keyword extraction, minimum edit distance etc.[28]–[31]. Apart from that, it has been observed that common approaches to query based summarization can be classified into three ways :[25]
Document graph: Here, text is processing by converting texts into graphs which can be used to draw references or manipulate. Graphs of the document and the query are compared to obtain relevant and important sentences.

Linguistic: Linguistic approaches involve using of lexical rules and clues to select the sentences that would fall under the summary.[26]

Machine learning: These approaches use machine learning methods such as supervised learning, unsupervised learning and semi-supervised learning.[27]

Evaluation of summaries is just as important as well. However, since summaries are subjective, it is difficult to evaluate without a common standard. Due to which summaries are often evaluated using ROUGE [32] which are industry standards.

The various approaches discussed in the coming section use either the document graph approach(A) or a linguistic approach or the machine learning approach but also propose new representations of data to assist in working with data at scale. They also have shown an improvement over previous models in terms of ROUGE scores.

IV. NOVEL TECHNIQUES:

In this section, we look at recently proposed approaches that cover deeper issues such as dealing with data at scale, semantic context awareness in summarization, topic awareness, and redundancy in generation. These are the issues that will help make summaries more useful as interaction of machines with text content becomes more commonplace. It will also help in language understanding and in generating more user friendly and efficient responses from conversational agents and other systems such as search engines.

The first two methods focus on semantics and awareness of the topics. The third method is about dealing with large amount of unstructured text and extracting query based summary from the text. The final method involves an abstractive method that reduces redundancy and proposes an attention driven model that focuses on different parts of text at different times.

A. Query based summarization using non-negative matrix factorisation[6]

The paper presents an extractive summarization Algorithm that doesn't involve training but instead uses Non Negative Matrix Factorization (NMF) which is capable of extracting semantic features naturally. Because of its inner representation, the need for complexes processes such as transforming documents to graphs in invalidated as well. It summarizes document using semantic features and semantic variables. The use of NMF enables the algorithm to make the distinction between two statistically similar but semantically different sentences such as "John calls Alex" and "Alex calls John" which in turn increases the accuracy of the sentences chosen for the summary.

The approach used in here could provide semantically correct information without needing a ton of data for training. It is usable in systems such as chatbots that require immediate and semantically correct answers where errors of redundancy in the abstractive method would discourage usage of the bot.

B. Query based multi document summarization using linguistic knowledge and content word expansion[33]

The approach tackles the issue of extractive summary while considering the semantic relations between words and the syntactic relations. It uses two similarity metrics: Sentence2sentence(s2s) similarity score and sentence2Query(s2q) similarity score. It defines a method called QSLK which represents the document as a graph. Document sentences are the nodes of a graph whereas s2s similarity score and s2q similarity score are the edges.

The approach has two stages:

1. SSCM(Statistical Semantic Comparison Model) and
2. CM (Combination Model)

where CM represents the sum of similarity to other sentences and the sum of similarity to the question. Following assigning the scores, the sentences are ranked in the descending order. From this, the n high scored sentences are chosen. To remove redundancy only those that are not too similar to other candidates are selected. The approach uses Word expansion as well which bridges lexical gaps. The semantic inclusion reduces errors of context. The approach hence deals with semantical over fitting and lexical under fitting.

Further work can be done to incorporate active and passive voice awareness in the model. It could also be tested on a different knowledge base and even a field specific knowledge base. Here, it uses Wordnet as a semantic knowledge base which is claimed to limit its understanding.

This method is particularly suitable for systems such as search queries where the query has to be understood semantically for the meaning rather than just comparing words.

C. Multi-Dimensional, Phrase-Based Summarization in Text Cubes[34]

This approach deals with scale and accessibility. A Generalized platform to support efficient online and offline computational optimization along with an architecture to store and analyse raw text data are proposed which would make analysing text and accessing specific information easier. It presents a way to structure, explore and extract information from large amount of raw text data. It holds an edge over relational databases since they lack support for analysing free text the way it is available on the internet. It is claimed that neither do they have support for Datacube technologies, and integrated analysis of traditionally structured and raw text data. The paper consists of an approach based on multidimensional attributes and where each cell is a subset of documents. Semantically close cells are found using context which is a function of parent, child and sibling score. Representative phrases that are used to define or represent the content of a cell are chosen if they hold the following characteristics:

- Multiple phrases collocate together more frequently than by random chance
- Phrase is a complete semantic unit rather than a subsequence of another equally frequent phrase

Apart from that, metrics such as popularity (multiple occurrences) and distinctiveness (Less background noise such as 'earlier this month') are used as well. Score between 0 and 1 is used to characterize the degree of each phrase satisfying the criteria

The algorithms names two major benefits:

It allows analysing statistical features of all documents together along with logical categories such as correlation between word frequency and publishing time. It also allows for contextualized analysis and uses semantic clusters instead of phrases to reduce semantic redundancy

We choose this because it provides a framework to represent and extract relevant data at scale via an online platform. This makes data access at scale easier and more accurate. It could be built upon to be integrated into real time social media analytics platforms such as Hootsuite or could be used for fetching query results from text based sites like Wikipedia or for on the spot

summarization/ question answering of text heavy social media platforms such as Reddit and Facebook.

D. Diversity driven Attention Model for Query-based Abstractive Summarization[8]

The paper proposes an abstractive model that overcomes the challenges of traditional encoder-decoder approaches to build a non-redundant abstractive summary. Traditional Encode-attend-decode approach generates repeated phrases which is dealt with using

1. Query attention model (focuses on different portions of the query at different times, hence a dynamic representation of the query)
2. Diversity based attention model to remove the issue of repeated phrases.

Typical encoder-decoder produce word by word contextual summary where a new context vector to the decoder at each time step by attending to different parts of the document and query. This often causes repetition of words in the produced summary. The model prevents the same words by assuring successive context vectors are orthogonal to each other. However, only exactly previous words are considered while checking for the orthogonal nature. The components that cause the two vectors to be in the same direction are removed. At each time step, query representation is dynamically computed. Doing so improves the results- meaning the model learns to focus on different portions of the query at different time steps

The experiment is performed on a custom dataset called Debatepedia. LSTM based diversity model is seen to give the best results in terms of redundancy removal and a 28% gain in ROUGE-L scores in summarization. The diversification model proposed here can be useful for generic natural language generation tasks as well.

The approach is particularly interesting for its attempt at reducing redundancy and the way it deals with the query. This could be useful in creating generated responses that are dynamic and spontaneous rather than extraction based responses.

V. CONCLUSION:

In this paper, we have explored various types of summarization and identified the need for query based summarization. Apart from that, issues such as context awareness and scale which will be faced by conversational systems interacting with the raw data are identified. Following which, approaches that solve the identified issues are discussed and the possible implementations of the given problem are discussed. It is hoped that the approaches will give other researchers possible problem areas to explore such as extracting relevant information from vast amount of unstructured data and will provide a starting point to find insights about possible solutions.

VI. REFERENCES

[1] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," Artif. Intell. Rev., vol. 47, no. 1, 2017.

[2] A. Nenkova, "Automatic Summarization," Found. Trends® Inf. Retr., vol. 5, no. 2, pp. 103–233, 2011.

[3] S. Fisher and B. Roark, "Query-Focused Summarization By Supervised Sentence Ranking and Skewed Word Distributions," Proc. 6th Doc. Underst. Conf. . DUC, 2006.

[4] K.-F. Wong, M. Wu, and W. Li, "Extractive Summarization Using Supervised and Semi-supervised Learning," Proc. 22nd Int. Conf. Comput. Linguist. 1. Assoc. Comput. Linguist. 2008., no. August, pp. 985–992, 2008.

[5] L. Logeswaran, H. Lee, and D. Radev, "Sentence Ordering using Recurrent Neural Networks," pp. 1–15, 2016.

[6] S. Park and B. R. Cha, "Query-based multi-document summarization using non-negative semantic feature and NMF clustering," Proc. - 4th Int. Conf. Networked Comput. Adv. Inf. Manag. NCM 2008, vol. 2, pp. 609–614, 2008.

[7] V. Gupta, "Hybrid Algorithm for Multilingual Summarization," pp. 717–727, 2013.

[8] P. Nema, M. Khapra, A. Laha, and B. Ravindran, "Diversity driven Attention Model for Query-based Abstractive Summarization," 2017.

[9] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive techniques," J. Emerg. Technol. Web Intell., vol. 2, no. 3, pp. 258–268, 2010.

[10] M. Allahyari et al., "Text Summarization Techniques: A Brief Survey," no. 1, 2017.

[11] J.-G. Yao, X. Wan, and J. Xiao, "Phrase-based Compressive Cross-Language Summarization," Conf. Empir. Methods Nat. Lang. Process., no. September, pp. 118–127, 2015.

[12] D. Wang, S. Zhu, and T. Li, "SumView: A Web-based engine for summarizing product reviews and customer opinions," Expert Syst. Appl., vol. 40, no. 1, pp. 27–33, 2013.

[13] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal, "Probabilistic question answering on the Web," J. Am. Soc. Inf. Sci. Technol., vol. 56, no. 6, pp. 571–583, 2005.

[14] G. Carenini, R. T. Ng, and X. Zhou, "Summarizing email conversations with clue words," Proc. 16th Int. Conf. World Wide Web - WWW '07, p. 91, 2007.

[15] A. Nenkova and A. Bagga, "Facilitating Email Thread Access by Extractive Summary Generation," Recent Adv. Nat. Lang. Process. III, Sel. Pap. from RANLP'03, vol. 260, pp. 287–296, 2003.

[16] O. Rambow, L. Shrestha, J. Chen, and C. Lauridsen, "Summarizing Email Threads," Proc. HLT-NAACL 2004 Short Pap. XX - HLT-NAACL '04, pp. 105–108, 2004.

[17] M. A. H. Khan, D. Bollegala, G. Liu, and K. Sezaki, "Multi-tweet summarization of real-time events," Proc. - Soc. 2013, no. September, pp. 128–133, 2013.

[18] L. Shou, Z. Wang, K. Chen, and G. Chen, "Sumblr: continuous summarization of evolving tweet streams," Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '13, p. 533, 2013.

[19] X. Liu, Y. Li, F. Wei, and M. Zhou, "Graph-Based Multi-Tweet Summarization using Social Signals.," Coling, vol. 2, no. December 2012, pp. 1699–1714, 2012.

[20] T. Mondal, P. Pramanik, I. Bhattacharya, A. Saha, and N. Boral, "Towards development of FOPL based tweet summarization technique in a post disaster scenario: From survey to solution," 2017 51st Annu. Conf. Inf. Sci. Syst. CISS 2017, 2017.

[21] J. B. S. Ong, Z. Wang, R. S. M. Goh, X. F. Yin, X. Xin, and X. Fu, "Understanding Natural Disasters as Risks in Supply Chain Management through Web Data Analysis," Int. J. Comput. Commun. Eng., vol. 4, no. 2, pp. 126–133, 2015.

[22] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2714–2721, 2013.

[23] A. Fuji and T. Ishikawa, "A System for Summarizing and Visualizing Arguments in Subjective Documents: Toward Supporting Decision Making," in Proceedings of the Workshop on Sentiment and Subjectivity in Text, 2006, vol. 69–72, no. July, pp. 15–22.

[24] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies," Inf. Process.

Manag. 40.6 919-938., vol. 40, no. 6, p. 10, 2000.
[25] M. Damova and I. Koychev, "Query-Based Summarization : A survey."
[26] J. M. Conroy and J. G. Stewart, "CLASSY Query-Based Multi-Document Summarization," Proc. DUC2005, 2005.
[27] Q. M. Summarization, S. D. Silva, N. Joshi, S. Rao, S. Venkatraman, and S. Shrawne, "Improved Algorithms for Document Classification &," vol. 3, no. 4, 2011.
[28] D. J. Brenes, D. Gayo-Avello, and K. Pérez-González, "Survey and evaluation of query intent detection methods," Proc. 2009 Work. Web Search Click Data - WSCD '09, pp. 1–7, 2009.
[29] H. Daumé, "Bayesian Query-Focused Summarization," 2009.
[30] L. Wang, H. Raghavan, V. Castelli, R. Florian, and C. Cardie, "A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization," 2016.
[31] S. Gupta, A. Nenkova, and D. Jurafsky, "Measuring importance and query relevance in topic-focused multi-document summarization," Proc. 45th Annu. Meet. ACL Interact. Poster Demonstr. Sess. - ACL '07, no. June, p. 193, 2007.
[32] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," Proc. Work. text Summ. branches out (WAS 2004), no. 1, pp. 25–26, 2004.
[33] A. Abdi, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev, "Query-based multi-documents summarization using linguistic knowledge and content word expansion," Soft Comput., vol. 21, no. 7, pp. 1785–1801, 2017.
[34] F. Tao et al., "Multi-Dimensional, Phrase-Based Summarization in Text Cubes," Data Eng., p. 74, 2016.