Chapter 1: Introduction

Abstract

Recently, data is growing fast in every domain such as news, social media, banking, education. With the increasing amount of information, it has become diﬃcult to take out concise information. There is a need of automatic summarizer which will be capable to present human quality summaries and summarize the data especially textual data in original document without losing any incisive purposes.

This Automatic text summarization is a tool that provides summaries of a given document. This Tool can be use on news stories, blog articles and social media posts. Most Important Thing is this supports a wide variety of languages. This Summarization tool can be used to summarize one single document on a percentage. Summarization base on a keyword and generate a summary report can be considered as some extra options that user can select.

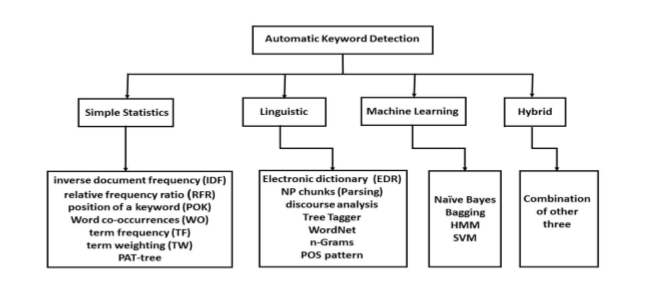
This will be ideal for any user as this produce short length text which comprises all the key information of the document. This is also appropriate for users who need to summarize the entire document based on one specific keyword. This will reduce using several applications such as word and google translator because this itself can generate a summary report and also can summarize documents with own native languages.

Text summarization is emerged as an important research area in recent past. In this regard, study about existing work on text summarization process is useful for build a new solution. It was also a necessity to effort to examine the needs of the solution addressed problems would enable a wider range of people to be aware of important information.

1.1: Project Background

The goal of text summarization Tool is to Publish the contents of a document in a concise form of meeting a user's needs. Online information is available for readers in the form of News information, journal articles, technical reports, biographical information it isn't possible to read everything and There are vast number of documents available in these digital media and extracting only relevant information from all these media is a dull job for the individuals in specify time. so some form of information condensation is needed that can extract only pertinent information from these data sources. To Realize this, from the documents texts mining is necessary The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. It includes tasks like automatic keyword extraction and text summarization. Automatic keyword extraction is the process of selecting words and phrases from the text document that can at best project the core sentiment of the document without any human intervention depending on the model. [1]. Summarization is a process where the most salient features of a text are extracted and compiled into a short abstract of the original document [2]. The important keywords extraction is the primary phase of text summarization. In keyword extraction as in the literature different methodologies used for keyword extraction process and different algorithms used under each methodology as shown in Figure 1.

Automatic Keyword Extraction for Text Summarization: A Survey. [5]



**Figure 1**: Classification of automatic keyword extraction on the basis of approaches used in existing literature

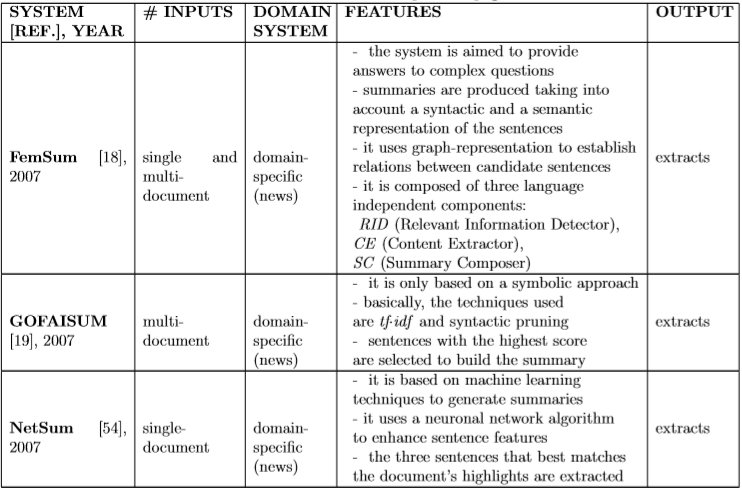
**Simple Statistical Approach -** These strategies are rough, simplistic and have a tendency to have no training sets. They concentrate on statistics got from non-linguistic features of the document, for example, the position of a word inside the document, the term frequency, and inverse document frequency. These insights are later used to build up a list of keywords. Cohen [6]

**Linguistics Approach -** This entry uses linguistic features to identify and extract the key words of the documents in the text. It incorporates the lexical analysis [7], syntactic analysis [8], discourse analysis [9], etc.

**Machine Learning Approach -** This approach requires manually annotated training data and training models. Hidden Markov model [10], support vector machine (SVM) [11], naive Bayes (NB) [12], bagging [8], etc. are commonly used training models in these approaches

**Hybrid Approach -** These approaches combine the above two methods or use heuristics, such as position, length, layout feature of the words. [13]

There are current uses that can be summarized as shown in Figure 2. Following is brief description about former Text Summarization Systems.



**Figure 2:** Text Summarization Systems [14]

Automatically summarized procedures can be divided as follows,

**Interpret -** This is where a representation of the document to be summarized is produced.

**Transform** - This is produced from where the representation of the document is turned into one of a summary of the document.

**Generate** - Here, summary text is the summary representation. Also known as synthesis.

The different dimensions of text summarization can be generally categorized based on its input type. [19]

**Single Document summarization** -Single document summarization produces summary of single input document. [19]

**Multi Document Summarization** – On the other hand, multi document summarization produces summary of multiple input document. [19]

Base on purpose summarization can be divided in to different categories.

**Generic Summarization** - Generic summarization purpose is to summarize all texts regardless of its topic or domain. Generic summaries make no assumptions about the domain of its source information and view all documents as homogenous texts. [19]

**Domain Specific Summarization** - this type of summarization requires domain specific. [19] knowledge bases to assist its sentence selection process.

**Query focused summarization** - Query-based summary contains only information which are queried by the user. [19]

Also Summaries can be divided in to Different categories base on output type.

**Extractive** - extracts are produced by identifying important sentences which are directly selected from the document. [19]

**Abstractive** -In abstractive summarization, the selected document sentences are combined coherently and compressed to exclude unimportant sections of the sentences. [19]

Also there are categories of summarization Based on language.

**Mono-lingual summarization** -Here input and Target document in the same language.

**Multi-lingual summarization** -Here if the input document in multiple languages summary is also contain these languages.

**Cross-lingual summarization** -Here input document in one language and output document to be in another language.

In this project automatic text summarization is subject to a single document summarization with multiple languages. Also, the summaries produced are largely extracts of the document being summarized, rather than newly generated abstracts there for this produce summaries base on extractive method. Also this follows Simple Statistical Approach. Since this summarizer text in queried by the user this is a Query focused summarization. Mono-lingual summarization, Cross-lingual summarizations are two type of summarizations which are based on language.

Create a table here.

1.2: Problem Landscape

1.3: Objective

Chapter 2: Project Management Tools

The Project Management approach selected should be that which will deliver the best value for money in terms of getting the job done and ensuring adequate control. So clearly, with a small project, it is both practical and sensible to adjust the project management approach to the size of the project. project management method offers a convenient and effective structure for the management of IS projects. [18]. Below are the main tools that help to carry out the project.

* + Trello
  + TODO plugin
  + GitHub
  + GitKraken
  + NetBeans
  + Google Drive

2.1: Project Management

The Automatic Text summarization project is a result that comes from a recent strategic review. The analysis showed that automatic text summarizer has a good reputation in its market place.

2.1.1: App Development Methodology

There is a considerable effect of app development methodology on each factor of the star model of project management.

**Budget** - Selected methodology determines whether it can assist to complete the project within the budget and to provide a good return on investment. Since this ITS is a prototype this is free of budget.

**Schedule** - Selected methodology determines whether it can assist to project to deliver on time, to project to get satisfactory requirement, to follow workflow as described in the scheduling. ITS project need to be deliver on time. As there were limited time period to this ITS project there should be a proper schedule. There must be greater transparency of the project. Sometimes Due to external factors had to do rescheduling. As the scheduling always provide plan B this was possible.

**Scope** - Selected methodology determines whether it can assist to have the well-defined scope and to have very clear scope of the project. In order to successfully get the final outcome of ITS project which is expected, also had to avoid the project is being way behind the schedule there for a well-defined scope were very important.

**Risk** - Selected methodology determines whether it can assist to manage project risks and opportunities and to meet the business objectives.

**Resources** - Selected methodology determines whether available material resources and whether it can assist to achieve maximum utilization of available resources.

**Quality** - Selected methodology determines whether it can assist to meet quality requirements and to meet client satisfaction to successful the project. [20]

collecting requirements is more signiﬁcant in light weight and in heavy weight methodologies. Also that in-time delivery is equal in both methodologies. The project scope is well deﬁned and clear compared to the heavyweight methodologies. Lightweight methodologies and heavyweight methodologies are of greater signiﬁcance for the availability and better utilization of resources. Since ITS project is a prototype which is individual I used agile development methodologies continue the development of the project.

2.1.2: Project Planning

Project planning is very important for the successful completion of the project. Those involved in the project need to know exactly what their role is, and what they expect to produce and, when it is wanted. This information will inform by The project plan. This gives the opportunity to think about what the project is about and how it is to be achieved, deliverables and when are they wanted, what skills we need and where do we get them, the problems we are likely to encounter and how shall we tackle them, the risks involved and how to overcome them. Since this automatic text summarizer is an individual project I thought about above all mentioned criteria. Following is the sequence of the process of project planning to automatic text summarizer.

2.1.2.1: **Understanding the requirement**

After going through several current use of automatic text summarization, it is observed that till now most of the summarization tools have developed with less features. Most of them can only use for summarize a given entire document at a predefine level. There for user does not have a chance to give a summarization level which user is preferred. User need to strike to a one specific summarization. Also some of current uses could not able to summarize multiple language documents. When go through these several current uses I could found no current summarization tool has the feature of summarizing a given document based on user preferred key word. To overcome these, I proposed a prototype of new intelligent automatic text summarizer which provide accurate and concise and ﬂuent summaries of longer text using diﬀerent summarization techniques by Retrieval of most important information with more usability in response to a user document. Following are the business objectives in this Proposed ITS approach.

1. Summarize a given document based on a user preferred key word.
2. Summarize a given document according to many summarization level.
3. Summarize documents with multiple languages.
4. Generate a summary report.

2.1.2.2: **Breaking down the work**

By looking at objectives of the project it is need to consider what need to be done to achieve that goal. What is trying to produce and how to archive it. The work breakdown structure and product break down structure are two basic approaches to do this. Both structures will be illustrated in Figure 3 and Figure 4.

2.1.2.2.1: Work breakdown structure.

**Figure 3.**

2.1.2.2.2: Product breakdown structure.

**Figure 4.**

2.1.2.3: **Understanding dependencies**

Dependencies are fundamental to planning a project and, later, in understanding the effects of any problems encountered. [18]. The dependencies between the product/deliverable are illustrated in Figure 5.

2.1.2.3.1: Dependency diagram.

**Figure 5.**

2.1.2.4: **Using planning tools.**

It may take time to create various diagrams and charts for a project. As plans are requiring adjustments and revision during project progresses they cannot be produce at the start of the project. Use of computerized tools make replanning less of a chore. Many project planning tools are there on the market which are created for use on computers. Advantages of planning tools helped in following ways.

**Make Easy re-planning** – Normally planning tools produce a project breakdown, create estimates, produce a schedule. With planning tools, I could do some adjustments to my plan in some points due to not working the plan. As planning tools have ability to review estimates and reschedule this was possible.

**Presentation quality were high** – Ability to have high quality output in a variety of formats as they have flexible presentation and reporting facilities.

**Keep track with the progress** - Now a days most planning tools have facilities to track progress on a project. Ability to compare actual progress with the plan. Then could identify where problems seem to be arising and to investigate them and decide how to respond them.

From the start of the project to the implementation level and testing level variety of project management tools dynamically alter to manage schedule, scope, risk, Resources, Quality. Below are the main tools that helped to carry out the ITS project to the success of the software project.

* Trello
* TODO plugin

Below is a screen shot of Trello boards that demonstrating tasks of the ITS project.

**Figure 6.**

2.2: File Management.

Around 90% of the files in ITS related with the development code base which were maintained by using a private git repository on GitHub. The remaining files were managed through a google drive cloud storage, where were in sync with local drive, as well as git repository files.

Advantages of using a git repository to ITS project shown as below:

* View the code changes across versions.
* If an error occurred there were ability to revert the code files/ revert the entire project back to the previous state.
* Due to flexibility it efficiency independent from the scale of the project.

Below is a screen shot of Git repository of the ITS project.

**Figure 7.**

2.1.2: Challenges

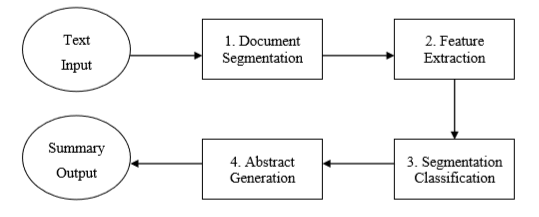
* During the project planning phase could discover following challenges.
* At the start scope of the project was unclear.
* The planed schedule for some tasks not accurate.

Chapter 3: Design

3.1: Requirements

3.2: Architecture

The summarizer is written in Java following the MVC architecture. This uses the following external library itextpdf-5.1.0, commons-logging-1.2.jar, httpclient-4.5.5.jar, httpcore-4.4.9.jar and java-json.jar. The ITS project followed the general summary process and totally there are four steps which come as document segmentation(**Interpret**), feature extraction **(Transform)**, segmentation classification(**Transform**), and abstract generation(**Generate**). As shown in the Figure 3.



**Figure 4:** Summarizer architecture [15]

The process of summarization begins with processing of input document using my naive summarization algorithm. This algorithm broken down into paragraphs and subsequently into sentences. The tokenization includes tasks such as paragraph detection base on no of consecutive new line feed characters, extract sentences base on appearance and position of periods in the entire text. This maintains a list of sentences of the document and list of paragraphs in the document Each paragraph is responsible to store the sentences contained in it. To conduct this process, give two numbers which indicated which paragraph the sentence is apart and second which indicated sentence respect to the entire text. The second step in the summary creation process is feature extraction which is features are extracted from the text and intermediate values calculated to them. For example, scores assigned to features such as capitalization, phrases, or the sentence length, words population, word importance (key word). Also Summarizer maintains intersection matrix to keep intermediate values. Thirdly, for each sentence in the context, its score is computed as a weighted sum of feature intermediate values. Last, in step four, the highest scoring sentences are used to generate the document summary. This uses a sentence comparator on score to compare two sentence base on score and Sentence comparator for Summary compare two sentence base on its sentence number. To reads byte stream from a file system FileInputStream. java.io. FileInputStream used. My native algorithm is such an important subject which is under 3.6 itself later in the report.

The Translation API's-TO DO

3.2.1: UML Class diagram

**Figure 5:**

Here design the classes according to the MVC architecture. There for this propose three types of objects in an application the Model, Views and Controllers.

**Model** - Objects hold data and define the logic for manipulating that data [16].

In this summarization application Paragraph, Sentence, GCPTranslator, are model classes as shown in the Figure 4.

**View** - objects represent something visible in the user interface [16].

In this summarization application View is the view class as shown in the Figure 4.

**Controller** - object acts as a Mediator between the Model and View objects. A Controller object communicates data back and forth between the Model objects and the View objects [16].

In this summarization application Algorithm, SentenceComparatorForSummary, SentenceComparatorOnScore, Validator are controller classes as shown in the Figure 4.

3.2.1: Class Diagram

3.3: Challenges

3.3.1 Implementation Challenges

1.To get the quality summary, quality keywords are required for text summarization.

2. There is no standard to identify quality keywords within or multiple documents. The extracted keywords are varying for applying different approaches of keyword extraction.

3. Multi-lingual text summarization is another challenging task.

ISSUES AND CHALLENGES OCCURS IN TEXT SUMMARIZATION In the area of text summarization, there are following research issues and challenges occurs during implementation

1.In the case of multi-document text summarization, several issues occurs frequently while evaluation of summary such as redundancy, temporal dimension, co-reference or sentence ordering, etc. which makes very difficult to achieve quality summary. Some other issues occurs such as grammaticality, cohesion, coherence which is harmful for summary.

2. The quality of summaries are varying from system to system or person to person. Some person feels some

set of sentences are important for summary, at the same time other person feel the other set of sentences are important for required summary.

3.3.1: Scalability

3.4: Programming Languages

3.5: Classification of Summarization Tasks

3.6: Methodology

3.7: Summarization Algorithm

3.8: Text Translate API

3.9: Unit Testing

Chapter 4: Development

4.1: Tools

4.1.1: IDE

4.1.2: Debugging

4.1.3: Other tools

4.2: Application

4.2.1: UI

4.2.2: Content

4.2.3: Features

Chapter 5: Testing

5.1: Test Plan

5.1.1: Market Fit

5.1.2: Performance and Reliability

5.2: Test Results

5.2.1: Market Fit

5.2.2: Performance and Reliability

Chapter 6: Experimental Results

6.1: Comparison of Summaries Generated with Hand Written Summaries

Chapter 7: Evaluations

7.1: Project Planning

7.2: Design

7.3: Future Development

7.4: Gained Lessons

7.5: Conclusion