CHAPTER 1

INTRODUCTION

1.1 INFORMATION EXTRACTION

Information Retrieval refers to the human-computer interaction (HCI) that happens when we use a machine to search a body of information for information objects (content) that match our search query. A Person's query is matched against a set of documents to find a subset of 'relevant' documents, while Information Extraction (IE) is a type of information retrieval whose goal is to automatically extract structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Systems that perform IE from online text should meet the requirements of low cost, flexibility in development and easy adaptation to new domains. Information extraction was originally applied, to identify the desired information from a natural language text, and convert it into a self-defined presentation, e.g., a database for particular fields. With the huge and rapidly increasing amount of available information sources and electronic documents on the world wide web, information extraction has been extended to the identification of relevant information from structured and semi-structured web pages. Recently, more and more research groups have begun to concentrate their attention on the development of information extraction systems. Researches on information extraction could be divided into two subareas: the extraction patterns used for the identification of target information from a given text, and using machine learning techniques to automatically build such extraction patterns for the sake of avoiding expensive construction by hand. Actually, many information extraction
systems have been successfully implemented, and part of them perform very well; i.e., they operate much faster than humans and have an accuracy comparable to that of manual work.

1.1.1 Information Extraction and Text Summarization

At a high level, the goal of Information Extraction (IE) and text summarization is the same: find those portion(s) of the given text(s) that are relevant to the user’s task, and deliver that information to the user in the form most useful for further (human or machine) processing. Considering them more closely reveals the fact that IE and Summarization are two sides of a coin, and that a different emphasis of output and techniques results in two quite different-looking branches of technology. In both cases, the input is either a single document or a (huge) collection of documents.

The differences between IE and Summarization lie mainly in the techniques used to identify the relevant information, and in the ways that the information is delivered to the user. Information Extraction is the process of identifying relevant information, where the criteria for relevance are predefined by the user in the form of a template that is to be filled. Typically, the template pertains to events or situations, and contains slots that denote who did what to whom, when, and where, and possibly why.

1.2 WHAT IS TEXT SUMMARIZATION?

Text summarization (or rather, automatic text summarization) is the technique in which a computer automatically creates a summary of one or more texts. The early interest in the automatic shortening of texts was initiated during the sixties in American research libraries. A large amount of scientific papers and books were to be digitally stored and made searchable. However, the storage capacity was very limited, and full papers and books
could not be fitted into databases in those days. Therefore, summaries were stored, indexed and made searchable. Sometimes, the papers or books already had summaries attached to them, but in cases where no readymade summary was available, one had to be created. Thus, the technique has been developed for many years (Luhn 1958, Edmundson 1969) and in recent years, with the increased use of the Internet, there has been an awakening interest in summarization techniques.

Digitally stored information is available in abundance, and in a myriad of forms to such an extent, as to making it nearly impossible to manually search, sift and choose which information one should incorporate. This information must instead be filtered and extracted, in order to avoid the data being lost forever. The approach to and the end-objective of summarizing documents determine the kind of summary that is generated. Summaries can be indicative, informative or critical, based on their purpose. Depending on its form, a summary can be classified as extractive or abstractive, the former being an actual representation of paragraphs, sentences or phrases from the original document, and the latter being a concise summary of the central subject matter of a document.

**1.2.1 Need for Text Summarization and its Approaches**

With massive amounts of textual data available, an important problem is how to efficiently process these data, to meet the user’s information need. There has been an increasing interest recently in automatically processing the text data, including summarization, and other understanding tasks in the research community.

Automatic summarization is a very useful technique to facilitate users to browse large amounts of data efficiently. Summarization can be divided into different categories along several different dimensions (Evans
Based on whether or not there is an input query, the generated summary can be query-oriented or generic; based on the number of input documents, summarization can use a single document or multiple documents; in terms of how sentences in the summary are formed, summarization can be conducted using either extraction or abstraction — the former only selects sentences from the original documents, whereas the latter involves natural language generation.

Summaries generated by both humans and machines can be of different types. Though, studies on text mining do not provide an exhaustive list of the summary types, the following list provides the means for knowing the existing summary types:

- an **Extract** is a selection of some material of the original, while an **Abstract** is a condensation and reformulation of the original;

- a **Generic** summary provides the author’s point of view, while a **Query-based** summary focuses on matters of interest to the user;

- an **Informative** summary reflects the content of the original text, possibly spelling out the arguments, while an **Indicative** summary merely provides an indication of what the original was about;

- a **Just-the-News** summary provides just the newest facts, assuming the reader is familiar with the topic, while a **Background** summary teaches about the topic;

- a **Neutral** summary tries to be objective, while a **Biased** summary extracts and formulates the content from some point of view.
Overall, automatic summarization systems aim to generate a good summary, which is expected to be concise, informative, and relevant to the original input. A lot of techniques and approaches have been proposed for automatic text summarization in the past decades, and there are some benchmark tests such as TIDES, AQUAINT, MUC, DUC and TAC (Reitter 2003, Teich & Fankhauser 2004, Saggion 2005, Miller & Fellbaum 2007).

1.2.2 Applications of Text Summarization

Nowadays, man is surrounded by summaries of information that one often takes them for granted. Imagine a newspaper without headlines! Books and movies are described in blurbs and reviews, scholarly articles begin with abstracts, and search results are summarized with a few snippets from each page. Summaries are of great potential value to journalists, lawyers, students, and casual browsers of the Internet.

Summaries aid users to evaluate the relevance of a document without reading the full text. Since a summary is not cluttered with detail, a user can quickly recognize its relevance. Summaries could be displayed in search results as an informative tool for the user. For example, the Google search engine provides snippets of search results. Digital libraries and journals could make use of summaries. Users of the library or readers of the journal could benefit from summaries and find the relevant text easily. News portals could provide precise summaries about a news collected from multiple source articles. Columbia University has a system called Newsblaster (Fattah & Ren 2008) for this purpose. Web browsing with web site summaries could change our browsing habits, and enable us to filter irrelevant web pages.

Ongoing research investigates the summarization of email, legal proceedings, customer reviews, search results, meetings, and videos. Many of the standard techniques, however, have their origins in document
summarization, where the goal is to convey the most important information from a set of documents within a length constraint using natural language. Though a number of attempts have been made on text summarization, the field requires a lot more research as to how the statistical and prediction models improvise the approaches that exist in text summarization. The following sections investigate the roles of the popular paradigms, namely, regression, classification and refinement. This thesis gives a new insight into these paradigms, and exploits their benefits in extractive text summarization.

Applications that make use of documents of a specific domain or those that belong to known collections of data, make use of supervised methods. These methods need classified datasets for training, that in the specific case are pairs of documents and trusted summaries. These classifiers trained over a collection can hardly be used for another one. For summarizing generic documents that belong to unknown collections of data or blogs that are posted on the internet sites such as those that are dynamic in nature, supervised methods will be of little use, while unsupervised methods like clustering will be of better choice.

1.3 ROLE OF REGRESSION IN SUMMARIZATION

Text Summarization is a vital aspect of data mining; the most relevant data mining tasks are: association, sequence or path analysis, clustering, classification, regression, and visualization. The problem of regression consists in obtaining a functional model that relates the value of a target continuous variable $y$ with the values of variables $x_1, x_2, ..., x_n$ (the predictors). This model is obtained using samples of the unknown regression function. These samples describe different mappings between the predictor and the target variable. Regression estimators are optimal in the sense that they are unbiased, efficient, and consistent. An unbiased estimator has the
expected value of the estimator equal to the true value of the parameter. An efficient estimator has a smaller variance than any other estimator. An consistent one has the bias and variance of the estimator approach zero as the sample size approaches infinity (Ostrom 1990).

There are two related but independent methods for decomposing the information content in a set of variables into information about an inherent set of latent components. The first method is called the principal component analysis. The aim of this method is to decompose the variation in a multivariate data set into a set of components, such that the first component accounts for as much of the variation in the data as possible, the second component accounts for the second largest portion of the variation, and so on. In addition, each component in this method of analysis is orthogonal to the others; that is, each component is uncorrelated with the others: as a direction in space, each component is at right angles to the others.

In factor analysis, which is the second method for decomposing the information in a set of variables, the decomposition approach is different. One is not always interested in the orthogonality of the components (in this context, called factors); neither do the proportion of the variance accounted for by the factors decreases as each factor is extracted. Instead, meaningful factors in terms of the particular application at hand are looked for. The factors sought are the underlying, latent dimensions of the problem. The factors summarize the larger set of original variables.

Since every inductive learning algorithm uses some biases, it behaves well in some domains where its biases are appropriate while it performs poorly in other domains. One algorithm cannot be superior, in terms of generalization performance to another one among all the domains (Kotsiantis et al 2005, Kotsiantis & Pintelas 2005). Given that no one regression method is the best in all tasks, a variety of approaches have been
evolved to prevent poor performance due to a mismatch of capabilities. One approach to overcome this problem is to decide when a method may be appropriate for a given problem, by selecting the best regression model according to cross validation (BestCV) (Sharkey et al 2000). A second, more popular approach is to combine the capabilities of two or more regression methods (Hjort & Claeskens 2003). To date, relatively limited research have addressed ensembles for regression (Breiman 1996, Zemel & Pitassi 2001, Hjort & Claeskens 2003). The combination of regression models is attractive for several reasons. The most significant reason is that the combination of complementary regression models assures to increase the robustness and the overall performance of the regression system (Kotsiantis, SB & Pintelas, PE 2005). Regression models are successful in text summarization systems, and pose an open challenge to improve their prediction performances (Fattah & Ren 2008). In this thesis, the role of regression in predicting summary sentences is investigated, and its strength in estimating weights for features is exploited.

1.4 ROLE OF CLASSIFICATION IN SUMMARIZATION

The extraction of important information from a large pile of data and its correlations is often the advantage of using machine learning. New knowledge about tasks is constantly being discovered by humans, and the vocabulary changes. There is a constant stream of new events in the world, and the continuous redesign of Artificial Intelligent (AI) systems to conform to new knowledge is impractical, but machine learning methods might be able to track much of it (Aker et al 2010). For classification, assuming each hypothesis has a finite set of possible outputs, the hypothesis space is always finite, since there are only a finite number of ways to label any finite training set, while for regression, even relatively simple hypothesis spaces, such as linear functions constructed using weighted least squares, consist of an
uncountable infinite set of hypotheses. There is a substantial amount of research with machine learning algorithms, such as Bayes network, Radial basis function, Decision tree and pruning, Single conjunctive rule learner, and Nearest neighbor’s algorithm.

A classification system performs two tasks: learning and classification. Rule inducing methods put considerable effort into the first task. Classification is quite a very simple problem, as the systems needs only to test a new example against each rule (Martin 1995). The drawback is that the system cannot classify the new example, if it does not fit any of the rules. In this situation, a heuristic may be adopted, such as “If the new example does not satisfy any rules, choose the class with the highest *apriori* probability”. In instance-based learning, the effort required to learn from a new case is roughly equal to that required to classify it and the learning effort is very small compared to many other approaches. Additionally, instance-based learning has the advantage that a conclusion is drawn for every new example, without needing a heuristic to handle examples that do not fit the learned rule set (Mitchell, TM 1997). The disadvantage is that the effort required to classify a new example is much higher than that of rule-based systems. This effort can become fruitless, if the number of examples in memory grows too large, or the number of features is very large.

In supervised methods for summarization, the task of selecting important sentences is represented as a binary classification problem, partitioning all sentences in the input into summary and non-summary sentences. A corpus with human annotations of sentences that should be included in the summary is used to train a statistical classifier for the distinction, with each sentence represented as a list of potential indicators of importance. The likelihood of a sentence to belong to the summary class, or the confidence of the classifier that the sentence should be in the summary, is
the score of the sentence. The chosen classifier plays the role of a sentence scoring function, taking as an input the intermediate representation of the sentence and outputting the score of the sentence.

Machine learning approaches to summarization offer great freedom, because the number of indicators of importance is practically countless (Lin & Hovy 1997, Osborne 2002, Zhou & Hovy 2003, Leskovec et al 2005, Fuentes et al 2007, Hakkani & Tur 2007). It is hardly an exaggeration to say that every existing machine learning method has been applied for summarization. One important difference is whether the classifier assumes that the decision about inclusion in the summary is independently done for each sentence. This assumption is apparently not realistic, and methods that explicitly encode dependencies between sentences, such as Hidden Markov Models and Conditional Random Fields, outperform other learning methods (Conroy & O’Leary 2001, Galley 2006, Shen et al 2007). A problem inherent in the supervised learning paradigm, is the necessity for labeled data on which classifiers can be trained. Asking annotators to select summary-worthy sentences is a reasonable solution (Ulrich et al 2008), but it is time-consuming, and even more importantly, annotator agreement is low, and different people tend to choose different sentences, when asked to construct an extractive summary of a text (Rath et al 1961).

Partly motivated by this issue, and partly because of their interest in ultimately developing abstractive methods for summarization, many researchers have instead worked with abstracts written by people (often professional writers). Researchers concentrated their efforts on developing methods for the automatic alignment of the human-generated abstracts and the input (Marcu 1999, Jing 2002, Zhou & Hovy 2003, Barzilay & Elhadad 2003, Daume & Marcu 2004), in order to provide labeled data of summary and non-
summary sentences for machine learning. Some researchers have also proposed ways to leverage the information from manual evaluation of content selection in summarization, in which multiple sentences can be marked as expressing the same fact that should be in the summary (Copeck & Szpakowicz 2005, Fuentes et al 2007). Alternatively, one could compute similarity between sentences in human abstracts and those in the input, in order to find very similar sentences, not necessarily doing full alignment (Chali et al 2009). Another option for training a classifier is to employ a semi-supervised approach. In this paradigm, a small number of examples of summary and non-summary sentences are annotated by people. Then two classifiers are trained on that data, using different sets of features which are independent given the class (Xie et al 2010), or two different classification methods (Wong et al 2008). After that one of the classifiers is run on unannotated data, and its most confident predictions are added to the annotated examples to train the other classifier, repeating the process until some predefined halting condition is met.

Several modifications to standard machine learning approaches are appropriate for summarization. In effect, formulating summarization as a binary classification problem, which scores individual sentences, is not equivalent to finding the best summary, which consists of several sentences. This is exactly the issue of selecting a summary, that we shall discuss in the next section. In training a supervised model, the parameters maybe optimized to lead to a summary that has the best score against a human model (Aker et al 2010, Lin et al 2010) and the most highly scoring sentences are selected to form the summary, possibly after skipping some, because of their high similarity to already chosen sentences. In this thesis an attempt is made to propose a modified classification based summarization model, that improves the task of summarization and the prediction performance of the classifier.
1.5 ROLE OF REFINEMENT IN SUMMARIZATION

The biggest challenge for text summarization is to summarize content from a number of textual and semi-structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user. Most summarization approaches generate a summary sentence by sentence: they include the most informative sentence, and then if space constraints permit, the next most informative sentence is included in the summary and so on. Some process of checking for similarity between the chosen sentences is also usually employed, in order to avoid redundancy in terms of the inclusion of repetitive sentences.

The text summarization software should produce the effective summary in less time and with the least redundancy. A general approach to improving the quality of the produced summary involves post-processing extracts, for example, replacing pronouns with their antecedents, replacing relative temporal expressions with actual dates, eliminating redundancies, etc. A sentence may contain several pieces of relevant information, alongside some not so relevant facts, which could be considered as noise. Including such a sentence in the summary will help maximize the content relevance at the time of selection, but at the cost of limiting the amount of space in the summary remaining for other sentences. In such cases, it is often more desirable to include several shorter sentences, which are individually less informative than long ones, but which taken together do not express any unnecessary information. In order to refine a generated summary, global optimization algorithms can be used to solve the new formulation of the summarization task, in which the best overall summary is selected (Filatova & Hatzivassiloglou 2004). Given some constraints imposed on the summary, such as maximizing informativeness, minimizing repetition, and conforming to the required summary length, the task would be to select the best summary.
The semantic similarity between sentences and a few existing sentence similarity factors have, in some cases, been useful in refining summaries (Meng et al 2005). A summary refinement model is proposed in this thesis which refines a rough summary so that the final summary produced is concise, informative and free from redundancy.

1.6 MAIN CONTRIBUTIONS

The greatest challenge for text summarization systems is to summarize information collected from single or multiple sources with different formats such as the textual, web pages, etc., still satisfying the needs of users. The text summarization system should produce an effective summary in an acceptable amount of time and with minimum redundancy. In this thesis, different statistical methods for the task of extractive text summarization are exploited, including unsupervised and supervised approaches.

Feature selection is an important task that enables one to choose a set of features, which are apt for representing a summary sentence. Deciding proper weights for individual features for the sentences in a document is very important as the quality of final summary depends on it and hence deciding features and feature weights is a crucial concern. An attempt is made to use multivariate regression approach to text summarization so that accurate feature weights can be used in producing the extractive summary. Multi Variate (MV) approach is a Mathematical Regression technique. In regression analysis, each independent variable is associated with a regression coefficient describing the strength of that variable's relationship to the dependent variable. The performance of the Multivariate based summarization system is evaluated using a set of religious documents and the DUC 2002 dataset.
Further, the problem of text summarization is viewed as a task of classification and the role and the influence of many classification techniques in the task of summarization is investigated. The classification based summarization models were evaluated in terms of accuracy in classifying sentences as summary sentences. In this research, these machine learning based classifiers are analyzed and their performance in predicting summary sentences is evaluated. Our analysis led us to the observation that decision trees have the true potential to act as summarizer. Decision trees are called optimal if it correctly classifies the data set and has minimal number of nodes. The decision tree algorithms use local greedy search method by means of information gain as target function to split the data set. The decision trees generated by these methods are efficient with classification accuracy but they often experience the disadvantage of excessive complexity. Construction of optimal decision tree is identified as NP-Complete problem (Murthy 1998). This fact leads the use of stable algorithms such as regression that provide global search through space in many directions simultaneously.

Our next contribution is in using Multivariate Regression model to enhance the performance of the decision tree based summarizer. Regression is one of the most used methods to estimate the regression parameters and it’s purpose is to minimize the differences between the observed and predicted values (Guerra & Donaire 2000). There are three primary reasons why regression analysis is of maximum use:

1. To model some phenomena in order to better understand it and possibly use that understanding to affect policy or to make decisions about appropriate actions to take. Basic objective: to measure the extent that changes in one or more variables jointly affect changes in another.
2. To model some phenomena in order to predict values for that phenomenon at other places or other times. Basic objective: to build a prediction model that is consistent and accurate.

3. The fitting of a regression model allows to make inferences like hypothesis tests of the model reliability and confidence intervals to the prediction;

A variation of Classification based Summarization algorithm named CBS-ID3MV that utilizes a regression ensemble to preprocess the training data used for decision tree construction is proposed. The proposed algorithm trains a regression ensemble and the trained ensemble is used to produce a new training data set. It substitutes the preferred attributes of the original training tuples with its values augmented by the output from the trained ensemble. The new modified training set is used for training the CBS-ID3MV model. The refined training data improves classification accuracy of the decision tree classifier in the CBS-ID3MV model.

The CBS-ID3MV model is compared with other existing methods such as one using fuzzy approaches (Suanmali et al 2009 & Hannah et al 2011), the baseline summaries of DUC 2002 and the Microsoft word 2007, in terms of recall, precision and F-measure. The proposed hybrid classification based summarization model is able to produce summaries that are better in predicting summary sentences based on the feature weights learnt after training the model on the DUC2002 dataset.

Though the proposed ensemble model is effective in predicting summary sentences, the final extractive summary suffers from redundancy, since the model is based on feature weights learnt through multivariate regression. Having redundancy in summaries are not desirable since summaries should be short, concise and should contain only informative sentences, not repetitive in nature. In order to refine summaries, context based
indexing is done, thus removing sentences that have similar meaning, thereby eliminating redundancies.

1.7 OBJECTIVES OF THE STUDY

This thesis focuses on exploring the role of classification and regression in the task of text summarization along with an improvement that provides refinement of summaries. This section lists down the main objectives behind this research.

- Study the known approaches to extraction-based automatic summarization and understand the difficulties and challenges on obtaining a high quality summarizer.

- Investigate the suitability of multivariate regression models in the task of summarization by identifying the importance of features in sentences present in a text document.

- Analyze the role of classification based representations and algorithms in solving the task of extractive summarization and study the performance of machine learning classifiers in predicting summary sentences.

- Develop a classification based summarization model that exploits the benefits of better feature weight learning through the use of multivariate regression and the prediction capabilities of the classification approaches.

- Design, implement and evaluate an automatic summary refinement framework. Apply the summary refinement framework on regression, classification and fuzzy based summarization system and evaluate its performance in terms of precision, recall and F-score.
1.8 THESIS OUTLINE

In Chapter 2, necessary background information and related work in summarization research is outlined. Previous researches on supervised and unsupervised approaches to text are briefly introduced.

Chapter 3, provides the set of features that are extracted from input text documents and used in the rest of the dissertation. The feature selection and the algorithms that helped to select those features are discussed in detail. The methods and functions used in extracting these features are detailed. The corpus used for the research and the evaluation metrics used in the overall research are discussed.

In Chapter 4, a detailed discussion of how the task of summarization can be achieved using a popular mathematical technique namely the multivariate regression is provided. Multivariate regression is successfully applied on two sets of datasets namely the religious dataset and the Document Understanding Conference (DUC) 2002 dataset. The evaluation and results of the said model are explained in detail.

Chapter 5, explores a classification approach to summarization. In this chapter, the summarization task is used from a classification perspective. The performances of various machine learning classifiers are investigated and an improved classification based summarization model that makes use of the multivariate regression for learning weights for features extracted that are discussed in chapter 3 is proposed.

In Chapter 6, one of the issues of extractive summarization namely redundancy is discussed and a possible solution to redundancy elimination using binomial distribution are discussed in detail. Machine generated summaries are prone to contain redundant sentences and such summaries
require refinement. The proposed refinement is applied to rough machine generated summaries produced from Regression Based Summarization (RBS) System, Classification Based Summarization Using ID3 (CBS-ID3) system, Classification Based Summarization Using ID3 and Multivariate (CBS-ID3MV) system, and the Fuzzy Based Summarization (FBS) system as a case study. The results of the refinement show a reasonable increase in the precision of the refined summary.