Chapter 2: Literature Survey

This chapter carries out an in depth study of the existing techniques, methods and models that have been developed in the context of text summarization. The first section takes a look about history of summarization. The next section goes deeper into summarization techniques with different approaches for the same.

2.1 History of Summarization

The initial research on summarization tasks dates back to several decades and continues to be a steady subject of research. Systems that were developed in the early 50s exploited thematic features such as term frequency, term occurrence, a product of term frequency and inverse document frequency, location-based features, and presence of background terms like title, cue words, and phrases. They were termed as surface-level approaches. This was followed by an entity-level approach based on syntactic relations, similarity relationships, co-occurrence and co-reference that were developed during 60s. Later, during 70s, the entity level approaches also called discourse-based approach, which used the rhetorical structure of text and format of the document, was developed.

These earlier approaches to automated Text Summarization were developed based on the principles in which the significant portion of a text can be determined to summarize with one or many assumptions:

- In a text, significant sentences include words, which are used frequently (Luhn 1958).
- Significant sentences contain words, which are used in the section and title headings (Edmundson 1969).
- Significant sentences are positioned at the beginning or at the end of paragraphs (Baxendale 1958, Mitra et al 1997).
- Significant sentences are located in a text, which depends on the genre, and these locations can be determined automatically with the help of training techniques (Kupiec et al., 1995, Lin and Hovy 1997, Teufl and Moens 1997).

- Important sentences use words such as “greatest” & “significant” or indicator phrases such as “the main aim of this paper” and “the purpose of this article”, while unimportant sentences use words of stigma such as “hardly” and “impossible” (Rush et al 1971).
- Important sentences and concepts are the highest connected entities in elaborate semantic structures (Skorokhod 1971, Lin 1995, Barzilay and Ellahald 1997, Mani and Bloedorn 1997)
- Important and unimportant sentences are derivable from a discourse representation of the text (Jones 1993).

Numerous summarization systems that are robust in nature have opted for statistical sentence extraction. Various systems have been designed that can objectively extract important sentences from the text in which the importance of the sentence is inferred from low-level properties. Hence, the result of any extraction process which leads to the formation of a summary contains a collection of sentences that are selected precisely from the text. All the current works (late 90s) represent the revival of different types of approaches and are being explored very aggressively due to high commercial value and government interest. Recent work focuses exclusively on extracts rather than abstracts. As natural language generation work has started focusing on Text Summarization, the focus on extracts is likely to be dwelt upon in the coming years.

The emergence of new areas such as multi-document summarization (Tjhi and Chen 2007), multilingual summarization (Mihalcea and Paul 2005) and multimedia summarization (Murray et al., 2009) are also being seen as lucrative areas nowadays. They are not only similarity measures based on current sentence extraction (Aliguliyev 2009, Qiu and Pang 2008), but also based on the sentence clustering approach (Alguliev and Aliguliyev 2005). It is also a challenging task to identify sentences for generating summary with focus on reducing similarity among the sentences (Binwahlan et al 2009, Hendrickx et al 2009).

2.2 A Deeper look into Summarization

Research on automatic Text Summarization is the need of the present scenario with respect to Information Retrieval, Information Extraction and Internet

Surfing being the most popular applications. Many methods and approaches are available for information retrieval from various sources. As an application of information retrieval, many techniques have been developed till date on document summarization. The various existing methods are explained below.

As the research in Text Summarization started gaining momentum in the research community, extensive research work and competitions on this subject was started by DUC (Document Understanding Conference) and TAC (Text Analysis Conference) since 2011.

In the subsequent sections, the recent work that has been carried out has been described in context of abstractive and extractive Text Summarization. It also explains various earlier methods and techniques such as graph-based, cluster-based, term frequency, Latent Semantic Analysis, unsupervised , supervised etc. to name a few.

2.3 The Input Source

In automatic Text Summarization, the system takes one or more articles/documents as input to produce a summary. The generation of summary from one document is called single document Text Summarization. On the other hand, the summary produced from multiple documents is referred to as Multi-document Text Summarization. To retrieve text summary efficiently from multiple sources is definitely more difficult than generating a single source summary. This is because of the variations in data redundancy in the summary generated.

2.3.1 Single Document Text Summarization

In single document Text Summarization, the system takes one document to produce a readable and concise yet complete summary so that the meaning of the document can be retained. IBM was the first to introduce the single document Text Summarization focused on technical articles (Lunh 1958). The “Bag of words” model is also available that is generated from the sentences in the documents and has been proved important for summarization. For better results, few other operations of pre-processing procedures such as stemming, stop word removal etc. are performed before generating the summary. In addition, each sentence is assigned a score as per its importance and that score is used to generate summary.

2.3.2 Multi-Document Text Summarization

In Multi-Document Text Summarization system, the primary aim is to extract information from multiple sources, which have the same context. The goal of multi-document summarization is to provide a brief summary of different resources as a common subject.

SUMMONS (McKeown & Radev 1995) system proposed to measure the similarity between sentence pairs in the document that works in strict domains. The Maximum Marginal Relevance, measure in Multi-Documents (MMR-MD, Carbonell & Goldstein 1998) is a multi-document extractive summarization system that works on relevance factor and produces summary that has good quality and low redundancy. MEAD system performs very well for extracting Text Summarization for a single as well as multi-document Text Summarization (Radev et al 2000). MEAD works on the idea of centroid-based extractive summarization.

2.4 Summarization Approaches

As stated earlier, Text Summarization can be achieved with two different approaches: extractive and abstractive. Extractive summarization approach extracts sentences from the input documents whereas in abstractive summarization approach, sentences are reformulated or reconstructed from the original text. The following subsections take a look into various approaches regarding how the regression, classification and deep learning techniques play a vital role in Text Summarization.

2.4.1 Text Summarization with Supervised Approaches

Supervised algorithms are the machine learning based algorithms that are driven by some given training data. These algorithms learn by example, which are provided in the form of training data set and testing data set. The algorithm learns from training and experience and on that basis, it predicts the class where that data belong to. All the proposed supervised techniques of Text Summarization (Mihalcea & Tara 2005, Fung & Ngai 2006, McDonald & Chen

2006), divide the summarization into two classes and represent it as a classification problem. If a sentence is classified as a part of summary then such sentence is considered as positive class, negative otherwise.

Kupiec et. al. in 1995 presented a method that has been derived from Edmundson's method which can learn from data. With the help of a Naïve Bayes classifier, each sentence is functionally checked and decided if it is worth to take that sentence as a part of the extractive summary.

Aone et. al in 1998. Introduced a model named Dim Sum with the use of Naïve-Bayes classifier based on special features like term-frequency (TF) and inverse document frequency (IDF) to get signature words. The IDF is computed from a large corpus of the same domain for the given documents. In this model, a name entity tagger is used to find a single token for each entity. This model also implemented shallow discourse analysis to maintain cohesion in the text. For linking name as aliased within a document, references are determined at a low level. For example, the linking of US to "The United States" and IBM to "International Business Machines". Considering the lexical terms, synonyms and morphological variations are combined using Wordnet (Miller, G. A., & Fellbaum, C. (2007).

For the Text Summarization system, a significant contribution to get the position of the sentences against the keyword was given by Lin and Hovy (1997) using the concept of Optimal Position Policy (OOP). In this, they used the position of the topic on newswire corpus and Ziff Davis text. Lin (2011) proposed a method on the basis of the Decision trees to extract the sentences instead Naïve Bayes classifier. It was tested on TREC dataset. In this, they included few features such as the signature of IR, query, adjective and quotation.

In 2002, Osborne proposed a Log-linear model which shows better performance over the Naïve-Bayes classifier. To evaluate the summaries, they used an F-score, a harmonic mean of precision and recall. They also used different types of features like the length of the sentences, the position of the sentences etc.

2.4.2 Classification in Text Summarization

Text summarization and classification are core techniques to analyze a huge amount of text data in the big data environment. Moreover, as the need to read texts on smart phones, tablets and television as well as personal computers continues to grow, text summarization and classification techniques become more important and both of them do essential processes for text analysis in many applications.

Traditional text summarization and classification techniques have individually been considered as different research fields. However, we find out that they can help each other as text summarization makes use of category information from text classification and text classification does summary information from text summarization. Text classification is the process of assigning tags or categories to text according to its content. It is one of the fundamental tasks in Natural Language Processing (NLP) with broad applications like sentiment analysis, topic labeling, spam detection, intent detection and many more.

Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more. Text can be an extremely rich source of information, but extracting insights from it can be hard and time-consuming due to its unstructured nature. Businesses are turning to text classification for structuring text in a fast and cost-efficient way to enhance decision-making and automate processes. Researchers are working on discovering new techniques which propels major changes every day. Though, Artificial Intelligence (AI) has shown better results, machine learning methods perform better (Aker et al 2010). In classification, there are a number of outputs as hypothesis; where the hypothesis space is always finite. There are a number of methods to label the training data. There are various machine learning based classification techniques developed such as Naïve Bayes Classifier, Decision Tree, Support Vector Machine and K-Nearest Neighbor.

Classification performs major two tasks: learning and classification. Classification approach is one of the easiest and simplest tasks. A new input is checked against the rules in the system (Martin 1995). The major disadvantage of this type of system is that, if any example does not match to any rule then it cannot classify new input. In the heuristic approach, in such situation, if any new input does not match to any rule then a label with the highest probability is chosen.

In some approaches, the selection of sentences is carried out based on binary classification problem, called supervised approach. It divides the sentences into two parts: selected sentences and non-selected sentences to form the summary. The documents are trained on model along with statistical measures where each sentence is represented as potentially important for summary. These sentences can belong to the summary class in the form of confidence interval. The sentences can be part of the summary and each sentence is assigned a score on the basis of probability. The classifier plays a major role in scoring the sentences.

In Text Summarization, machine-learning techniques provide great support and freedom (Lin & Hovy 1997, Osborne 2002, Zhou & Hovy 2003, Leskovecet al 2005, Fuentes et al 2007, Hakkani & Tur 2007). It is difficult to say that every classification problem works for Text Summarization. The assumption is that the decision of classifier to include the sentences, in summary, is independent for each sentence. The assumption does not work for realistic contents; hence the sentences are encoded based on dependencies (Conroy &, Galley 2006, Shen et al 2007). In supervised learning, the system has to train the classifier to label data. The human annotators select sentences for the summary as a possible solution (Ulrich et al 2008). The problem in this approach is that it is time-consuming and different annotators choose different sentences. In this thesis, one of the models produces extractive summarization using Naïve Bayes Classifier which has been discussed in detail further in subsequent chapters.

2.4.3 Regression in Text Summarization

Data Mining as mentioned in the first chapter is the foundation of Text Mining wherein the data is text which is unstructured and a subject of it is Text Summarization. The most frequently used data mining techniques are: association, path analysis, regression, classification and prediction, clustering, visualization and so on.

Linear methods are well suited for working with sparse and dense data, for example, when working with texts. This can be explained by a high rate of training and a small number of parameters, making possible to avoid retraining. Linear regression is one of the basic and most simple machine learning methods. The method is used to restore the relationship between independent and dependent variables.

Regression analysis is a statistical technique for estimating the relationships among variables. In case of Text Summarization the relationship is amongst the individual words and sentences in a given document. The words or sentences in this case may be considered as the variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables Over here the independent variables are called the predictors or the endogenous variables and the dependent variables are called the criterion or the exogenous variables. In fact using regression we can find how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Work on Time Series Linear Regression Analysis has been done on textual data (Ostrom 1990).

Every inductive algorithm uses some biases. For, some algorithms perform very well with biases and predict excellent results whereas, some algorithms do not perform well. Therefore, any single algorithm or technique cannot be perfect in terms of performance for all domains (Kotsiantis et al 2005, Kotsiantis & Pintelas 2005). No single regression method has been proven best for all tasks. One must go for a trial and error approach to get the best results from it. Various approaches have been used to prevent over-fitting and worst performance. Cross-Validation (Sharkey et al 2000) is one of those approaches to overcome this problem. Another approach is to combine two or more regression models (Hjort & Claeskens 2003) for performance reason. The combination of multiple regression models is best, for, it increases the robustness and performance of the system (Kotsiantis, SB & Pintelas, PE 2005). For Text Summarization systems, the regression model has shown excellent performance however, the challenge is to improve the results (Fattah & Ren 2008).

2.4.4 Text Summarization with Unsupervised Approaches

In the unsupervised approach, the model doesn't generate the results during training but it prepares clusters based on the statistical characteristics. It is contrary to the supervised approach, where it generates the result during training and performs labeling at same time. The labeling is done to all data even they have a small subset of a particular representative of the desired classes (Liritano & Ruffolo 2001, Xi Quan Y 2008, Ming, Z Jianli, W & Guanjun, F, 2008). For Text Summarization, the unsupervised techniques have shown good result.

In the labeling of sentences that form the summary, fuzzy approach is also being used. Fuzzy-rough sets are used to extract a sentence from the document as part of summary by predicting the significance of the sentence. This approach removes the issue of the sentences that have similar meaning but written using synonyms and are treated differently. One of such model was constructed for Text Summarization and tested on the standard corpus. There are many approaches which use fuzzy logic based clusters and rough-rule based clusters for Text Summarization. Suanmali et. al. (2009) and Hannah et. al. (2011) used fuzzy approach for Text Summarization to generate summaries from the input document by using different features.

In automatic Text Summarization, the cluster-based approaches have been used to get an effective summary from the input document. If the documents contain different topics then clustering techniques become more vital to generate the summary.

2.4.5 Cluster-Based Methods

The cluster based approach is related to grouping or clustering multiple or single documents and to producing cluster wise summary based on feature profile oriented sentence extraction strategy. The related documents are generally grouped into same cluster using threshold-based document clustering algorithms. Feature profile is generated by considering word weight, sentence position, sentence length, sentence centrality, proper nouns and numerical data in the sentence. Judith D. Schlesinger (J M Conroy, J D Schlesinger, and J G Stewart 2005) introduced a method for multi-document summarization named CLASSY (Clustering, Linguistics and Statistics for summarization). CLASSY can be used for single and multi-document Text Summarization as well. In this, summaries are generated along with topic of the contents. In this, English corpus has been used for trimming and statistical approach has been used for scoring the sentences. With the help of trimming, it reduces the distance between the sentences i.e. how the two sentences differ in terms of meaning, and identification of the sentence on the basis of significance of that sentence to be included in summary. The summary is produced for individual document and a later stage; they are rearranged for the final summary and combined into one. The CLASSY has five steps starting from preparation of document using stemming and stop word removal, calculating the score of each sentence, removal of duplicate sentences, and finally sorting of the sentences based on scores.

Xiao-Chem Ma et. al. (X.-c. Ma, G.-B. Yu, and L. Ma, 2009) introduced the technique for Text Summarization with three stages named pre-processing, clustering and summary generation. Clustering is the core and important part of this technique. For clustering, first, it constructs VSM (Vector Space Model) and then, it prepares the matrix of relationship. In the third step, it sets the value of initial parameters to form the clusters. Finally, to generate the summary, it uses MMR (Maximal Marginal Relevance) technique by selecting the content of the multi-set of documents and by querying, it finally deliver the summary.

Virendra Gupta (V. K. Gupta and T. J. Siddiqui, 2012) proposed an approach of multi-document summarization with phrase clustering to generate summary of the document. The syntactic and semantic analysis is used to find the similarity between sentences for clustering. To generate summary, various features like index of sentence reference, similar features of location and concept. The summaries are composed into an individual cluster for each document. Later, best sentence is taken out from each cluster to build a multi-document summary.

2.4.6 Term Frequency Based Methods

G. Salton (2005) introduced method based on term frequency and inverse document frequency model (TF-IDF). Here, the term of a document is the proportion between the numbers of terms in the document to the occurrence of the number of documents that include those terms. The expression is evaluated by the formula TFI X IDFI, where TF is term frequency of ‘I’ in the document and IDF (inverted document frequency) in that term occurs. The sentences are assigned the score and as per importance of the score, the summaries generated.

Jun’ichi Fukumoto (2004) proposed a method of extraction by using TF-IDF techniques to generate extract from multiple documents. In this model, the summary of the individual documents are prepared. Then, all summaries are merged as a part of multi-documents summarization. The system divides the document into three categories called high-frequency nouns, object names and others. First, the sentence is extracted from each document on the basis of score of TF-IDF. Next, the sentences those are not important are removed. All extracted sentences are sorted as per the actual order as given in the document to generate summary efficiently for each document. All these sentences are grouped into one cluster and repetitive sentences are removed. Finally, all sentences are sorted to generate summary.

2.4.7 Latent Semantic Analysis Methods

Shuchu Xiong et. al. (2014) introduced a method using Latent Semantic Analysis. Here, in this method, the sentences are evaluated based on its similarity of prediction on the vector of latent singular. The steps that are executed in LSA are as follow: The first step is to create an input matrix which gives the occurrence of the terms in document. In second step, it performs singular value decomposition (SVD) on the input matrix. Finally, the sentence selection algorithm selects the sentences to produce the summaries. Here, the authors have used the Maximal Marginal Relevance (MMR) and a centroid based algorithm, Mead.

Josef Steinberger et. al. (2014) has experimentally demonstrated the issues related to LSA. To overcome these issues, they proposed a method with a variation in the existing method based on Singular Value Decomposition (SVD). In this new technique, they carried out recalculation of Singular Value Decomposition (SVD) of the sentence matrix. For evaluating the summaries, they used similarity and term significance.

2.4.8 Graph-Based Methods

There are a number of different models which are based on graph theory and one of the most popular being the TextRank model (R. Mihalcea and P. Tarau). Graph-based ranking algorithms like Kleinberg’s HITS algorithm (Kleinberg, 1999) or Google’s PageRank (Brin and Page, 1998) have been successfully used in citation analysis, social networks, and the analysis of the link-structure of the World Wide Web. Arguably, these algorithms can be singled out as key elements of the paradigm-shift triggered in the field of Web search technology, by providing a Web page ranking mechanism that relies on the collective knowledge of Web architects rather than individual content analysis of Web pages. In short, a graph-based ranking algorithm is a way of deciding on the importance of a vertex within a graph, by taking into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information. Applying a similar line of thinking to lexical or semantic graphs extracted from natural language documents, results in a graph-based ranking model that can be applied to a variety of natural language processing applications, where knowledge drawn from an entire text is used in making local ranking/selection decisions. Such text-oriented ranking methods can be applied to tasks ranging from automated extraction of keyphrases, to extractive summarization and word sense disambiguation (Mihalcea et al., 2004).

Julin Zhang (2005) introduced a method by using graph theory. That method is called hub / authority framework. In this technique, surface features and content features are merged with each other, such as length and location of sentence etc. The model extracts important features of sub-topic underneath the framework of hub / authority. At last, the sentences are assigned the scores and on that basis, the final summary is generated.

S. Hariharan and R. Srinivasan (2009) introduced two methods with differences such as with or without omitting the nominated sentences. In their research, they generated summaries on the news articles using graph-based method. With the help of adjacency matrix, the representation was done with similarity measures between the sentences of documents. This was the first step in their graph based approach. From the two techniques that they developed, the first one proposes cumulative sum and the second one concentrates on the degree of centrality. By using these two values, an adjacency matrix is generated. In this, precision and recall have been used for calculating extractive summaries in the form of matrices. It is evaluated based on two metrics: Effectiveness-1 and Effectiveness-2. With the help of discounting method while testing the generated summaries for single and multi-documents, it was concluded that the second approach is better than the first one, however, there was scope of improvement in it.

Khushboo and her colleagues (K. S. Thakkar, R. V. Dharaskar, and M. Chandak, 2010), introduced the methodology of Text Rank using few variances. This method uses the shortest path algorithm for generating summaries. The sentences are selected from the path where each sentence may be similar to previous sentences for generating summaries and to be selected as top ranking sentences based on its Text rank. As a first step, it builds a graph model in which the Text units such as words, phrases, collocation, sentence or others are considered as vertices in the graph. Later, for each vertex, the score is calculated with help of any graph based ranking algorithm such as HITS, Page Rank etc. Finally, the shortest path algorithm is applied to generate summaries.

Shuzhi Sam proposed a hybrid approach using weighted graph model for Text Summarization which includes two concepts: the sentences clustering and ranking. It means the method depends on the cluster as well as Graph-based approaches for generating summaries from the given text. Following steps are executed in this approach:

1 As stated above, there are two techniques followed: first is Graph model for sentence ranking and the second is clustering to merge the similar sentences

2 Clustering of sentences can be completed on the basis of Singular nonmatrix factorization. In this, Latent Semantic Analysis can be applied.

3 The weighted graph model reflects the association between sentences in order to formulate clusters and rank the sentences in a document.

Tu-Anh Nguyen-Hoang (T. Nguyen-Hoang, K. Nguyen, and Q.-V. Tran, 2012) proposed method which consisted of three steps. In the first step, for the data set, the specific structure is added into every document. The undirected weighted graph is used as a structure. The title and sentences plays a major role for the construction of the graph. In the second step, a weighted page rank is calculated using graph based ranking algorithm for each sentence of the document. The scores are given based on the relevant features of the sentences in the given document. Few sentences are extracted from the documents to generate summaries based on the rank of sentences. In the third and final stage, all different summaries are merged into a single summary using MMR (Maximal Marginal Relevance) algorithm.

2.4.9 Text Summarization with Natural Language Processing

To generate an effective summary, various Natural Language Processing (NLP) techniques such as a stop word removal, stemming, analysis of language etc. are also being used.

Ono et. al. (1994), a Japanese researcher, developed a computational model that extracts sentences using linguistic techniques. For evaluating this model, 30 articles of newspaper were used as a dataset. Boguraev & Kennedy (1996) presented an idea based on local salience, which uses a mixture of syntactical, grammatical and contextual parameters. Barzilay & Elhadad (1997) developed an alternative method of linguistic analysis for Text Summarization. In this, the semantically related documents are identified and many lexical chains are extracted for the presentation of the original document.

Marcu (1998), presented method where using traditional feature along with heuristic discourse. Here, in the original document to generate a valid discourse presentation, used rhetorical parsed to generate a tree of discourse. Carbonell & Goldstein (1998) presented a method of information novelty in the association of query relevance, which primary aim to reduce data redundancy and re-ranking retrieve sentences from Text Summarization by using Maximal Marginal Relevance (MMR).

Lam et. al. (2001) developed and tested a system called Financial Information Digest (FID), which retrieves financial news online. The system uses content-based classification to retrieve documents simultaneously after the understanding from different domains. It also integrates the information from all the articles. The scattered information is integrated as one article and allows the users to access complete information using cross-validation. This system performs excellent and obtains an accuracy of 91%.

To generate an informative summary, semantic-based approaches are bieng used. Word dictionary can also be used to get information about the semantics for words which are in the document. Mengwang et. al. (2005) presented a method in which they used HowNet lexical to extract most relevant sentences to generate a summary. The dictionary produces words to retrieve and recognize the conceptual vector space. Initially, rough summaries are produced and gradually the redundancies are removed from summary to obtain better summary.

To identify the relationships between entities, term co-occurrence graph makes the task easier. This has been proved by Gean et. al. (2008) to present the idea of summarization for information on various subjects based on the graph of term co-occurrence and linkage of different subjects.

To present the text contiguous structure, lexical chain structure can be used. To calculate lexical chains, grouping of words that are semantically related, Silber & McCoy (2002) presented a lexical chain generation method. In Information Retrieval and correction of English grammar correction, these lexical chains are used.

Gupta & Lehal (2010) investigated that the summaries that are being generated without the use of Natural Language Processing (NLP) have less consistency and more redundancy.

2.4.10 Deep Learning in Text Summarization

Deep learning has shown tremendous performance in Text Summarization. There are a number of techniques introduced to achieve efficient Text Summarization using deep learning. Artificial Neural Network (ANN) with hidden layers, Recurrent Neural Network (RNN) are examples of supervised methods of deep learning whereas Self-organizing maps, Auto Encode, Boltzmann Encoding are considered as unsupervised methods of deep learning. For automatic Text Summarization, shallow network has also been introduced by

Kaikhah (2004). Netsum is a neural network-based system, which is introduced by Svore, Vanderwende, and Burgs (2007). To improve the result of Text Summarization, a mixture of MLP (Multi-Layer Perceptron) in association with fuzzy logic has been proposed by Shardan and Kulkarni (2010). In similar attempt, to improve the result of summaries, RNN and feed-forward neural network have been applied. Deep Learning methods can be applied to extractive as well as abstractive Text Summarization. Companies like IBM and Facebook have established successful models of abstractive summarization built on RNN and convolutional neural network (CNN) respectively (Nallapati, Zhou, Nogueira dossantos, Gulcehre, & Xiang, 2016; Rush, Chopra, & Weston, 2015). In this thesis, a hybrid approach has been developed based on deep learning to achieve efficient Text Summarization. For that, Self-organizing Maps (SOM) has been used as unsupervised approach along with Artificial Neural Network in a combination of hidden layers and gradient descent as a supervised approach.

Nowadays, researchers have been trying to provide efficient Text Summarization using Artificial Neural Network (ANN) and it has been proven effective in many applications. Kaikhah (2004) introduced an ANN based technique that learns the characteristics of the sentence. Then these sentences are included in the summary as according to their characteristics. They evaluated the model for 50 news articles of different type of subject such as politics, sports, technology etc. and achieved 96% accuracy.

In another similar, ANN based attempt, the relevance score is assigned to each extracted sentence from the input document. With this score, it becomes easier to evaluate the importance of a sentence for the summary (Alguliev 2005). The F1 score was used to evaluate this model. Yong et. al. (2005) presented an idea for Text Summarization to produce summary by combing two different types of approach, statistical and neural network. The system develops learning ability by merging statistical approach along with neural network. The model learns from the classified sentences with well-trained and sufficient text samples. The English subject experts evaluated the summary in the context of English language with an accuracy of 83.03%. The system produces an effective summary from the input document with proper readability and maintained important concepts of document.

Svore et. al. (2007) gave the concept of Text Summarization using ANN with third-party datasets to produce an efficient summary of the original document. The model used ANN based RankNet algorithm with gradient descent techniques to train inputs. It uses the CNN dataset with 1365 documents. The summaries were evaluated using ROUGE-1 and ROUGE-2, developed by Lin in 2004. The statistical (Mihalcea & Tarau 2005, Fung & Ngai 2006 and McDonald & Chen 2006) techniques enhance the system performance by selecting the important sentences. This trainable summarizer considers many features like position of the sentence, positive and negative keyword, the centrality of the sentence, the inclusion of name entity, numerical data in a sentence, relative length for each sentence against generating summaries etc.

The effect of every sentence feature on the summarization task is examined to get suitable feature weights for a proper combination of Genetic Algorithm (GA) and Mathematical Regression (MR) to train the summarizer. The Feed Forward Neural Network (FFNN), Probabilistic Neural Network (PNN), and Gaussian Mixture Model (GNM) have been used to the train the model for the text summarizer. To test this model, the authors used 100 Arabic political articles and 100 English religious articles at different compression rate with different parameters. The result of the summarizer was remarkable. In another attempt of automatic Text Summarization, Chen et. al. (2008) proposed a model called “AutoTextSumm” to extract sentences for oil and gas drilling with help of statistical method.

SUMMARY

This chapter gave a deeper look about the work that has been carried out in the area of Text Summarization starting from traditional techniques to ANN and NLP based approaches. The next chapter discusses about the preprocessing phase and various methods that have been used for the same purpose.