6. APPROACHES TO TEXT SUMMARIZATION: A SURVEY

With the wide spread use of internet and the emergence of information exploration era, quality text summarization is essential to effectively condense the information. Text summarization is the process of producing shorter presentation of original content which covers non-redundant and salient information extracted from a single or multiple documents. Approaches to automatic text summarization involves

(i) *Elimination of redundancy*: The sentences in the text which convey the same meaning are said to be redundant and can be eliminated in the summary.

(ii) *Identification of significant sentences*: Summary being a shorter representation of text requires to include only salient sentences from the original document.

(iii) *Generation of coherent summaries*: Sentences selected for summarization needs to be ordered and grouped so that coherence and readability is maintained.

(iv) *Metrics for evaluating the automatically generated summaries*: In most of the cases the quality of the summary is judged by humans and hence automatic evaluation is a desirable feature.
To solve these issues several methods have evolved ranging from statistical to semantic methods (Carbonell J and Goldstein J 1998; Braziley et al 1999; Dejong G 1982; Radev D et al 1998). The approaches to text summarization can be broadly classified as (i) Abstractive and (ii) Extractive. These methods further involve either (i) Elimination of redundant information or (ii) identification of salient information.

### 6.1 Abstractive Summarization

Abstractive summarization methods (Dejong G 1982; Radev D.R et.al 1998; Radev D.R et.al. 2000; Braziley et al 1999) extract phrases and lexical chains from the documents by using language understanding tools to generate a summary (or abstraction) (Reiter et al 1997). Some of the major steps involved in abstractive summarization technique are **extraction of basic features, identifying the relevant information, and Refining and reducing information.**

The process of abstraction is complex because it cannot be formulated mathematically or logically (Jing H et.al 99). The quality of abstractive summaries depends on the deep linguistic skills. Abstractive summarization techniques are classified into two categories.

**Structured based approaches,** encode most important information from the document(s) through cognitive schemas such as frames, scripts, and templates (Dejong 1982; Radev D R et.al. 1998][Radev D R 2000; Radev D.R et.al 2004]. Script for example is a structured template
with slots used to identify common important events over the domain. Each domain has its own script.

Semantic based method, semantic representation of document(s) is used to feed into natural language generation (NLG) system. This method concentrates on identifying noun phrases and verb phrases by processing linguistic data. Phrases thus obtained are then linked to concepts, attributes and relations of a domain-specific ontology (Brazilay R et al 1999; Saggion H et al 2002). The important document regions (sentences, paragraphs) are selected by using ontology-based annotations and clustering techniques. Resultant information is used to convert regions into semantic representation. This representation is then fed to an NLG system which produces abstracts.

The techniques involved in abstractive summarization include word by word parsing, rhetorical parsing, statistical parsing and a mixture of all. The limitations involved in abstractive summarization techniques are

- Sentence synthesis is not a well-developed field yet, and hence machine generated automatic summaries would result incoherence even within a sentence. In case of extractive summaries, incoherence occurs only at the border of two sentences.
- Abstractive systems are difficult to replicate, as they heavily rely on the adaptation of internal tools to perform information
extraction and language generation. In fact, extractive systems are generally easy to implement from scratch, and this makes them appealing when sophisticated NLP tools are not available.

- Abstraction is not easy to do since it requires semantic understanding of text.

### 6.2 Extractive Summarization

Extractive summarizers (Luhn 1958; Endumen 1969; Dalniais et al 1988; Hsin hai C 2000; I mani et al 1999; Marcu 1997) aim to pick out most relevant sentences in the document and maintain low redundancy of information in summary. Though anti-redundancy was not explicitly documented in older systems, most of the current systems account for it in their own novel ways. The key ideas in this approach require performing a study on human generated abstracts, and determine characteristics expected in automatically generated abstracts. Mathematical and logical formulations are designed to score and pick sentences from the documents that match with manually generated abstracts. Iteratively sentence-scoring is applied to match manually generated abstracts with automatic abstracts.

In extractive-based methods, most important document regions (phrases, sentences, paragraphs are ranked high. The highest ranked regions from all documents can then be combined and re-ranked using
similarity measures which in turn minimize redundancy. Finally rules are applied to produce summary. There are different approaches to implement extractive summaries.

**Term frequency-Inverse document frequency** (TF-IDF) (Luhn 1958) is a word distribution method in which two measures namely term frequency (tf) and document frequency (df) are calculated for each non-stop-word (w) in the document. Term frequency (tf), indicate number of times a word appears in the text which measures salience of word within that document. Document frequency (df) indicate number of documents in which the word appears. The frequent occurrence of a word in a document is treated as informative word, which is calculated by document frequency measure. Thematic words are obtained by comparing the ratio between two frequencies, referred as (if*idf) measure.

Once (tf-idf) score has been computed for each word the next step is to calculate number of such thematic words per sentence. With this value sentences in the input text are ranked and highest scored sentences are picked to be part of summary. Relying on TF-IDF score for summary generation would “insist” that each sentence should consist of thematic words, to mark them as informative sentence. It is unlikely that every sentence in the document would contain theme words, which would cause some of the most important sentences to be
excluded from a summary. Redundancy of information is extremely high in this method.

**Classical method** (Edmundson, H.P., 1969) in these approaches sentences are scored according to four factors namely, the presence of (i) high-frequency content words (key words), (ii) pragmatic words (cue words) (iii) title and heading words and (iv) sentence location. Score of a sentence is calculated as summarization of scores obtained for each of these features. The issue of redundancy is not accounted in this approach.

**Cluster based methods** measures relevance or similarity between each sentence in a document with that of sentences selected for summary. Summaries address onto different “themes” appearing in the documents, which is incorporated through clustering. Clustering based methods become essential to generate a meaningful summary.

Maximum marginal relevance multi-document (MMR-MD) summarization is a cluster based extractive summarization method. This method aims at generating summary of high relevance, to the given query or document topic. Passage clustering forms a main component in this system, to extract most relevant sentences of a documents by keeping the summary non-redundant (Gold Stein et al 1999; Goldstein et al 2000; Carbonell J et al 1998).

Similarity measure is used to initially cluster given document or documents. Sentence closest to centroid of a cluster may be chosen to
include it in the summary. Sentences are chosen for inclusion in summary are such that they are maximally similar to the document or query, while maintaining maximal dissimilarity to the sentences already included in the summary. This ensures that most representative sentences of the document are chosen, while ensuring minimum redundancy in the summary.

*MEAD is a centroid based extractive summarization method* (Carbonell J et al 1998; Radev et al 2004) calculates centroid of words which play a crucial role in documents classification and identifying salient sentences in a cluster. Scores are calculated for each sentence, which plays a significant role in selecting sentences into a summary. Score calculation for each sentence is dependent on three factors namely centriod value ($C_i$), positional value ($P_i$), first-sentence overlap ($F_i$). The overall score ($S_i$) of a sentence $i$ is a weighted sum of these three factors. The scores are calculated for each sentence by taking a cluster of $d$ documents with $n$ sentences and compression rate ($r$) as input. The summary generated by this method extract ($n^*r$) sentences from the cluster, whose scores are high.

*Graph theoretic* (Xiaojun Wan, Jianwu Yang 2006) representation is an extractive summarization model, which provides a method to identify themes in the document. Preprocessing steps, namely, stop word removal and stemming are done before, to obtain graphical view of the documents. Sentences in the documents form nodes of an undirected
An edge is drawn among the nodes if sentences share some common words, or whose (cosine, or such) similarity is above some threshold. Graphic representation yields partitions indicating distinct topics covered in the documents. Identification of nodes with high cardinality forms higher preferred sentences to be included in the summary.

**Rhetorical structure theory (RST)** (Mann et al. 1988) exploits relations with the help of tree representation. Sentences form nodes of the tree, which are connected using RST relations such as elaboration, anti-thesis, etc. Salient sentences are retrieved by traversing the tree, in order to build a partial ordering of sentences in terms of their importance. According to the target compression rate, top n sentences can be extracted and presented as a summary.

**Machine learning techniques** are viewed as a classification problem in which the sentences are classified as summary and non-summary sentences. Features that are used to distinguish summary sentences are obtained by employing statistical techniques (Kupiec et al. 1995; Edmundson, 1969; Conroy and O'leary 2001; Osborne M. 2002).

### 6.3 Summary

This chapter determines state of art summarization technique discussing about different approaches evolved to generate a summary. Broadly the approaches are classified as extractive and abstractive.
Abstractive methods are difficult to implement because it requires a lot of semantic interpretation and sentence synthesis. Formulation of mathematical and logical models is complex for abstractive methods. Comparatively, extractive summarization techniques are of fewer complexities. A survey of these approaches has been examined to develop a new model for summarization.