Chapter 3

Single and Multi-document Summarization

Overview

Single document summarizers have proved their role and significance in creating short outline of single documents. Document Understanding Conferences (DUC, 2001-2007) working under NIST (National Institute of Standards and Technology) conducted Workshops on Text Summarization for single as well as multi-document summaries for different tracks. In 2008, DUC became a summarization track in the Text Analysis Conferences (TAC). The Text REtrieval Conferences (TREC) started since 1992; co-sponsored by NIST and DARPA (Defense Advanced Research Projects Agency) also conducted on-going series of workshops to evaluate new technologies in text retrieval including summarization tracks. Recently multi-document summarization has become a hot topic in research community. Thus the efforts presented at these worldwide recognized summarization workshops shows the growing needs and increasing research interest in this field. This chapter provides the details for single and multi-document summarization briefing different summary types, how it would be carried out? What are the various features needed to summarize multiple documents as in our case research papers? What are the issues and challenges summarizing research papers? What are the data mining strategies to overcome this?

3.1 Summary Generation

Summary requirements may be anything consisting applications like news articles, books, web

documents, meetings, presentation, notes, e-mails, speeches, lectures, thesis, scientific articles, headlines, outlines, minutes, biographies, abridgments, sound bites, movies, chronologies, etc [41].

But according to user needs summaries are generated in two ways; single document summaries and multi-document summaries depending on the factor dimensions.

### 3.1.1 Single Document Summarization

Text summarization is a data reduction process. As summary is concise, accurate and explicit, it becomes more and more important. One document can be composed of some subdocuments. The described contents in each of the documents laid special emphasis on different aspects although these documents were all surround the same topic. Summarization process identifies highly salient units (usually words, phrases, sentences, or paragraphs) within a cluster of documents. When a cluster consists of one document, the process is called single-document summarization; otherwise the name is multi-document summarization.

Single document summary systems will generate a summary based on a single source document. Single document can be composed of some subdocuments with multiple paragraphs. The described content in each of these subdocuments emphasis on different aspects all surrounding the same topic. Generally single document is composed of different side information and the different side information is related to the local topic only. In terms of the logical structure, a single-document set can also be considered as a set of local topics. The multiple semantic paragraphs comprise of similarity sentences, and each semantic paragraph is also a local topic itself.

Existing methods for single document summarization usually make use of only the information contained in the specified document. For evaluating single document summaries, DUCs issued the task of creating N-words summary of a single news article. The best performing systems [13] used various learning and semantic-based methods, although no system could outperform the baseline with statistical significance. After 2002, the single-document summarization task was
dropped by DUC saying single document extracts may be naturally more difficult than multi-document summarization [39].

### 3.1.2 Multi-Document Summarization

Multi-document summary is the process of dealing with a large amount of information present in multiple related source documents by comprises only the essential material or main ideas in a document in less space. Recently, multi-document summarization has become a major topic in automatic summarization [22]. Multi-document summarization by sentence extraction is almost the same as single-document summarization, except that we need to consider some extra features as degree of redundancy, temporal dimension, compression ratio and co-reference problem for the former [64]. It can be viewed as either as an extension of single document summarization of a collection of documents covering the same topic, or information extracted from several sources. Handling these related features from single document to multi-document is the main problem for researchers.

Multi-document summarization is an automatic procedure for extracting the information from multiple texts written about the same topic. Resulting summary report allows individual users to quickly familiarize themselves with information contained in a large cluster of documents. In such a way, multi-document summarization systems are coping with the problem of information overload. An ideal multi-document summarization system not only shorten the source texts but also presents information organized around the key aspects to represent wide range of views on the topic.

Thus the multi-document summarization task,

2. A short summary for each individual document is generated;
3. Multiple single document summaries are aggregated to form a complete summary of multiple related documents.
The multi-document summary quality should possess following criteria:

- Clear structure, including an outline of the main content, from which it is easy to navigate to the full text sections.
- Text within sections is divided into meaningful paragraphs.
- Continuous evolution from more general to more specific aspects
- Good readability etc.

3.2 Summary Categorization

Summarization tasks differ one from another and lead to various summary types depending on what the summarization process focuses on stating the factor context, for example **generic summaries** or **query related summaries** (sometimes called **query-biased summaries, user-focused**, i.e. tailored to the requirements of a particular user or user-group).

Basically following summary types are marked out [21]:

- Detail: indicative/informative
- Granularity: specific events/overview
- Technique: extraction/abstraction
- Content: generalized/query-based
- Approach: domain/genre specific/independent

Broadly, one distinguishes two approaches of summarization as Extract or Abstract:

**Extractive or Abstractive Summaries**- **Extraction** is the process of identifying important material in the text while **abstraction** is the process of reformulating it in novel terms. Extractive summaries are created by reusing portions (words, sentences, etc.) of the input text exactly. Extraction often uses standard statistically based information retrieval techniques using natural language processing and heuristics. Passages, often sentences or phrases, are extracted and pasted together to form a non-redundant summary which is shorter than the original document with as little information loss as possible. Sometimes the extracted fragments are post-edited, for

---

example by deleting subordinate clauses or joining incomplete clauses to form complete clauses [30, 66]. Text abstraction is the more challenging task. It parses the original text in a linguistic way, interpret the text into a formal representation, find brief concepts to describe the text and then generate a new shorter text, an abstract, with the same information content. Abstractive summaries are created by regenerating the extracted content. Within and across these two categories, summaries differ according to function and target reader as indicative, informative, or critical.

Various algorithms and strategies are used for summary generation. The generated summaries can also be divided into different types depending on their purpose.

a. *Indicative and Informative Summaries* - Indicative summary provide enough content to users for related sources, which users can read in more depth, providing details of the text without conveying specific content. While informative summaries act as substitutes for the source, collects related information in a brief. Critical summaries (or reviews), besides containing an informative idea, incorporate opinion statements on content.

b. *Generic and User Focused Summaries* - Generic summaries are content based addressing a broad community; there is no focus on special needs because the summarizer is not targeting any particular group. It is for general use reflecting the author’s point of view. While topic oriented or user focused summaries are approach based, concentrates on specific needs or interest of an individual or a specific group. Until recently, generic summaries were more popular, but with the full-text searching and personalized information filtering, user-focused summaries are gaining importance.

This hypothesis generates extraction based, domain specific topic oriented user focused summary. It falls under the indicative summary type as it extracts the most related information according to scholar query.

### 3.3 Summary Generation Issues

There are some issues producing multi-document domain dependent topic oriented summaries [64]. These issues are related to grouping or clustering of similar documents, identifying most
related and similar sentences, finding RRN terms, reducing redundancy, document ranking, sentence ranking, sentence categorization, summary accuracy and user interface as described in details below:

i. **Clustering**: The ability to group or cluster similar documents and sentences.

ii. **Coverage**: The ability to find and extract the ‘Related Novel’ research oriented sentences across documents.

iii. **Anti-redundancy**: The ability to minimize redundancy between sentences in the summary.

iv. **Summary cohesion criteria**: The ability to combine ‘Related Novel’ research oriented sentences in a useful manner for the scholar. This may include:
   – **Document ordering**: Arranging all ‘Related Novel’ research oriented sentences of highest ranking documents, then all sentences from the next highest ranking documents etc.
   – **Rank ordering**: Presenting most Related Novel and Dissimilar information first so that the reader gets the maximum information content
   – **Topic-cohesion**: Categorizing sentences together by topic clustering using sentence similarity and sentence ranking.
   – **Occurrence ordering**: Documents and sentences ordering based on the occurrence at the time of IR/search engine retrieval.

v. **Coherence**: Summaries generated should be readable and most related to guide the scholars.

vi. **Effective user interface**:
   – Easy access to the source of a given sentence or select or eliminate various sources.

### 3.4 Use of Data Mining Strategies

This hypothesis addresses all the above summary generation issues by the use of various data mining techniques such as **preprocessing, term identification, clustering, sentence extraction and sentence scoring**.
The sentence extraction is the data mining process for automatically removing most related sentences from multiple documents depending on query relevance. This process helps in identifying important material from text which has a significant and complete relative meaning. This research replaces conventional iterative extraction methods, which extract one sentence at a time without considering the rest of the sentences in the summary, with more holistic ones, where the decision to extract a sentence is determined not only by the content of a sentence, but also by the rest of the sentences extracted.

The clustering is the data mining process for automatically grouping most related multiple documents into clusters or bunches. If a collection is well clustered, the clusters containing related documents only be searched thus searching a smaller collection improves effectiveness and efficiency. Clustering can work to give users an overview of the contents of a document collection and can also reduce the search space. Thus the use of data mining strategies such as extraction and clustering for summarizing multiple research papers achieves remarkable progress avoiding system redundancy.

Optimization is the process by which one object, property, concept, theory, etc., is shown to be entirely dispensable in favor of another. Thus this system enhances effectiveness of automatic summary, to make it function at its best or most effective, and give its best advantage through optimization. The optimized summary presents user innovative contribution through most useful sentences from papers. It significantly improves the conciseness of automatic summaries by cutting unwanted, repeated, non-related material and presenting only the abstract. This process saves scholar time by avoiding his attention from reading and understanding unnecessary material for which he is not interested. This process also shortens system output and minimizes space requirements.