Chapter 9
Multilingual Summarization Approaches

Kamal Sarkar
Jadavpur University, India

ABSTRACT

As the amount of on-line information in the languages other than English (such as Chinese, Japanese, German, French, Hindi, etc.) increases, systems that can automatically summarize multilingual documents are becoming increasingly desirable for managing information overload problem on the Web. This chapter presents an overview of automatic text summarization with special emphasis on multilingual text summarization. The various state-of-the-art multilingual summarization approaches have been grouped based on their characteristics and presented in this chapter.

INTRODUCTION

Summarization is a kind of human ability. Since the time immemorial, some form of summarization has been used to store knowledge, transmit knowledge and memorize the key facts about anything.

Today, in the age of electronic media, it is hard to imagine everyday life without some form of summarization. News headlines are summaries, written in a terse stylized language, of material in a news articles. Abstracts of the scientific articles are a traditional form of author-written summaries. Other form of summaries include reviews of books or movies, Minutes of a meeting, a stock market bulletin, an abridgement of a book, a resume, an obituary, initial search hits returned by the search engines and so on.

The human experts for an area can write good summaries for the articles related to that area. This indicates that the human summarizers require background knowledge while writing the summaries. The summaries are very much subjective in nature. The human summarizers may differ in their viewpoints for summarizing the same article. So, the human produced summaries suffer from the bias of the authors. But, there is no question of bias when summaries are produced by the machines. Ideally, the machine-generated summaries should contain the best and important information.

Due to the lack of proper mathematical model of human cognition, it is difficult to explain how human learns to summarize. Since the human cognition is not computable or the present paradigms of computation are inadequate for computing the

DOI: 10.4018/978-1-4666-6042-7.ch009

human cognition, it is hardly possible for the machines to generate summaries as the human does. However, given a compression ratio, a machine can distinguish between the more relevant textual units and the less relevant textual units and select or reformulate the more relevant textual units for the generation of the summaries. Thus, the machine can produce sub-optimal results without depending on the exact mathematical formulation of the cognition process that humans apply to summary generation.

The World Wide Web (WWW) has introduced us with a new paradigm of knowledge sharing, transmission and consumption. The number of web pages available on the Internet almost doubles every year. Though English is dominating language on the web, the number of documents written in the languages other than English (such as Chinese, Japanese, German, French, Hindi etc.) is also increasing daily at unprecedented rate. With this rapid growth of the World Wide Web, information overload is becoming a problem for an increasingly large number of people. The traditional search engines such as Yahoo, Google etc. have revolutionized searching capabilities along with the intelligent visualization of search results to facilitate browsing on the Web. A keyword-based search using the engines results in links to web pages along with snippet of its contents where these keywords have occurred. This snippet actually helps the reader decide whether the retrieved link is worth reading in detail or not. So, in some sense these snippets serve as indicative summaries. It is also very difficult for the users to go through all the hits that the traditional search engines return and find the relevant information from the collection. This has created a growing need for the development of a new way of managing a vast hoard of information. Automatic summarization can be an indispensable solution to reduce the information overload problem on the web. Not only for the online application, but also for the offline applications such as searching information on a large corpus of offline documents, the text summarization tools can be useful.

Definition of Summary

It is very difficult to define what a summary is. The following are the definitions of "summary" found in various sources.

- **Oxford Dictionary Online (2009):** "A brief statement of the main points of something."
- **Cambridge Dictionaries Online (2009):**"A short clear description that gives the main facts or ideas about something."
- **Radev et.al. (2002):** "A text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that. Text here is used rather loosely and can refer to speech, multimedia documents, hypertext, etc."

Definition of Automatic Summarization

The following are the definitions of "automatic text summarization" found in various sources.

- **Mani and Maybury (1999):** "To take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs."
- **Mani (2001):** "A process to produce a condensed representation of the content of its input for human consumption."
- **Sparck Jones (1999):** "A reductive transformation of source text to summary text through content condensation by selection and/or generalization on what is important in the source."