Chapter 12

Automatic Query-Focused Summary Generation System for Tourism Discourse Using Rhetorical Structure Theory: Cognitive and Multimodal Approach to Tourism Discourse

Subalalitha C. N.
SRM University, India, India

Balaji J.
Wipro Technologies, India

ABSTRACT

Summary generation systems when integrated with Information Retrieval (IR) can give an idea about the retrieved web pages to the user before even the user opens the web page. The summary could be generated for a single web page or for a set of web pages retrieved for a given query. When such a system is built for tourism web sites, the user can get a summary of a particular tourist spot or about the tourist spots present in a particular place. This chapter describes about such a summary generation system which is built using Rhetorical Structure Theory (RST). RST is a well-known discourse theory which is used for discourse analysis of text documents. The RST makes use of another semantic representation namely, Universal Networking Language (UNL) to find the coherent text fragments. These coherent text fragments are indexed and linked with an IR system. When a user gives a query, the web pages along with a single document and multi document summary.

DOI: 10.4018/978-1-5225-2930-9.ch012
INTRODUCTION

Tourism websites play a major role in attracting the tourists and apparently, Information Retrieval systems (IR) play a major role in retrieving the appropriate tourism websites for the user when they give a query. Since IR systems retrieve thousands of tourism specific websites for a single query and in turn, a single tourism website may contain variety of information in it, the users may find difficulty in locating their desired web page and even if they locate, they may find difficulty in locating the desired information.

An automatic query-focused summary generation system may alleviate these difficulties by retrieving the query based summary for each web page retrieved by the IR system. By reading the summary, the user will come to know if the web page associated with the summary is relevant to their query. This will help the users to avoid reading many irrelevant web pages. This chapter proposes such a query focused summary generation system which generates a summary for each web page retrieved by the IR system. The query-summary relevance increases when the web page is semantically analyzed, indexed. This chapter makes uses of two semantic representations namely, Rhetorical Structure Theory (RST) and Universal Networking Language (UNL).

Semantic analysis of text can be done at various levels. Semantics can be incorporated for a word, clause, sentence, paragraph present in a text document. Semantic analysis adds semantical relations using a Lexical Knowledge Base and increases the comprehension of the text. This leads to the meaningful understanding of the text which reflects in the applications that are built on top of such text analysis. Discourse analysis is also equivalent to semantic analysis but refers to text analysis done beyond words level. Discourse analysis may focus on resolving the reference expressions, relating the coherent texts. Discourse Representation Theory (DRT), RST are few popular theories which aids in the discourse analysis. The text representation that results from the discourse analysis is called a discourse structure.

This chapter makes use of RST to construct a discourse structure for the tourism web pages. The web pages need to be indexed efficiently in order to be retrieved by the summary generation system. The RST based discourse structure formed for the web pages aids the indexing as the texts are coherently related by RST relations. Since the text fragments of the web pages are semantically coupled with RST relations, the text fragments are indexed by the semantic relations along with the topic specific words as they indicate the importance of the web page.

The summary generation methodology discussed in this chapter makes use of one more text representation namely, Universal Networking language (UNL) which forms a language-independent graph connecting the words using semantic relations. This RST representation is built on top of the UNL graphs by exploring the similarities between the UNL and RST (Subalalitha and Ranjani Parthasarathi, 2015). The UNL links two words, whereas RST links two clauses, sentences, paragraphs and documents. The summary is generated from the UNL-RST representation.

The summary generation system is evaluated using Forum For Information Retrieval (FIRE) documents using ROUGE score which is the standard metric to analyse summaries (Lin & Chin-Yew, 2004). The chapter describes the summary generation system which can be integrated with an Information Retrieval system (IR) system.