Discourse Marker Based Topic Identification and Search Results Refining

Nipun Suwandaratna† and Udayangi Perera‡
Informatics Institute of Technology
Colombo, Sri Lanka
email: †nipun.sa@gmail.com, ‡udayangi@iit.ac.lk

***Abstract*—Most research oriented search queries consist of multiple topics belonging to one or more domains of knowledge. The objective of such search queries is to find the relationship or impact that each topic has on the other(s). Though web search engines provide easy means to retrieve information off the web, search engines are mainly key word oriented and do not consider the different topics and relationships between such topics present in the results.**

This project presents a software solution, 'SearchDroid' that acts as an intermediate user between a search engine and the end user, refining search results based on different topics identified in the search query and their presence and relationships depicted in search results. A discourse parsing approach has been used to build discourse structures that represent the rhetorical relations of text in search results; this is used to re-rank results based on topics identified in the search query. The project combines linguistics research under discourse parsing with web information retrieval techniques.

The lack of literature combining discourse parsing techniques with web information retrieval has been compensated for by introducing a fresh algorithmic approach. An abstract information retrieval mechanism has been created with the use of discourse parsing techniques, and can be integrated into any web information retrieval approach.

The system was evaluated by Linguistic Experts, Technical Experts and End users. All experts agreed that a discourse parsing approach was suitable for addressing the problem at hand and that the project had high research value in the aspect of linking linguistics research with web information retrieval research. Hands-on testing of the system by End Users produced high user acceptance of the proposed system.

I. Introduction

Research oriented queries are generally the combination of one or more sub topics formulated with the intention of identifying the impact of the scenario mentioned in one topic on the other(s). Topics being considered under a single research may belong to the same larger field (domain) of research or belong to completely different fields of research. In such research the queries required to find information on the web may contain multiple keywords that represent multiple topics. In such scenarios, the results returned by search engines focus on the presence of keywords and not the relationship between different topics contained in the query. This may result in search engines producing results that are relevant to the query only by means of keywords, but do not contain value to the users when considering their research objectives.

This research attempts to model a software application that acts as an intermediate user between a commercial search engine and the end user, refining search results returned by the search engine to better match user information requirements. The SearchDroid system initially addresses the problem domain of difficulties faced by users when carrying out research that contain multiple topics, and require information relating to relationships between such topics. A discourse parsing approach is used to build a discourse structure that represents the rhetorical relations of text which enables the ranking of each search result based on the level of relevance it shows between topics identified in the search query. Therefore, this project combines linguistics research under discourse parsing with web information retrieval techniques.

II. Robust Discourse Parsing for Web Information Retrieval

Discourse refers to pieces of language that are longer than a sentence [1]. Discourse Parsing is a method used mainly in text summarization to summarize text by building up discourse structures. A discourse structure represents the rhetorical information of text [2]. Rhetorical relations are relations that hold across two or more text spans [3]. Katja Jasinskaja [4] explains Rhetorical Relations as, "Rhetorical relations (RR) hold between sentences and clauses in a coherent discourse and may reflect the contentful relations between events or situations described (e.g. cause, temporal succession), or the presentational strategy pursued by the speaker in order to produce a certain effect on the hearer (e.g. contrast, evidence)."

Discourse Markers are connectors in the language that help maintain the flow. They can show the connection between what is being said and what has already been said [1], thus, allowing the clear expression of the idea meant to be conveyed. Words such as *due to, because of, therefore,* and *as a result* are examples of discourse markers and can be used in Discourse Parsing.

Recent approaches to achieve robust discourse parsing require an extensive set of training data corpus [2]. For

978-1-4244-8551-2/10/$26.00 ©2010 IEEE