DR-LINK: Document Retrieval Using Linguistic Knowledge

Elizabeth D. Liddy, Sung H. Myaeng

School of Information Studies
Syracuse University
Syracuse, NY 13244

PROJECT GOALS

DR-LINK is a modular information retrieval system which takes a conceptual-linguistic approach to document detection by satisfying two apparently opposing task requirements: the need to handle large numbers of documents efficiently and the need to represent and retrieve on well-specified information needs. DR-LINK's approach is to enrich the semantic representation of the texts, while focusing its processing on those documents which have real potential of being relevant to a user's query. DR-LINK consists of six modules which, in combination, produce textual representations that capture great breadth and variety of semantic knowledge which will be used to improve retrieval effectiveness, in terms of both recall and precision. To produce this enriched representation, the system uses lexical, syntactic, semantic, and discourse linguistic processing techniques for distilling from documents and topic statements all the rich layers of knowledge incorporated in their deceptively simple textual surface and producing a representation which has been shaped by all these levels of linguistic processing. Specifically, these modules: 1) create summary-level content-vector representations of each text; 2) assign conceptual categories to all proper-noun entities; 3) delineate each text's discourse-level structure; 4) detect relations among concepts; 5) expand lexical representations with semantically-related terms, and; 6) represent and match concepts and relations via Conceptual Graphs.

CURRENT STATUS

The system's six modules are:

- * Subject Field Coder
- * Proper Noun Interpreter
- * Discourse-level Text Structurer
- * Relation-Concept Detector
- * Conceptual Graph Generator
- * Conceptual Graph Matcher

Although our system is now functional, it was run with incomplete knowledge bases, partial implementation of some modules, absence of some important functionalities, and only minimal integration of the output from early system modules by later modules.

RECENT RESULTS

At the 18th month TIPSTER evaluation meeting, the full DR-LINK System was run on 25 Topic Statements against the Wall Street Journal corpora for the ad hoc testing. In addition, the first three system modules were tested in the routing situation on a equal footing with the other systems. For the ad hoc testing, our 11-point precision was .2638. However, the cut-off criterion algorithm which will determine for each individual query how many of the top-ranked documents by the Subject Field Coder (SFC) and Proper Noun Interpreter (PNI) ranking should be processed by the remaining modules was not implemented. Therefore, the full system simply ran against the top 2,000 ranked documents for each query. Once the algorithm is in place, there will be a reasonable mathematical means for determining how many documents should be passed on to later modules so that the set will contain all the relevant documents. In addition, some of the modules were tested alone or in combination as system runs. For example, the 11-point average precision of the SFC + PNI run was a respectable .2245. And although the cut-off criterion was not implemented, by simply ranking the documents in terms of their SFC + PNI similarity to the Topic Statements, all of the relevant documents were ranked in the top 28% of the database.

PLANS FOR THE COMING YEAR

Our major thrust in the months ahead is to complete the system's unfinished knowledge bases and algorithms, and to fully integrate the rich representations which the various system modules produce. We are still analyzing results which will suggest necessary adjustments. Our goal is to accomplish the very rcfined level of matching which the system is capable of producing.