DR-LINK System: Phase I Summary

Elizabeth D. Liddy & Sung H. Myaeng
Syracuse University
4-206 Center for Science & Technology
Syracuse, New York 13244-4100
liddy@mailbox.syr.edu; shmyaeng@mailbox.syr.edu

1. Description of System

1.1 Approach

The underlying principle of the DR-LINK System is that retrieval must be at the conceptual level, not the word level. That is, a successful retrieval system must retrieve on the basis of what people **mean** in their query, not just what they say in their query. The same is true of documents - their representation needs to capture the content at the conceptual level of expression. To accomplish this human-like goal, DR-LINK aims to represent and match documents and queries at all of the available levels of linguistic expression at which meaning is conveyed. Accordingly, we have developed a modular system which processes, represents, and matches text at the lexical, syntactic, semantic, and discourse levels of language. In concert, these levels of representation permit DR-LINK to achieve a level of intelligent retrieval beyond more traditional approaches.

The DR-LINK system takes an innovative approach to dealing with the specific characteristics of the information retrieval tasks required in TIPSTER, focusing on the development of a retrieval system where documents as well as queries are enriched with multiple levels of annotation, with the final representation being a network of concepts and relations expressed in a conceptual graph (Sowa, 1984), thereby enabling retrieval based on conceptual relations. Relations are extracted and represented throughout the system at many levels, ranging from relations between words, to case relations between arguments of a verb, to discourse level relations involving whole sections of text.

The system's conceptual processing was particularly motivated by various semantic restrictions often found in the TIPSTER topic statements. A retrieval system needs to be able to process natural language sentences and extract key concepts and the implicit relations among them, which cannot be expressed as a set of keywords or phrases. For example, it may be crucial to detect documents that talk about a dispute between Airbus and an aircraft company (i.e. the specific relationship between the two concepts), not just about dispute, Airbus, and aircraft company in isolation. Additional relations, e.g. the discourse level relations of 'pending' or 'consequence of' are essential requirements of topic statements that need be fulfilled in relevant documents.

In order to achieve conceptual level representation, we have implemented a range of methods for detecting concepts and extracting relations from natural language sentences in a large text database, by detecting domain-independent linguistic patterns that reveal relations between concepts, which are contained in the set of knowledge bases. Our efforts at knowledge base construction were geared toward general-purpose use in a variety of text processing applications and were guided by corpus statistics, machine-readable lexical resources, and linguistic theories.

1.2 Processing Flow

The DR-LINK system employs sophisticated, linguistically-oriented text processing techniques throughout in order to capture the necessary conceptual information in texts. Since various modules in the system require different annotations of the texts, we opted for staged processing rather than a single-stage full parsing. The system was developed in a modular fashion and functional modularity has been achieved. As currently configured, DR-LINK performs a staged processing of documents, with each module adding a meaningful annotation to the text. For matching, a topic statement undergoes analogous processing to determine its relevancy requirements for documents at each stage.

Among the many benefits of staged processing are: improvements and adjustments can be easily made within any module; the contribution of the various

stages can be empirically tested by simply turning them on or off; modules can be re-ordered for better utilization of document annotations, and; individual modules can be incorporated in other evolving systems. The full flow is described in detail in the following section.

1.3 Description of Key Modules

As currently configured, the DR-LINK System (Figure 1) consists of two stages of processing. The system description and results will be reported according to this division. All of the documents in a corpus are processed by the modules of Stage One, which produces a ranked list of documents according to their predicted degree of relevance to an individual query. In the TIPSTER environment, this ranking was used in two ways: one was for submission of test-runs for the twenty-four month relevance assessment evaluation; the second was for provision of a selected set of documents for further analysis and matching by Stage Two. As seen in Figure 1, Stage One consists of the modules beginning with the Text Structurer which lead into the Integrated Matcher, while the modules beginning with the Relation Concept Detector which lead into the Conceptual Graph Matcher comprise Stage Two.

The **Preprocessor** provides a clean, standardized set of documents and queries for use in both Stage One and Stage Two. The Preprocessor subdivides newspaper texts into their subtexts on the basis of orthographic clues. This means that long newspaper articles which concatenate numerous reports will be accurately processed as separate stories. Part-of-speech tags are added to the pre-processed texts using our C version of BBN's POST (Meteer et al., 1991). Since several modules require segmentation of individual sentences, we apply our constituent boundary bracketer to the tagged texts to identify boundaries of clauses and noun phrases using a relatively simple pattern-driven processing.

1.3.1 Stage One Modules

The **Text Structurer** is based on discourse linguistic theory which suggests that texts of a particular type have a predictable text-level structure which serves as an indication of how and where certain information endemic to a text-type will be conveyed. We have implemented a Text Structurer for the newspaper text-type, which produces an annotated version of a news article in which each clause or sentence is tagged for the specific slot it instantiates in the news-text model, e.g. MAIN EVENT, EXPECTATION, CONSEQUENCE. The structural annotations are used to respond more precisely to information needs expressed in Topic Statements, where some aspects of relevancy requirements such as time, source, intentionality, and state of completion can only be met by understanding a Topic Statement's discourse requirements (e.g. the **consequences** of automation; a **proposed** theme park development).

The Text Structurer assigns news-text component labels to document clauses/ sentences on the basis of four types of linguistic evidence learned from text. We have reduced the matching complexity via a function that maps the thirty-eight news-text components which are recognized in documents to seven meta-component requirements which are recognized in Topic Statements. This allows the system to impose fine-level structure on newspaper articles with excellent precision and to map this fuller set of text components to the appropriate level of discourse requirement specificity typically expressed in Topic Statements.

The **Subject Field Coder** (SFCoder) uses an established semantic coding scheme from the machine-readable Longmans Dictionary of Contemporary English (LDOCE) to tag each word in a text with its disambiguated subject code (e.g. Agriculture, Military, Political Science) and to then produce a fixed-length, subject-based vector representation of each document's and query's contents. Using the SFCoder, each text or sub-text is represented as a vector of the normalized frequencies of the SFCs for that unit's constituent words. The normalized SFC vectors represent the implicit semantics of text at a level of abstraction above the word level.

The **V-8 SFC Matcher** combines the annotations of the Text Structurer and the SFCoder in a representation that captures the distribution of SFCs across the discourse meta-components that occur in a document. Up to eight SFC-vectors are produced for each document - one for each of the seven Text Structure meta-components, plus one for the combined categories. Experimentation with several formulas for combining the similarity values of the meta-component SFC vectors which are responsive to a particular query indicated that the Dempster-Shafer algorithm is superior. The V-8 Matcher's unique combination of discourse semantics and lexical semantics produced a 13% improvement in precision over matching on just lexical semantics.

The **Proper Noun (PN) Interpreter** uses a variety of processing heuristics and knowledge bases to

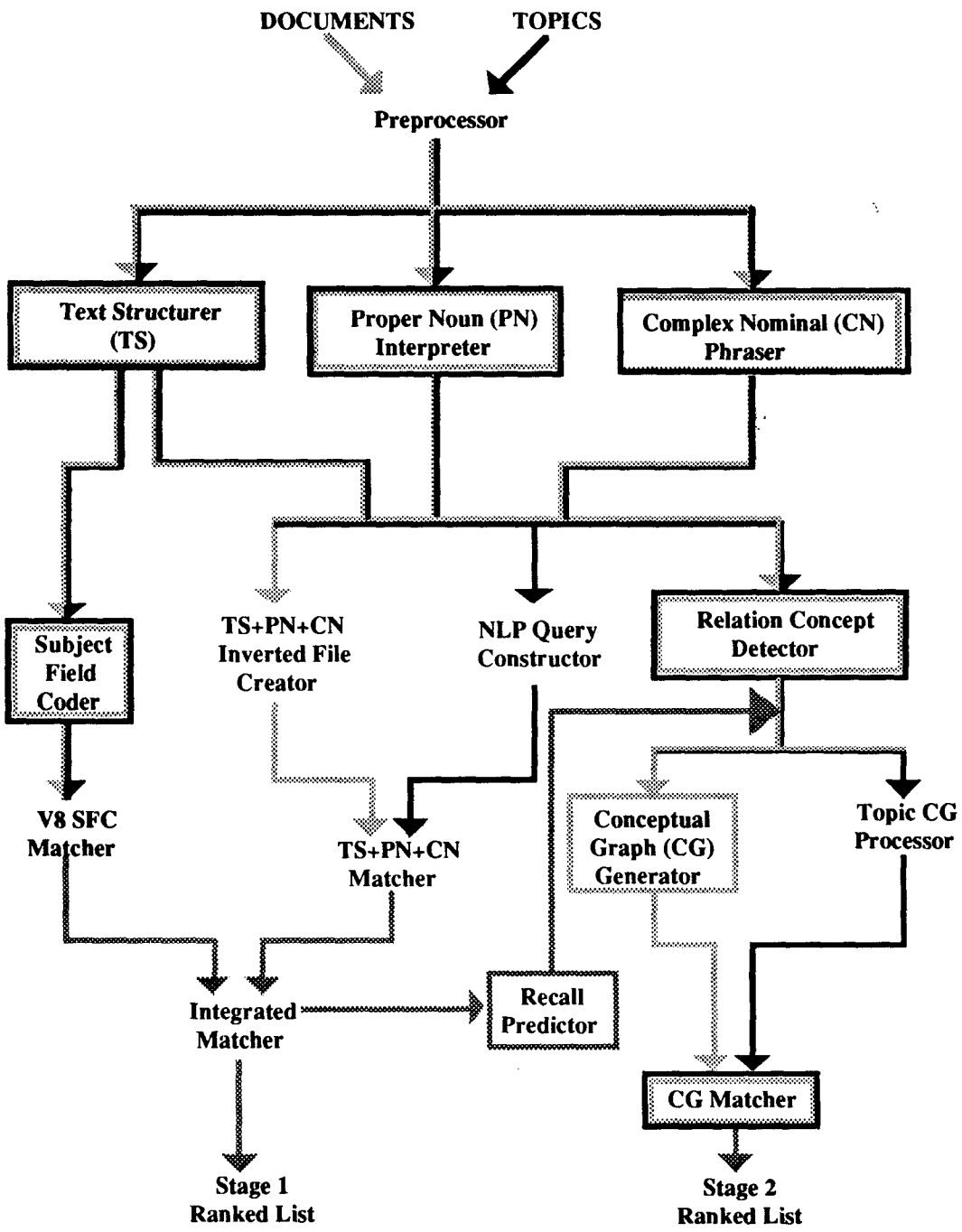

Figure 1. DR-LINK System

produce: a canonical representation of each PN; a classification of each PN into one of thirty-seven categories (e.g. organization, country, company), and; an expansion of group nouns into their constituent PN members (e.g. European Community to all member countries). In addition, the accurate indexing of multiple references to a proper noun entity throughout a document permits complete representation of the multiple relational links implicitly contained in the article. The module recognizes and categorizes proper nouns with 93% accuracy using 37 categories as tested on a sample of 545 PNs from newspaper text. The representations produced by the PN Interpreter are used in Stage One matching and also provide rich relational

information later used by the CG Matcher.

The **Complex Nominal Phraser** provides a means for precise matching of complex semantic constructs when expressed as either adjacent nouns or a non-predicating adjective + noun pair. We have focused on complex nominals because they were observed to convey much of the conceptual content of a text (e.g. debt reduction, campaign financing, electronic theft). In addition, we have been experimenting with computational means that permit the system to produce conceptual matches when the constructs are expressed in synonymous phrases, thereby addressing issues of both recall and precision. Substitutable phrases for each CN can be found by statistical corpus analysis which identifies all second order associations between each CN constituent and terms in the corpus. Terms that exhibit second order associations (terms used interchangeably in certain contexts) are compiled into equivalence classes for use by the matching algorithms.

In addition, each complex nominal and its assigned relation in a query provides a concept-relation-concept triple to the later Relation-Concept Detector module for use in Stage Two matching. Semantic relations between the constituent nouns of each complex nominal in the query are stored in a knowledge base, using an ontology of forty-three relations. The existence of both case-frame relations and complex nominal relations makes it possible for the system to detect conceptual similarity even if expressed in different grammatical structures by means of a relation-similarity table that assigns a degree of similarity across the two grammatically-distinguished sets, and a degree of similarity between pairs within the same set. The relation-similarity table is used in Stage Two matching to allow concepts that are linked by a relation in a document that is different from the relation that links the same two concepts in a Topic Statement, to still be awarded some degree of similarity.

The **Natural Language Query Constructor** for Stage One takes as input a natural language Topic Statement and produces a query which reflects the appropriate logical combination of the Text Structure, Proper Noun, and Complex Nominal requirements of a Topic Statement. The basis of the Query Constructor (QC) is a sublanguage grammar which is based on a generalization over the regularities exhibited in the Topic, Description, and Narrative fields of the one hundred and fifty TIPSTER Topic Statements. The QC sublanguage grammar relies on items such as function words, meta-text phrases, and punctuation to recognize and extract the logical combination of relevancy requirements of Topic Statements. The QC sublanguage interprets a Topic Statement into pattern-action rules which translate each sentence into a first order logic assertion, reflecting the boolean-like requirements of Topic Statements, including NOT'd assertions and resolved definite noun phrase anaphora.

The **TS+PN+CN Matcher** evaluates each logical assertion produced by the Query Constructor against the entries in the inverted document file and assigns a weight to each document segment if it matches the PN + CN Boolean requirements of the Topic Statement. If the document segment also matches the Topic Statement's Text Structure requirement, this weight is increased. Depending on which field in the Topic Statement the assertion came from, the preliminary value will be weighted by a co-efficient reflecting importance as indicated in the Topic Statement. The highest similarity value for a single assertion in the document is selected as that document's explicit similarity match to the Topic Statement.

The **Integrated Matcher** combines the 'explicit similarity' value as determined by the TS+PN+CN Matcher with the 'implicit similarity' value as determined by the SFC V-8 Matcher and an integrated similarity score for each document is produced. This similarity value is used in several ways: 1) to provide a full ranking of all the documents for the Stage One Ranked List, and; 2) as input to the Recall Predictor, a filter which determines for each new query how many documents from the ranked list should be passed to Stage Two in order for this set to contain 100% of all the relevant documents for that query.

The **Recall Predictor**'s filtering function is accomplished by means of a multiple regression formula that successfully predicts a ranked-list cut-off criterion for individual queries based on the similarity of documents to the query in terms of their SFC, Proper Noun, Complex Nominal, and Text Structure requirements (Liddy et al, 1993b). The real power of the Recall Predictor is its sensitivity to the varied distributions of similarity values for individual queries. For a few queries, a good portion of the ranked list may need to be passed to Stage Two. However, for most queries, a relatively small portion of the database needs to be passed to Stage Two in order to guarantee the potential of 100% recall. For example, to achieve 100% recall for Topic Statement 42, the regression formula predicts a cut-off criterion similarity value which requires that only 13% of Stage One's ranked output be processed by later modules. The available relevance judgments have shown that this pool of documents contained 99% of the documents judged relevant for that

query. Fuller performance results on the Recall Predictor are included in Section 5.2.

1.3.2 Stage Two Modules

The **Relation-Concept-Detector** (RCD) provides building blocks for the Conceptual Graph (CG) representation by generating concept-relation-concept (CRC) triples based on the domain-independent knowledge bases which have been constructed with machine-readable resources and corpus statistics. In this module, there are several handlers that are activated selectively depending on the input sentence. In addition, the rich relational annotations received from the Proper Noun Interpreter and the relations between constituents supplied by the Complex Nominal Phraser contribute to the CRC output of the module.

The Case Frame (CF) Handler generates CRC triples where one of the concepts comes usually from a verb. It identifies a verb in a sentence and connects it to other constituents surrounding the verb. Since the relations (about 50 are used currently) included in our representation originate from theories of linguistic case roles, and are all semantic in nature, this module consults the knowledge base containing 13,786 case frames we have constructed, each of which prescribes a pattern involving a verb and the corresponding concept-relation-concept triples.

The CF Handler selects the best case frame by attempting to instantiate each case frame and determine which one is satisfied most by the sentence at hand. This can be seen as a sense disambiguation process using both syntactic and semantic information. The semantic restriction information contained in the case frames were obtained from LDOCE, and when the sentence is processed, the CF Handler also consults LDOCE to get semantic restriction information for individual constituents surrounding the verb in the sentence and compares it with the restrictions in the case frames of the verb as a way to determine which case frame is likely to be the correct one. For example, with the sentence fragment, ... the chairman declined to elaborate on the disclosure ..., the CF Handler chooses an appropriate case frame and produces

[decline] -> (AGENT) -> [chairman]
[decline] -> (ACTIVITY) -> [elaborate]
[elaborate] -> (AGENT) -> [chairman].

The Nominalized Verb (NV) Handler consults the NV case frames to identify a NV in a sentence and create CRC triples based on the rule. At the same time, it converts the NV into its verb form. In this way, we can allow for a match between a CG fragments generated from a phrase containing verb and another fragment generated from a noun phrase containing the corresponding nominalized verb. For example, the NV Handler converts the sentence fragment, ... the company's investigation of the incident ... , into

[investigate] -> (AGENT) -> [company]
[investigate]-> (PATIENT) -> [incident].

This process is much more than a sophisticated way of performing stemming in that we canonicalize concept-relation-concept triples rather than just concept nodes.

The Ad-Hoc Handler looks for lexical patterns not covered by any of the other special handlers. Its processing is also driven by its own knowledge base of patterns to infer relations between concepts. The knowledge base contains a small number of simple patterns involving BE verbs and more than 350 pattern rules for phrasal patterns across phrase boundaries (e.g. `... for the purpose of ...' reveals the relation GOAL), by which important relations are extracted. The pattern rules specify certain lexical patterns and the order of occurrences of words belonging to certain part-of-speech categories, and the CRC triples to be generated. These patterns require a processing capability no more powerful than a finite state automaton.

The **Conceptual Graph Generator** merges individual CRC triples generated for a document to form a set of conceptual graphs, each corresponding to a clause in most cases. Since more than one handler can generate different triples for the same concept pairs (e.g. a prepositional phrase handled by the CF handler and the NP/PP handler) based on independently constructed rules and on independent processes, a form of conflict resolution is necessary. In the current implementation, we simply order the execution of different handlers based on the general quality of the rules and the resulting triples so that more reliable handlers have higher precedence.

For semi-automatic processing of topic statements for CG generation, the current system first applies the same RCD and CG generator modules to produce topic statement CGs. Several topic statement-specific processing requirements have been identified, some of which have been implemented as post-processing routines and others are under development. These include: elimination of concept and relation nodes corresponding to contentless meta-phrases (e.g.

"Relevant document must identify ..."); handling of negated parts of topic statements; automatic assignment of weights to concept and relation nodes, and; merging common concept appearing in different sections of topic statements.

The **RIT Coder** adds Roget's International Thesaurus codes to individual concept nodes as a way to make our current representation more "conceptual", so that the label on the nodes is not a word but a position of the hierarchy of RIT. The lowest level position beyond individual lexical items in the RIT hierarchy consists of several terms within the delimiter of semi-colons, which represents a concept.

The mapping from a word (called target) in text to a position in RIT requires sense disambiguation, and our approach is to use the words surrounding the target word as the context within which the sense of the target word is determined and one or more RIT codes are selected. The algorithm selects minimal number (i.e. one or more) of RIT codes, not just the best one, for target words since we feel that some of the sense distinctions made in RIT are unnecessarily subtle, and it is unlikely that any attempts to make such fine distinctions would be successful and hence contribute to information retrieval.

The **Conceptual Graph Matcher**'s main function is to determine the relevance of each document against a topic statement CG and produce a ranked list of documents as the output of Stage Two of the system. Using the techniques necessary to model plausible inferences with CGs, this module computes the degree to which the topic statement CG is covered by the CGs in the document.

While our approach has the ability to enhance precision by exploiting the structure of the CGs and the semantics of relations in document and topic statement CGs, and by attempting to meet the specific semantic constraints of topic statements, we also attempt to increase recall by allowing flexibility in node-level matching. Concept labels can be matched partially (e.g. between `Bill Clinton' and `Clinton'), and both relation and concept labels can be matched inexactly (e.g. between `aid' and `loan' or between `AGENT' and `EXPERIENCER'). For both inexact and partial matches, we determine the degree of matching and apply a multiplication factor less than 1 to the resulting score. For inexact matching cases, we have used the relation similarity table, described above in the Complex Nominal Phraser section, to determine the degree of similarity between pairs of relations. Although this type of matching slows down the matching time, we feel that until we have a more accurate way of determining the conceptual relations and a way to represent at a truly conceptual level (e.g. our attempt to use RIT codes), it is necessary. More importantly, the similarity table reflects our ontology of relations and allows for matching between relations produced by different RCD handlers whose operations in turn are heavily dependent on the domain-independent knowledge bases.

1.4 Hardware/Software Requirements

The DR-LINK System has been running in a university environment on a Sun Sparc workstation, running UNIX, with a C compiler. The only special-purpose software required is a part-of-speech tagger that is a C version of the POST tagger (Meteer et al, 1991). We are currently developing our own tagger based on morphological statistics (Liddy & McHale, in press). No special-purpose hardware is required to run the DR-LINK system.

1.5 Efficiency/Speed/Throughput Statistics

There are two points to be made regarding the notion of efficiency. First, there is no question that the sophisticated text processing and document retrieval using the rich representations described above can be done only at the expense of processing time requirements. The primary goal of the project to date has been to focus on developing and implementing new ideas in a prototype without much concern for efficiency. Whenever there was a choice between efficiency and a potential for improved effectiveness through richer representation and more sophisticated processing, we chose the latter. In other words, our design goal was to include as many features as possible so that they can be tested not only as a whole but also individually.

Secondly, since we conduct our research in a university environment, DR-LINK does not have a dedicated server and must time-share computing resources with all students and faculty. Therefore the throughput reported here is the lower bound to be expected of even this prototype system. Additionally, since the text processing and matching were done on different machines shared by multiple processes, it was not possible to generate statistics on efficiency of the system which can be used reliably as a prediction of throughput of a future operational system using the algorithms in the current experimental system.

There are a few points we can report in this regard. Stage One of the system can analyze and annotate the representation of an estimated 1.5 Megabyte of documents an hour. We do not have reliable times on the matching for Stage One. Currently, we are putting our effort into extending and optimizing Stage One's matching process for our ongoing tech-transfer efforts.

In order to provide some indication of how much time was spent running Stage Two of the system for the 24th month evaluation, we gathered statistics using a small sample of documents on a SUN SPARC-10 workstation with 128 RAM for both concept-relation-concept generation and matching. The RCD module processed the already POS-tagged and bracketed text at the rate of approximately 12K bytes per second to produce concept-relation-concept triples.

The result of the sample runs of the CG matcher in the same computing environment shows a wide range of time requirements depending on topic statements. With a sample Wall Street Journal documents and three topic statements, the average matching times per document varied from 1 second to 12 seconds. It appears that the time required for matching depends on such factors as: the number of nodes in the topic statements; the frequency of topic statement nodes in document conceptual graphs (i.e. specificity of concept nodes); the extent to which similarity tables (concept term similarities, relation similarities, and complex nominal similarities) need to be looked up for inexact matching; and the connectivity and size of conceptual graphs for both topic statement and document.

1.6 Key Innovations of System

One unique aspect of DR-LINK is its ability to represent content at the conceptual level rather than the word level to enable intelligent information processing which mimics the multiple levels of language comprehension used by humans in understanding text and determining whether information is relevant to a query. The output of Stage One combines the lexical, syntactic, semantic, and discourse levels of understanding into a single prediction of a document's relevance for a query. The key innovations can be summarized as follows. Stage One of DR-LINK:

• Accepts as input, a user's lengthy, ambiguous, complex, natural language query, which it translates into a precise Boolean representation of user's relevance requirements.

• Produces summary-level, semantic vector representations of queries and documents which are used to quickly provide a reliable ranking of large sets of documents at the subject-domain level. The application of a tested regression formula determines a cut-off criterion for an individual query so that quick and accurate filtering of large data sets can be done.

• Fulfills the focused proper noun requirements of queries with highly accurate proper noun categorization and controlled expansion of proper names via all variants of proper noun entities and expansion of proper noun categories to their constituent members.

• Has the capability of providing high recall and precision simultaneously, via controlled expansion of complex nominals using lexical resources for expansion of each member of the complex nominal combined with corpus statistics on phrase contexts so that the substantive content of a query can be matched in its synonymous phrasings (e. g. capital spending -> capital expenditures -> equipment expenditures; trade ban -> trade sanctions -> export sanctions).

• Detects the implicit significance, temporal, credibility, state of completion, and intention aspects of information in documents by use of the inherent discourse-level structure of various text types. These more holistic information aspects regarding the content of a document are conveyed via discourse-level features and frequently cannot be detected by reliance on word or sentence level linguistic features.

• Combines evidence from the multiple levels of linguistic processing done on both the query and documents to assign a degree of belief, based on implicit and explicit semantics, represented as a single weight, that a particular document is likely to be relevant.

Stage Two of the system is unique and innovative in its attempt to explicitly satisfy the semantic restrictions contained in topic statements. With the RCD module, we attempted to extract relations between concepts so that documents containing the required concepts that are linked with the required semantic relation would receive a higher score than those without the specified link.

Unlike other rule-based retrieval systems developed for a small domain, the rules in the knowledge bases we developed for relation extractions are domain-independent and were constructed based on corpus statistics and machine-readable lexical resources such as LDOCE. Since the rules are domain-independent and based on general linguistic patterns, they can be applied to any

information retrieval contexts regardless of the types or domains of databases. In addition, the scope and coverage of the knowledge bases can be extended easily with the same methodology that relies on the linguistic patterns of the lexical resources and the corpus.

To our best knowledge, our work is the first attempt to represent document and query contents in conceptual graphs for information retrieval and model information retrieval as conceptual graph matching which is a form of inferencing. With this representation, furthermore, we incorporated Dempster-Shafer's theory of evidence (Shafer, 1976) as the basis for computing the similarity between a document and a query so that information retrieval can be modeled as a process of gathering evidence under uncertainty (Myaeng & Khoo, 1993). As a result, our development of the DR-LINK system has laid a firm foundation for using relations and structural representation of documents and user needs in information retrieval environments where a large volume of texts needs to be handled.

2. Original Project/System Goals

As required, we proposed to develop a system which would be portable, modular, extensible, and domain independent. With the domain and language independence requirements, one of the main emphases of the original project was to develop a learning module that would acquire structure-revealing and relation-revealing patterns over time by processing texts with or without human intervention. It became apparent at an early stage of the project that it would be more fruitful to de-emphasize the goal of making the learning module as autonomous as possible and shift our attention to developing linguistic knowledge bases and algorithms.We now find, at the end of Phase I, that certain modules are at the point where we can exploit machine learning. We now have an understanding of the meta-processes involved in these modules, including the nature of their underlying models, and the linguistic evidence which the system needs for learning.

To an extent, we believe that the original goals of portability, modularity and extensibility have been achieved in certain modules of the system. For example, the Subject Field Coder algorithms, including the semantic disambiguation algorithm, have been successfully ported to another machine-readable lexical resource, and are producing vectors which provide retrieval results equal to the original implementation with LDOCE. Modularity has quite demonstrably been achieved as evidenced by our re-ordering of the Stage One modules for the twenty-four month test runs in order to produce the highly successful V-8 vectors which combine SFC and Text Structure information. In terms of extensibility, the Subject Code vector approach to implicit semantic level representation of text content has been proposed for extension into a multilingual environment, where it will be used to provide cross-language semantic representation and access to documents in six languages (Liddy et al, 1993a).

Another important aspect of the original project was to explicitly deal with user information needs by developing a method of constructing and using user profiles for the purpose of better representing user information needs. The user profiles in this context are much more extensive than the notion of profiles for routing. With the practical difficulty of accessing real users for modeling and testing purposes, however, the idea was dropped at an early stage.

3. Evolution of System over Two Years

We had originally envisioned a less modular, more integrated processing of texts which would basically combine two major levels of representation: discourse level structure and semantic relations as expressed in Conceptual Graphs. Also, we had originally planned for Conceptual Graph representations to be constructed for all the documents. When it began to appear that this was unreasonable due to the processing requirements of the CG Matcher, we introduced the Subject Field Coder as a means of limiting the number of documents which would need to be processed by Stage Two. The SFCoder has proven to be not only a reliable predictor of relevant document sets, it also has proven to be a unique way of adding implicit semantics to our representations. In addition, the SFC vectors provide an excellent representation on which to cluster the collection for browsing (Liddy et al, 1992).

We had not originally intended to have a system with two such distinct stages as are present in the current system. The original goal was a continued enhancement of text as it passed through the full length of the system. However, the exigencies of ARPA testing required that the system produce ranked output for comparative evaluation for the eighteenth month testing. Since the Relation Concept Detector and CG Generator and Matcher (which were originally planned to produce the only ranking of documents) were not completed, we started providing ranked results from the first few modules, which eventually became known as Stage One. These results were surprisingly good even though many important aspects of retrieval had not yet been included.

Subsequently, additional levels of representation and matching were added to Stage One as separate modules. The modular system provided for an environment in which empirical testing of the contribution of various levels of linguistic processing could be performed, as well as an environment in which the results of our detailed analysis of Topic Statement requirements and retrieval results which guided our decisions, could be acted on with the addition of new levels of linguistic analysis. For example, since 85% of the Topic Statements are in some way concerned with proper noun entities, we began the second year of the project with a focus on development of the Proper Noun Interpreter which processes the distinctive linguistic constructions which modify proper nouns for the extraction of vital semantic information. In addition, our observation that both recall and precision could be positively impacted by a constrained expansion of topics via the addition of synonymous phrasings of complex nominals, resulted in the development of the Complex Nominal Phraser.

One particular module, the Text Structurer, has been through a very interesting evolution, including quite drastic shifts from a holistic model, to a distributed, attribute model, back to a model which combined aspects of both of these models. (Liddy, In press). In addition, the original Text Structurer implementation used eight sources of linguistic evidence which were evaluated by the Dempster-Shafer Theory of Evidence Combination (Shafer, 1976) to assign a discourse component label to each segment of text. Analysis of the 18th month results and the relative contribution of each of the evidence sources to the module's performance, led us to reduce the number of evidence sources to four, which also allowed for the implementation of a much leaner model of discourse-level processing.

4. Accomplishments

DR-LINK, which is still in the process of development, has just begun to achieve its potential, which is the provision of the depth of matching of information source to information need that now occurs only with the assistance of a human intermediary (e. g. librarian or information specialist) who can interview the user; comprehend their information need at the conceptual, not word level; understand the complexity of ways in which the relevant information might be expressed in various information sources, and; bring the user's need and relevant documents together. This is possible because the human intermediary's understanding of both the information need and the information content of documents is not limited to surface level matching. The intermediary interprets text (both queries and documents) at the multiple levels at which meaning is conveyed in human language: from pure lexical pattern matching; to the disambiguated semantic word-level representation where only the appropriate sense of an ambiguous word is considered by the intermediary; to the semantic relation level where not only the presence of requested concepts occurs, but these concepts also exist in the desired relations to each other (e.g. company A is the buyer, not the seller); to the discourse level where the structuring of the information content throughout the document conveys implicit relations, and the connections between concepts that are distant in text are brought into alignment.

Stage One of the system has succeeded in combining all levels of linguistic analysis in the provision of very promising retrieval performance. For instance, the rankings produced by Stage One combine both implicit and explicit similarity between documents and queries in a new and principled way. In addition, the Recall Predictor makes use of these combined similarities for the accurate prediction of precisely how many documents from the ranked list need to be reviewed by a user in order to achieve a particular level of recall. This capability is not available in other systems. Given the size of current databases and the real need of some users to review every potentially relevant document, the functionality and reliability of the Recall Predictor is a sizeable accomplishment.

One hard-earned achievement is the recent demonstration of the contribution of the discourse level information which is made available by the Text Structurer. One of the major tenets of the original proposal was that discourse structure would positively impact retrieval. During the two years of the project, our ability to provide discourse information about both documents and queries continually increased. However, our ability to utilize this enrichment was slow in developing. The recent implementation of the V-8 representation provides one very appropriate and successful way to incorporate Text Structure knowledge in Stage One. However, we believe that the full potential of this level of information will only be fully realized when its availability and functional capability is known and exploited by future users. Additionally, a recent realization in the field of IR is that retrieval of very long, full-text documents, such as those in the TIPSTER corpus, may require sub-document processing. Although it was originally believed that paragraph level computations might prove useful, results suggest that orthographic divisions in text are not necessarily appropriate. Discourse theory suggests

that the text level model would predict the appropriate subdivisions. We believe that the Text Structure which we can detect in documents will provide the appropriate sub-document units of analysis which are needed for high performance in the retrieval of long documents.

A major accomplishment in Stage Two of the system is that we have laid a foundation for operationalization of an unconventional information retrieval system that can handle a user's semantic restriction explicitly. For the Relation-Concept Detector, it means development of both algorithms and knowledge bases. In the former category are an efficient constituent boundary bracketer and the special handlers (e.g. case frame handler) for different types of pattern rules. For the latter, we have developed: 13,786 case frames, 15,053 nominalized verb patterns, and 350 ad-hoc rules.

Other algorithms have been developed and implemented for the purpose of retrieving documents based on conceptual graphs representations. The RIT coder was developed as an attempt to provide a truly conceptual representation although its efficacy has not been fully tested. The conceptual graph matching algorithm is a flexible retrieval engine capable of modeling uncertainty handling, evidence gathering, and retrieving texts of various sizes (e.g. sentences, paragraphs etc.), when texts are represented as a network of concepts and relations.

While we believe that our accomplishment of building and testing our system in the TIPSTER environment is a significant step toward a breakthrough in information retrieval, the complexity of the approach has forced us to leave many "loose" ends that need to be tightened up and improvements to be made, in order to take full advantage of the power of the algorithms and knowledge bases.

First of all, there are many errors made by individual modules and handlers in Stage Two in generating accurate conceptual graphs. Some errors are propagated through the stages of text processing. For example, the part-of-speech tagger often is confused between VBD (past tense verb) and VBN (past participle). The current bracketer relies on these tags to determine the main verb of a clause or a phrase and makes errors in determining phrase boundaries, which in turn leads the Case Frame Handler to assign incorrect case roles to noun phrases. Since conjunctions are not handled properly either within the bracketer or within the special handlers, primarily due to our design decision to make the algorithms non-recursive for simplicity and efficiency, many errors were made in the RCD output.

These errors had much more severe impact on query construction for Stage Two. We applied the same RCD module to the topic statements with a few additions such as eliminating "meta-phrases" and assigning weights automatically. The frequent occurrences of conjunctions and other sub-language features of topic statements such as parentheticals and examples were not handled by the Stage Two Topic Statement processor in a special way, sometimes resulting in inadequate representations. The negative impact of these errors manifest itself when the retrieval results from the 24th month evaluations are compared to those with manually constructed queries in 18th month evaluations (Myaeng & Liddy, 1993). Although they differ in terms of topics and databases, the 18th month results were superior to the 24 month results.

Another aspect of the RCD module that has not been fully developed is the handling of prepositions. There was no direct attempt in RCD to deal with the unsolved problem of prepositional phrase attachment. We adopted the simple default rule that unless a case role is assigned to a prepositional phrase by the expectation in the case frame (as specified in LDOCE), it is attached to the nearest constituent with a general relation. While the scheme of inexact matching between relations helped alleviating the inaccurate assignments of relations, we observed many remaining problems.

Based on our failure analysis of Stage Two, it appears that all the errors mentioned above and the incomplete nature of some of the handlers (e.g. ad-hoc handler) resulted in the lack of matches on relation nodes when document and query conceptual graphs are matched. In other words, there weren't as many connected graphs in the matching results as expected, and the payoff we expected from all the efforts we put in to extract relations weren't as great.

5. Evaluation Summary

The DR-LINK System was tested as a Type B system, meaning that it was run against a smaller, more homogenous corpora than the other systems. In many ways, therefore, our results are hard to compare to the other detection contractors' performance. Results were produced for the Ad Hoc Topic Statements against the Wall Street Journal collections and the Routing Topic Statements against the San Jose Mercury collection. Stage One queries were automatically constructed, while Stage Two queries were semi-automatically generated in that automatically generated queries were adjusted to merge a small number of common concept nodes appearing across different Topic Statement fields.

5.1 Official Results

Stage One produced seven runs in the Routing situation and three runs in the Ad Hoc situation. As can be seen in Tables 1 and 2, the various Stage One runs produced roughly similar results on the official precision measures. There is little observable differentiation in the results because the conditions tested in Stage One were not of the type to introduce much variation in the top of the ranked lists, but rather, were designed to improve the Recall Predictor's ability to pass to Stage Two a set containing as many of the relevant documents as possible using various methods for combining documents from four possible document groupings. The variation between runs occurred either further down the ranking of one thousand documents which were used to compute precision or, in some instance, even past the one thousand document point.

To understand the variations we were testing, consider each document's similarity value for a given Topic Statement as being composed of two elements. One element is the implicit similarity represented by the SFC vector value and the other element is the explicit similarity value that represents the combined Proper Noun, Complex Nominal, and Text Structure similarities. The system applies the multiple regression formula, and computes the mean and standard deviation of the distribution of the SFC similarity values for the individual Topic Statement. Using these statistical values, the system produces the cut-off criterion value. Since the eighteen-month results indicated that 74% of the relevant documents had an explicit similarity to the Topic Statement and the remaining 26% of the relevant documents had no such explicit similarity, this information was also used in predicting what proportion of the relevant documents should come from which segment of the ranked documents in order to achieve full recall. The groupings which are used to produce the final ranking can be envisioned as consisting of four segments, as shown in Figure 2.

Four groups are required to reflect the two-way distinction mentioned above. The first distinction is between those groups which have an explicit similarity (Groups 1 and 2) and which should contain 74% of the relevant documents and those documents without an explicit similarity (Groups 3 and 4), which should contribute 26% of the relevant documents. The second distinction is between those documents whose SFC similarity value is above the predicted cut-off criterion (Groups 1 and 3) and those whose SFC similarity value is not (Groups 2 and 4).

In Table 1, which presents the Adhoc results, the three runs all use the univector (as compared to V-8) version of the SFC representation. DRwus1 is the "straight

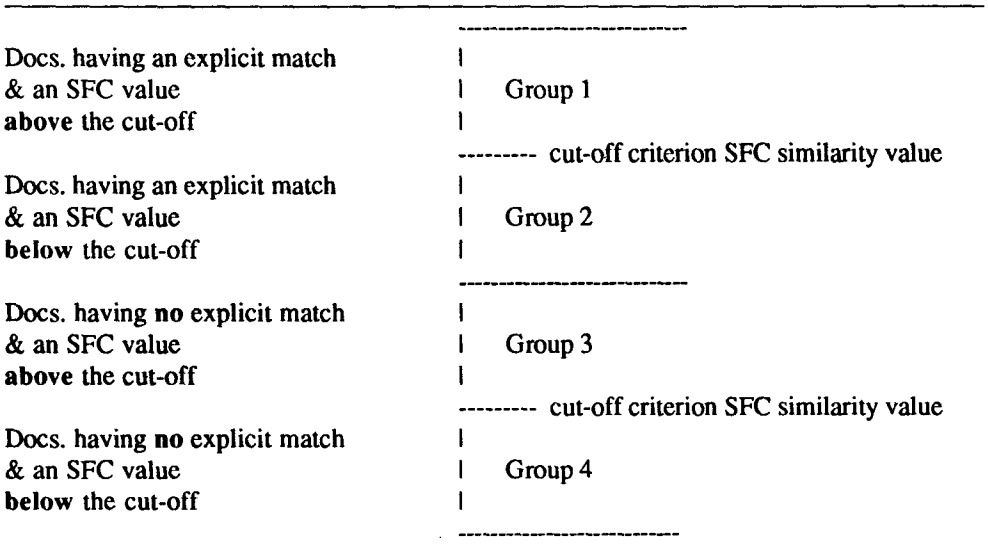

Figure 2: Schematic of Document Groupings

Run	Average Precision	Precision at 100	R-Precision	Relative Precision
DRwus1	0.2237	0.2672	0.2778	0.4004
DRwur1	0.2243	0.2706	0.2793	0.4115
DRwum1	0.2247	0.2682	0.2793	0.4052

Table 1. Stage 1, Adhoc

combination" of document similarity values. This means that the ranking is produced by concatenating the ranks from the groups in sequential order from 1 to 4. Therefore, all documents that have any match on the explicit Proper Noun or Complex Nominal requirements of the Topic Statement will precede documents without any such match. As a result, documents which have any explicit similarity, even though their implicit similarity might be very low, will outrank a document with very high implicit similarity. Therefore, this run emphasizes explicit similarity.

DRwur1 uses the "regression" formula, which incorpoates the predicted percentage of relevant documents with and without an explicit match, to rank the documents. That is, the system will produce the ranking by concatenating the documents above the appropriate cut-off from Group 1; then documents above the appropriate cut-off from Group 3, then documents from Group 2. Therefore, this run emphasizes implicit similarity.

DRwum1 uses a "modified regression" formula to produce rankings. This is an adaptation of the regression approach, since our test results show that there is a potential 8% error in the predicted cut-off criterion for 100% recall. Therefore, we used extrapolation to add the appropriate proportion of the top ranked documents from Group 2 to Group 1, before concatenating documents from Group 3.

Run	Average Precision	Precision at 100	R-Precision	Relative Precision
DRsus1	0.1715	0.1306	0.2047	0.3621
DRsur1	0.1431	0.1258	0.1782	0.2992
DRsum1	0.1638	0.1254	0.1984	0.3231
DRsds1	0.1689	0.1302	0.1980	0.4021
DRsdr1	0.1611	0.1244	0.1951	0.3458
DRsas1	0.1685	0.1290	0.1989	0.4031
DRsar1	0.1443	0.1160	0.1841	0.3178

Table 2. Stage 1, Routing

Comparison of the results shows that reliance on the regression formula, DRwur1, without the modification, produced the highest precision on three out of the four precision measures.

The results from the routing queries (Table 2), which were run against the San Jose Mercury, reflect the fact that the regression formula was developed on earlier queries using the Wall Street Journal as the training corpus. Therefore, it is not surprising that in the first three runs, the "straight" run (DRsus1) surpasses the runs based on the regression formula (DRsur1) and the modified regression formula (DRsum1).

The remaining four runs use the V-8 vectors, which use SFC similarity values based on those segments of the document whose Text Structure match the Text Structure requirement of the Topic Statement. A second variable which was tested in these runs is how to best combine the V-8 similarities. DRsdr1 and DRsds1 are

based on use of the Dempster-Shafer evidence combining algorithm, while DRar1 and DRas1 combine the evidence using a straight average. The Dempster-Shafer algorithm outperforms the averaging approach. The comparisons were run using both the "straight" approach to combining documents from the four groups described above, as well as the "regression" approach. As expected, since the regression formula was trained on the Wall Street Journal, it did not perform as well in the routing situation, as did the "straight" approach.

One result which we would like to point out is the 12.6% improvement in average precision which the V-8 use of discourse structure information provides, as can be seen by comparing DRsdr1 (V-8 combined using the Dempster Shafer algorithm) to DRsur1 (univector). Since the use of discourse level linguistic information is new to the field of information retrieval, this result is a very promising first effort.

Tables 3 and 4 provide document level averages for Stage One, as a common measure for comparison of Stage One and Stage Two output. That is, Table 3 can be compared with Tables 5 and 6 for Stage 2 output on the routing queries and Table 4 can be compared with Table 7 for Stage Two output on the ad hoc queries.

In conclusion, although the official results for Stage One may not be equal to the other contractors' system performance, it should be remembered that the performance goal of Stage One was to enrich the texts with the multiple levels of linguistic representation which would permit Stage One to pass the selected level of recall in the ranked lists provided to Stage Two for further refinement and matching. As discussed in the Section 5.2 on Unofficial Results for Stage One, this goal was achieved with resounding success.

	DRsus1	DRsur1	DRsum1	DRsds1	DRsdr1	DRsas1	DRsar1
At 5 docs	0.2560	0.2560	0.2560	0.2480	0.2480	0.2360	0.2400
At 10 docs	0.2400	0.2280	0.2300	0.2400	0.2320	0.2400	0.2300
At 20 docs	0.2240	0.2090	0.2140	0.2290	0.2210	0.2310	0.2160
At 30 docs	0.2073	0.1960	0.1993	0.2093	0.2033	0.2080	0.1927
At 100 docs	0.1306	0.1258	0.1254	0.1302	0.1244	0.1290	0.1160

Table 3: Stage 1 Routing, Document Level Average Precision

	DRwus1	DRwur1	DRwum1
At 5 docs	0.4000	0.4080	0.4000
At 10 docs	0.4080	0.4120	0.4080
At 20 docs	0.3890	0.3890	0.3890
At 30 docs	0.3573	0.3567	0.3560
At 100 docs	0.2672	0.2706	0.2682

Table 4: Stage 1 Ad Hoc, Document Level Average Precision

For Stage Two of the DR-LINK system, the official runs we submitted were different in many ways from those other systems produced. The input for the final matching was a relatively small set of documents selected for each topic (2838 documents on average) because Stage Two selected 81% as the recall-level prediction point. Another difference is that only 500 documents from the output for each topic was submitted for evaluation, as opposed to 1000 documents. Finally, due to the time constraints, the numbers of topics for routing and ad-hoc included in the official runs were 45 and 19, respectively.

Table 5 shows document level averages for the routing

case where S0, S1, S2, and S3 represent different scoring methods in the conceptual graph matching. S0 represents the basic scoring scheme explained in section 1.3. S1 is a scoring scheme designed to give higher scores those documents containing matched sub-graphs that are less fragmented. In other words, the more coherent the matching sub-graphs are with respect to the query graph, the higher score they get even though the number of matching nodes are the same. This was done by segmenting documents into several parts and computing the scores for them before they are combined. S2 and S3 are two different scoring schemes that take into account the maximum score obtained by a single sub-graph match which was averaged out in the case S1.

	S0	S1	S2	S3
At 5 docs	0.2800	0.2978	0.3111	0.2933
At 10 docs	0.2444	0.2489	0.2733	0.2822
At 20 docs	0.2222	0.2078	0.2222	0.2222
At 30 docs	0.1837	0.1852	0.1904	0.1933
At 100 docs	0.1287	0.1322	0.1309	0.1311

Table 5: Stage 2 Routing, Document Level Average Precision

Table 6 shows a set of corresponding values when RIT codes were used in the representation. Although the values in Table 6 are lower than those in Table 5 in general, it should be noted that for some topics (e.g. 53, 72, and 96), the average precision values obtained with RIT codes were significantly higher than those without RIT codes. This indicates that usefulness of RIT codes depends on the characteristics of topic statements.

	S0	S1	S2	S3
At 5 docs	0.2605	0.2698	0.2698	0.2698
At 10 docs	0.2395	0.2465	0.2465	0.2419
At 20 docs	0.1977	0.2151	0.2360	0.2377
At 30 docs	0.1860	0.1891	0.2031	0.2008
At 100 docs	0.1200	0.1240	0.1226	0.1221

Table 6: Stage 2 Routing, Document Level Average Precision with RIT

Because of the fact that only a subset of the entire databases (i.e. a subset of relevant documents) was used for each query and that only 500 documents were submitted for evaluation, we feel that the document level averages are better indicators of the system performance than the recall level averages. Table 7 shows the results for the Stage Two ad-hoc case without RIT codes.

	S0	S1	S2	S3
At 5 docs	0.4211	0.3263	0.4211	0.4105
At 10 docs	0.3579	0.3211	0.3737	0.3842
At 20 docs	0.3132	0.3000	0.3368	0.3553
At 30 docs	0.2807	0.2772	0.3140	0.3175
At 100 docs	0.2189	0.2116	0.2311	0.2316

Table 7: Stage 2 Adhoc, Document Level Average Precision

While there was about 8% average precision difference between S0 and S3 in our internal experiments with a smaller test collection, the differences among the four scoring schemes are not very apparent in the sets of top ranked documents as shown in the table.

5.2 Unofficial Results

There is an additional functionality exhibited by Stage One of DR-LINK which is not measured by the official precision and recall formulas. That is, the Recall Predictor can successfully apply a multiple regression-based cut-off criterion formula to the ranked list of relevant documents produced by Stage One to provide a set of documents which very accurately reflects a selected level of recall. As seen in Table 8, the baseline run (DRwus1) in which no cut-off prediction is made, post hoc evaluation of the documents judged relevant by the assessors, shows that all the relevant documents were contained in the top 29% of the ranked list, as averaged across the 50 Topic Statements. By comparison, on DRwur1, which uses the regression-based formula as the criterion to predict the cut-off for 100% recall, a 40% improvement in the portion of the database which is filtered out is achieved, as only 17% needs to be processed by Stage Two. Even more importantly, this system-predicted set of documents contains 97% of the relevant documents. Using the modified regression formula, DRwum1, 99% recall is achieved and only 23% of the database needs to be further processed, a 22% improvement over the baseline.

While these are average results across 50 queries, what should be remembered is the wide range amongst queries and the Recall Predictor's ability to provide **query specific** results. For example, for one Topic Statement, the regression formula selects only 63 documents for further processing, while for another Topic Statement, the formula selects 16,000. In each of these instances, 100% recall is achieved.

Baseline	Recall	%DB Searched	
DRwus1	1.0000	29.32	
Run	Actual Recall at Predicted 1.00 Recall	Average %DB Searched	%Change in %DB Searched from the Baseline
DRwur1	0.9670	17.50	-40.31
DRwum1	0.9864	22.77	-22.33

Table 8: Recall Prediction, Adhoc

Table 9 presents the same type of result for the routing queries. As noted in the Official Results section, the fact that the regression formula was trained on another corpus than the one it was tested on in the routing situation, produced somewhat poorer results. The actual recall achieved at the 100% predicted recall level is lower than on the adhoc queries, but the savings in the percentage (45%, 49%, 56%) of the database which needs to be further processed is higher. The lower actual recall level performance can be corrected by simply

recomputing the regression formula with training data from the appropriate corpus.

As discussed earlier, Stage Two did not elect to use the document set which Stage One predicted would contain

Baseline	Recall	%DB Searched	
DRsus1	1.0000	14.65	
Run	Actual Recall at Predicted 1.00 Recall	Average %DB Searched	%Change in %DB Searched from the Baseline
DRsur1	0.8497	6.45	-55.97
DRsum1	0.9032	7.99	-45.46
DRsdr1	0.8605	7.38	-49.62

Table 9: Recall Prediction, Routing

100% recall level , but instead chose a recall level of 81% across all queries. This recall level was based on results from the 18th month runs which showed that the average number of documents per query at the 81% recall level was 2000. Given Stage Two estimates of how long it would take to run the CG Matcher, an across-the-board 81% recall level was selected. However, an a priori selection based on average number of documents, rather than adaptive selection based on the recall predictor is not appropriate, given the fact that the number of documents which should be processed by Stage Two is known to vary a great deal by query and this number is predicted prior to Stage Two processing.

Run	Actual Recall at Predicted 0.81 Recall	Average %DB Passed to the 2nd Stage	Median #Docs Passed to the 2nd Stage
ADHOC DRwur1	0.8655	5.86	3,234
ROUTING DRsur1	0.6459	3.82	1,485

Table 10: Recall Prediction Statistics of Stage One at the 81% Recall Level

However, given Stage Two's selection of 81% as the desired recall level, Table 10 shows that for the Adhoc queries, Stage One actually passed a full 86% of the relevant documents to Stage Two. While for the Routing Queries, 65% of the relevant documents were contained in the set selected by Stage Two. The lower performance for routing again reflects the fact that the regression formula was trained on a different corpus.

5.3 Interpretation of Results

Stage One of DR-LINK provided very reasonable results, given that its function was not intended to be a stand-alone retrieval system, but was intended to be a filter for Stage Two, whose task was to improve precision by adding a finer level of matching using Conceptual Graphs. Given that, the precision achieved by Stage One based on the use of new and specialized types of linguistic evidence, is very promising. Additionally, it should be remembered that Stage One fared poorly on those Topic Statements in which single words mattered (e.g. cancer), since Stage One did not have any means for explicit matching of single words. Development efforts since the TIPSTER 24 month evaluation have greatly improved the performance of Stage One as a stand-alone retrieval system.

A detailed analysis of the Topic Statements on which

Stage One performed well, using the TIPSTER-provided data showing query by query average precision across systems, has shed some light on where Stage One is exhibiting real achievement and promise. Twelve of the twenty-six Topic Statements on which Stage One performed well, were ones in which Text Structure representations were able to place relevant documents higher in the rankings because the information being sought needed to possess a particular discourse-level attribute. This means that not only did information on an entity, person, or event need to be contained in a relevant document, but the information needed to match on another dimension or requirement. This might be, for example, that an event was pending, in which case the temporal relations encoded in the Text Structurer ranked more highly those documents in which the event was mentioned in a FUTURE component of text. Or, the Topic Statement might require information on the impact or outcome of an event. The Text Structure would, in response, rank more highly those documents in which the CONSEQUENCE or EXPECTATION relation was attached to the type of event sought.

An additional twelve Topic Statements on which Stage One performed well, were ones in which a 'not' was included in a subtle way. The Natural Language Query Constructor was able to interpret these logical requirements correctly in its construction of the query which was then matched against the inverted document file.

Given the errors and the incomplete nature of the various system modules as explained in section 4, the relatively low precision values in the tables for Stage Two are not surprising. Nonetheless, when the numbers are interpreted, there are factors that need to be taken into account. First, since Stage Two selected a Recall level of only 81% as its input from Stage One matching, and only top 500, as opposed to 1000, documents were submitted as official runs, smaller portion of all relevant documents were included in the output. In other words, an emphasis needs to be placed in precision of highly ranked documents rather than recall for the Stage Two. The average precision over all relevant documents and the exact R-precision are not necessarily good indicators of any systems whose major role is to enhance precision for the top portion of retrieved documents.

Second, because of the differences in the document databases, the number of topics (45 and 19), and the number of output documents submitted, the numbers in the tables are meaningful only for the purpose of comparing different strategies within the same system and thus are not to be compared with other systems' performances. The Table 5 and 7 show that even with the exactly same representation and retrieval methods, the results obtained from different databases and topics are radically different.

6. Future Work

There are two types of ongoing efforts to improve Stage One of DR-LINK. One effort is to extend the symbolic level of natural language processing developed during Phase I of TIPSTER by such efforts as the extension of the subject code vector algorithm into the multi-lingual environment (Liddy et al, 1993a); the development of new corpus analysis techniques for the expansion of complex nominals into their synonymous phrasings, and; the automatic construction of proper noun knowledge bases over time and multiple texts, for the purpose of cross-document inferencing in response to complex queries requiring extensive relational information (Paik, In press).

The second effort actually constitutes a major paradigm shift in the type of NLP which is used in Stage One. We are exploring ways to exploit the large amount of linguistically well-motivated, tagged text that is currently available for the various modules in the DR-LINK System as the necessary training data for the development of more adaptive techniques for NLP-based information retrieval. The research has commenced with an investigation into whether the functioning of each of the five modules in Stage One can be achieved via either probabilistic or neural-net linguistic processing. The distinctive aspect of our proposed use of the probabilistic approach is that it will use probabilities based on multiple levels of linguistic features.

For Stage Two, we have embarked on the work of increasing the number of relation node matches and thus maximizing the value of the representation. We not only need to correct the errors mentioned above but also need to incorporate the following strategies: 1) merging sentential conceptual graphs with the common concept nodes, which will benefit from anaphora resolution; 2) clean-up and refinement of knowledge bases; 3) extraction of "high-level" relations from a sub-graph to reduce the complexity of the conceptual graphs. We believe that the work to date provides a strong basis for accomplishing the tasks.

In addition to the works described in section 4, we are currently in the process of better understanding the semantic restrictions in the topic statements and how they can be translated into "high-level" relations. There

are at least two major advantages: this effort will both increase the value of the current representation and matching methods and reduce the complexity of the conceptual graph representation and hence matching. The second advantage is particularly important in that it will allow us to run many experiments to evaluate various features of the system, which were not possible in Phase I.

In conclusion, we believe that DR-LINK, Phase I has been a successful foray into the use of very rich representations of linguistic information for the retrieval of documents which match the requirements of queries on a wide range of the dimensions on which users' needs can be expressed. The innovative use of Conceptual Graphs shows potential for addressing many difficult issues in information retrieval research.

Acknowledgments

We would like to thank Ken McVearry and Bob DelZoppo, the software engineers from our subcontractor, Coherent Research Inc., and the following graduate students who worked on the research and development of DR-LINK: Margot Clark, Saleh Elmohamed, Chris Khoo, Tracey Lemon, Ming Li, Slawomir Marchinkowski, Mary McKenna, Woojin Paik, Stone Shih, Joe Woelfel, Edmund Yu, and Ahsan Zia.

References

Liddy, E.D. (In press). Development and implementation of a discourse model for newspaper texts. Proceedings of the Dagstuhl on Summarizing Text for Intelligent Communication. Saarbruken, Germany.

Liddy, E.D., Paik, W. & Woelfel, J. (1992). Use of subject field codes from a machine-readable dictionary for automatic classification of documents. Proceedings of the 3rd ASIS Classification Workshop. pp. 83-100.

Liddy, E.D., Paik, W. & Yu, E.S. (1993a). Conceptual Interlingua for Document Representation. Proposal submitted to National Science Foundation.

Liddy, E.D., Paik, W. & Yu, E.S. (1993b). Document filtering using semantic information from a machine readable dictionary. Proceedings of the ACL Workshop on Very Large Corpora.

Meteer, M., Schwartz, R. & Weischedel, R. (1991). POST: Using probabilities in language processing. Proceedings of the Twelfth International Conference on Artificial Intelligence. Sydney, Australia.

Myaeng, S. H. & Khoo, C. (1993). On Uncertainty Handling in Plausible Reasoning with Conceptual Graphs In Conceptual Structures: Theory and Implementation. H. D. Pfeiffer & T. E. Nagle, Springer-Verlag, 1993. An earlier version appears in the Proceedings of the 7th Conceptual Graphs Workshop, Las Cruces, NM.

Myaeng, S. H. & Liddy, E. D. (1993) Information Retrieval with Semantic Representation of Texts. In Proceedings of Symposium on Document Analysis and Information Retrieval. Las Vegas.

Paik, W. (In press). Chronological information extraction system. Proceedings of the Dagstuhl on Summarizing Text for Intelligent Communication. Saarbruken, Germany.

Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton, N.J. : Princeton University Press.

Sowa, J. (1984). Conceptual Structures: Information Processing in Mind and Machine. Reading, MA : Addison-Wesley.

IV. Published Papers

Liddy, E.D. (In press). Development and implementation of a discourse model for newspaper texts. Proceedings of the Dagstuhl on Summarizing Text for Intelligent Communication. Saarbruken, Germany.

Liddy, E.D. (1993). An alternative representation for documents and queries. Proceedings of the 14th National Online Meeting.

Liddy, E.D., McVearry, K., Paik, W., Yu, E.S. & McKenna, M. (1993). Development, implementation & Testing of a Discourse Model for Newspaper Texts. Proceedings of the ARPA Workshop on Human Language Technology, Princeton, NJ, March 21-24, 1993.

Liddy, E.D. & Myaeng, S.H. (In press). DR-LINK: A system update for TREC-2. Proceedings of Second Text Retrieval Conference (TREC-2). National Institute of Standards and Technology.

Liddy, E.D. & Myaeng, S. H. (1993). DR-LINK's linguistic-conceptual approach to document detection. Proceedings of First Text Retrieval Conference (TREC-1). National Institute of Standards and Technology.

Liddy, E.D. & Myaeng, S.H. (1992). DR-LINK Project Description. SIGIR Forum.

Liddy, E.D., Paik, W. & Yu, E.S. (1993). Document filtering using semantic information from a machine readable dictionary. Proceedings of the ACL Workshop on Very Large Corpora.

Liddy, E.D., Paik, W., Yu, E.S. & McVearry, K. (1993). An overview of DR-LINK and its approach to document filtering. Proceedings of the ARPA Workshop on Human Language Technology, Princeton, NJ, March 21-24, 1993.

Liddy, E.D., Paik, W. & Woelfel, J. (1992). Use of subject field codes from a machine-readable dictionary for automatic classification of documents. Proceedings of 3rd ASIS Classification Research Workshop.

Liddy, E.D. & Paik, W. (1992). Use of multiple knowledge sources for word sense disambiguation. In Proceedings of the 2nd Pacific Rim International Conference on Artificial Intelligence. Seoul, Korea.

Liddy, E.D. & Paik, W. (1992). Statistically-guided word sense disambiguation. In Proceedings of AAAI Fall '92 Symposium on Probabilistic Approaches to Natural Language. Boston.

Liddy, E. D. & Paik, W. (1992). Automatic recognition of thematic roles and semantic relations in text using the maximum coincidence search technique. In Proceedings of Informatics 11 Conference, University of York, England.

Myaeng, S. H. & Khoo, C. (1993). On uncertainty handling in plausible reasoning with conceptual graphs. In Conceptual Structures: Theory and Implementation. H. D. Pfeiffer & T. E. Nagle, Springer-Verlag, 1993. An earlier version appears in the Proceedings of the 7th Conceptual Graphs Workshop, Las Cruces, NM.

Myaeng, S. H., Khoo, C., & Li, M. (1993). Linguistic processing of text for a large-scale conceptual information retrieval system. Technical Report, School of Information Studies.

Myaeng, S.H. & Liddy, E.D. (1993). Information retrieval with semantic representation of texts. Proceedings of the Second Annual Symposium on Document Analysis and Information Retrieval.

Myaeng, S. H. & Li, Ming (1992). Building term clusters by acquiring lexical semantics from a corpus. In Proceedings of the First International Conference in Information and Knowledge Management. Baltimore.

Myaeng, S. H. & Lopez-Lopez, A. (1992). Conceptual graph matching: A flexible algorithm and experiments. Journal of Experimental and Theoretical Artificial Intelligence. Vol. 4, 107-126. An earlier version appears in Proceedings of 6th Annual Workshop on Conceptual Graphs. July 11-13, 1991.

Myaeng, S. H. (1992). Using conceptual graphs for information retrieval: A framework for adequate representation and flexible inferencing. Proceedings of Symposium on Document Analysis and Information Retrieval. Las Vegas.

Paik, W. (In press). Chronological information extraction system. Proceedings of the Dagstuhl on Summarizing Text for Intelligent Communication. Saarbruken, Germany.

Paik, W., Liddy, E.D., Yu, E.S., McKenna, M. (1993). Interpreting proper nouns for information retrieval. Proceedings of the ARPA Workshop on Human Language Technology, Princeton, NJ, March 21-24, 1993.

Paik, W., Liddy, E.D., Yu, E.S. & McKenna, M. (1993). Categorizing and standardizing proper nouns for efficient information retrieval. Proceedings of the ACL Workshop on Acquisition of Lexical Knowledge from Text.