AN OVERVIEW OF DR-LINK AND ITS APPROACH TO DOCUMENT FILTERING

Elizabeth D. Liddy [1], *Woojin Paik* [1], *Edmund S. Yu* [2], *Kenneth A. McVearry* [3]

[1] School of Information Studies
Syracuse University
Syracuse, NY 13244

[2] College of Engineering and Computer Science
Syracuse University
Syracuse, NY 13244

[3] Coherent Research, Inc.
1 Adler Drive
East Syracuse, NY 13057

1. MOTIVATION

DR-LINK is an information retrieval system, complex in design and processing, with the potential for providing significant advances in retrieval results due to the range and richness of semantic representation done by the various modules in the system. By using a full continuum of linguistic-conceptual processing, DR-LINK has the capability of producing documents which precisely match users' needs. Each of DR-LINK's six processing modules add to the conceptual enhancement of the document and query representation by means of continual semantic enrichments to the text. Rich representations are essential to meet the retrieval requirements of complex information needs and to reduce the ambiguities associated with keyword-based retrieval. 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 for producing a textual representation which has been shaped by all these levels of linguistic processing.

A vital aspect of our approach which is evidenced in the various semantic enrichments (e.g. Subject Field Codes, proper noun categories, discourse components, concept-relation-concept triples, Conceptual Graphs) added to the basic text, is the real attention paid to representation at a deeper than surface level. That is, DR-LINK deals with lexical entities via conceptually-based linguistic processing. For example, complex nominals are interpreted as meaningful multi-word constituents because the combination of individual terms in complex nominals conveys quite different meanings than if the individual constituents were interpreted separately. In addition, verbs are represented by case-frames so that other lexical entities in the sentence which perform particular semantic roles in relation to the verb are represented according to these semantic roles. Also, the rich semantic data (e.g. location, purpose, nationality) that are conveyed in the appositional phrases typically accompanying proper nouns, are represented in such a way that the semantic relations implicitly conveyed in the appositions are explicitly available for more refined representation and matching.

2. OVERVIEW

DR-LINK's system architecture is modular in design, with six processing modules, each of which enhance the document and query representation in terms of continual semantic enrichments to the text. Briefly overviewed, the system's six modules function as follows:

1. The Subject Field Coder uses semantic word knowledge to produce a summary-level topical vector representation of a document's contents that is matched to a vector representation of a topic statement in order to rank all documents for subject-based similarity to a query. All of the documents with their Subject Field Code vectors are passed to:

2. The Proper Noun Interpreter, which uses a variety of knowledge bases and context heuristics to categorize every proper noun in the text. The similarity between a query's proper noun requirements and each document's Proper Noun Field is evaluated and combined with the similarity value from the Subject Field Coder for a reranking of all documents in response to the query. Those documents with a mathematically determined potential for being relevant to the query are then passed to:

3. The Text Structurer, which sub-divides a text into its discourse-level segments in order to focus query matching to the appropriate discourse component in response to particular types of information needs. All of the structured texts, with the appropriate components weighted, are passed to:

4. The Relation-Concept Detector, whose purpose is to raise the level at which we do matching from a key-word or key-phrase level to a more conceptual level by expanding terms in the topic statement to all terms which have been shown to be 'substitutable' for them. Then, semantic relations between concepts are recognized in both documents and topic statements using separate handlers for the various parts of speech. This module produces concept-relation-concept triples which are passed to:

5. The Conceptual Graph Generator which converts these triples into the CG formalism (Sowa, 1984), a variant of semantic networks in which arcs between nodes are coded for relations. The resultant CGs are passed to:

6. The Conceptual Graph Matcher, which measures the degree to which a particular topic statement CG and candidate document CGs share a common structure, and does a final ranking of the documents.

In combination, these six stages of processing produce textual representations that capture breadth and variety of semantic knowledge. However, since the Conceptual Graph generation and matching are so computationally expensive, we also take care to eliminate from further processing for each query, those documents which have no likelihood of being relevant to a well-specified query or query-profile.

3. DOCUMENT FILTERING WITHIN DR-LINK

The fact that information-intense government organizations receive thousands of documents daily with only a relatively small subset of them being of potential interest to any individual user suggests that the routing application of information retrieval can be approached as a filtering process, with the types and optimal number of filterings dependent on the desired granularity of filtering. Our research demonstrates how a first, rough-cut, purely content-based document filter can be used to produce its appropriate preliminary ranking of an incoming flow of documents for each user. Using the similarity values produced by the SFC Filter, later system modules further refine the ranking and perform finer levels of analysis and matching.

The success of our filtering approach is attributable to the representation scheme we use for all texts, both documents and queries. The Subject Field Codes (SFCs) are based on a culturally validated semantic coding scheme developed for use in Longman's Dictionary of Contemporary English (LDOCE), a general purpose dictionary. Operationally, our system tags each word in a document with the appropriate SFC from the dictionary. The within-document SFC frequencies are normalized and each document is represented as a frequency-weighted, fixed-length vector of the SFCs occurring in that document (see Figure 1). For routing, queries are likewise represented as SFC vectors. The system matches each query SFC vector to the SFC vector of all incoming documents, which are then ranked on the basis of their vectors' similarity to the query. Those documents whose SFC vectors exceed a predetermined criterion of similarity to the query SFC vector can be displayed to the user immediately or passed on to the Proper Noun Interpreter for further processing and a second-level re-ranking.

The real merit of the SFC vectors is that they represent texts at a more abstract, conceptual level than the individual words in the natural language texts themselves, thereby addressing the dual problems of synonymy and polysemy. On the one hand, the use of SFCs takes care of the "synonymous phrasing" problem by representing text at a level above the word-level by the assignment of one SFC from amongst 124 possible codes to each word in the document. This means that if four synonymous terms were used within a text, our system would assign each of them the same SFC since they share a common domain which would be reflected by their sharing a common SFC. For example, several documents that discuss the effects of recent political movements on legislation regarding civil rights would have similar SFC vector representations even though the vocabulary choices of the individual authors might be quite varied. Even more importantly, if a user who is seeking documents on this same topic expresses her

A U. S. magistrate in Florida ordered Carlos Lehder Rivas, described as among the world's leading cocaine traffickers, held without bond on 11 drug-smuggling counts. Lehder, who was captured last week in Colombia and immediately extradited to the U.S., pleaded innocent to the charges in federal court in Jacksonville.

LAW	.2667	SOCIOLOGY	.1333
BUSINESS	.1333	ECONOMICS	.0667
DRUGS	.1333	MILITARY	.0667
POLITICAL SCIENCE	.1333	OCCUPATIONS	.0667

Fig. 1: Sample Wall Street Journal document and its SFC representation

information need in terms which do not match the vocabulary of any of the documents, her query will still show high similarity to these documents' representations because both the query's representation and the documents' representations are at the more abstract, semantic-field level and the distribution of SFCs on the vectors of the query and the relevant documents would be proportionately similar across the SFCs.

The other problem with natural language as a representation alternative that has plagued its use in information retrieval is polysemy, the ability of a single word to have multiple senses or meanings. Our SFCoder uses psycholinguistically-justified sense disambiguation procedures (Liddy & Paik, 1992) to select a single sense for each word. Ambiguity is a serious problem, particularly in regard to the most frequently used lexical items. According to Gentner (1981) the twenty most frequent nouns in English have an average of 7.3 senses each, while the twenty most frequent verbs have an average of 12.4 senses each. Since a particular word may function as more than one part of speech and each word may also have more than one sense, each of these entries and/or senses may be assigned different SFCs. This is a slight variant of the standard disambiguation problem, which has shown itself to be nearly intractable for most NLP applications, but which is successfully handled in DR-LINK, thereby allowing the system to produce semantically accurate SFC vectors.

We based our computational approach to successful disambiguation on current psycholinguistic research literature which we interpret as suggesting that there are three potential sources of influence on the human disambiguation process: 1) local context, 2) domain knowledge, and 3) frequency data. We have computationally approximated these three knowledge sources in our disambiguator. The disambiguation procedures were tested by having the system select a single SFC for each word. These SFCs were compared to the sense-selections made by an independent judge. The disambiguation implementation selected the correct SFC 89% of the time. This means that a word such as 'drugs', which might refer to either medically prescribed remedies or illegal intoxicants that are traded on the street would be represented by different SFCs based on the context in which it occurred.

4. PROCESSING IN THE SUBJECT FIELD CODER

In the Subject Field Coder, the following stages of processing are done:

In **Stage 1** processing, we run the documents and query through a probabilistic part of speech tagger (Meteer et al, 1991) in order to restrict candidate SFCs of a word to those of the appropriate syntactic category.

Stage 2 processing retrieves SFCs of each word's correct part of speech from the lexical database and assigns the SFCs.

Stage 3 then uses an ordered set of sentence-level context-heuristics to determine a word's correct SFC if multiple SFCs have been assigned to a word's different senses. First, the SFCs attached to all words in a sentence are evaluated to determine at the sentence level whether any words have only one SFC assigned to all their senses in LDOCE (unique-SFC), and; secondly, the SFCs which are assigned to more than three words in the sentence (frequent-SFC).

Stage 4 scans the SFCs of each remaining word to determine whether the unique-SFCs or frequent-SFCs discovered in Stage 3 occur amongst the multiple SFCs assigned by LDOCE to the ambiguous word. Those ambiguous words which have no SFC in common with the unique-SFCs or frequent-SFCs for that sentence are passed on to the next stage.

Stage 5 incorporates two global knowledge sources to complete the sense disambiguation task. The primary source is a correlation matrix which reflects stable estimates of SFC co-occurrences within documents. The second source is the order in which the senses of a word are listed in LDOCE which is based on frequency of use in the English language. In Stage 5, each of the remaining ambiguous words is resolved a word at a time, accessing the matrix via the unique and most frequent-SFCs of the sentence. The system evaluates the correlation coefficients between the unique and most frequent-SFCs of the sentence and the multiple SFCs assigned to the word being disambiguated to determine which of the multiple SFCs has the highest correlation with a unique-SFC or frequent-SFC. The system then selects that SFC as the unambiguous representation of the sense of the word.

Stage 6 processing produces a vector of SFCs and their frequencies for each document and for the query.

Stage 7 normalizes the vectors of each text, and at:

Stage 8, the document vectors are compared to the query vector using a similarity measure. A ranked listing of the documents in decreasing order of similarity is produced.

The assignment of SFCs is fully automatic and does not require any human intervention. In addition, this level of semantic representation of texts is efficient and has been empirically tested as a reasonable approach for ranking documents from a very large incoming flux of documents. For the 18th month TIPSTER evaluation, the use of this representation allowed the system to quickly rank 60 megabytes of text in the routing situation that was tested. All the later-determined relevant documents were within the top 37% of the ranked documents produced by the SFC Module.

A second level of lexical-semantic processing further improves the performance of DR-LINK as a reasonable document filter. That is, the Proper Noun Interpreter (Paik et al; this volume) computes the similarity between a query's proper noun requirements and each document's Proper Noun Field and combines this value with the similarity value produced by the SFCoder for a reranking in relation to the query. In the 18th month testing of our system, the results of this reranking based on the SFC values and the Proper Noun values placed <u>all</u> the relevant documents within the top 28% of the database.

5. DOCUMENT CLUSTERING USING SUBJECT FIELD CODES

These summary-level semantic vector representations of each text's contents produced by the SFCoder have also proven useful as a means for dividing a database into clusters of documents pertaining to the same subject area. The SFC vectors are clustered using Ward's agglomerative clustering algorithm (Ward, 1963) to form classes in the document database. Ad hoc queries are represented as SFC vectors and matched to the centroid SFC vector of each cluster in the database. Clusters whose centroid SFC vector exhibit high similarity to the query SFC vector can then be browsed by users who do not have a fully specified query, but who prefer to browse groups of documents whose optimum content they can only loosely define to the system (Liddy, Paik, & Woelfel, 1992).

A qualitative analysis revealed that clustering SFC vectors using Ward's clustering algorithm resulted in meaningful groupings of documents that were similar across concepts not directly encoded in SFCs. Two examples: all of the documents about AIDS clustered together, although AIDS is not in LDOCE. Secondly, all of the documents about the hostages in Iran clustered together even though proper nouns are not included in LDOCE and the word 'hostage' is tagged with the same SFC as hundreds of other terms. What the SFC representation of documents accomplishes, is that documents about the same or very similar topics have relatively equal distributions of words with the same SFCs and will therefore cluster together in meaningful groups.

6. CONCLUSION

Our implementation and testings of the SFCoder as a means for semantically representing the content of texts, either for the purpose of ranking a document set according to likelihood of being relevant to an individual query or for producing conceptually related clusters of documents for browsing are very promising. Particularly worthy of note is the observation that in a large operational system, the ability to filter out an average of 72% of the incoming flux of millions of documents will have a significant impact on any document detection system's performance with which this semantic-based document filter is combined.

REFERENCES

Gentner, D. (1981). Some interesting differences between verbs and nouns. Cognition and brain theory. 4(2), 161-178.

Liddy, E.D., McVearry, K.A., Paik, W., Yu, E.S. & McKenna, M; (In press). Development, implementation & testing of a discourse model for newspaper texts. Proceedings of the Human Language Technology Workshop. Princeton, NJ: March, 1993.

Liddy, E.D. & Paik, W. (1992). Statistically-guided word sense disambiguation. In Proceedings of AAAI Fall Symposium Series: Probabilistic approaches to natural language. Menlo Park, CA: AAAI.

Liddy, E.D., Paik, W. & Woelfel, J.K. (1992). Use of subject field codes from a machine-readable dictionary for automatic classification of documents. Advances in Classification Research: Proceedings of the 3rd ASIS SIG/CR Classification Research Workshop. Medford, NJ: Learned Information, Inc.

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

Paik, W., Liddy, E.D., Yu, E.S. & McKenna, M. (In press). Interpretation of Proper Nouns for Information Retrieval. Proceedings of the Human Language Technology Workshop. Princeton, NJ: March, 1993.

Sowa, J. (1984). Conceptual structures: Information processing in mind and machine. Reading, MA: Addison-Wesley.

Ward, J. (1963). Hierarchical grouping to optimize an objection function. Journal of the American Statistical Association. 58, p. 237-254.