Advanced Natural Language Processing Tools for Web Information Retrieval, Content Analysis, and Synthesis.

Final Scientific / Technical Report

Advanced Natural Language Processing Tools for Web Information Retrieval, Content Analysis and Synthesis.

This SBIR was granted in response to topic **41** and subtopic **a**:

41. DISCOVERY, SEARCH, AND COMMUNICATION OF TEXTUAL KNOWLEDGE RESOURCES IN DISTRIBUTED SYSTEMS a. Discovering and Utilizing Knowledge Sources for Metasearch Knowledge Systems

Recipient: Edgewater Technology Associates, Inc.

ID Number: DE-FG02-07ER84704

Principal Investigator: Antonio Zamora

Advanced Natural Language Processing Tools for Web Information Retrieval, Content Analysis, and Synthesis.

Executive Summary

The goal of this SBIR was to implement and evaluate several advanced Natural Language Processing (NLP) tools and techniques to enhance the precision and relevance of search results by analyzing and augmenting search queries and by helping to organize the search output obtained from heterogeneous databases and web pages containing textual information of interest to DOE and the scientific-technical user communities in general.

The SBIR investigated 1) the incorporation of spelling checkers in search applications, 2) identification of significant phrases and concepts using a combination of linguistic and statistical techniques, and 3) enhancement of the query interface and search retrieval results through the use of semantic resources, such as thesauri.

A search program with a flexible query interface was developed to search reference databases with the objective of enhancing search results from web queries or queries of specialized search systems such as DOE's Information Bridge. The DOE ETDE/INIS Joint Thesaurus was processed to create a searchable database. Term frequencies and term co-occurrences were used to enhance the web information retrieval by providing algorithmically-derived objective criteria to organize relevant documents into clusters containing significant terms. A thesaurus provides an authoritative overview and classification of a field of knowledge. By organizing the results of a search using the thesaurus terminology, the output is more meaningful than when the results are just organized based on the terms that co-occur in the retrieved documents, some of which may not be significant.

An attempt was made to take advantage of the hierarchy provided by broader and narrower terms, as well as other field-specific information in the thesauri. The search program uses linguistic morphological routines to find relevant entries regardless of whether terms are stored in singular or plural form. Implementation of additional inflectional morphology processes for verbs can enhance retrieval further, but this has to be balanced by the possibility of broadening the results too much. In addition to the DOE energy thesaurus, other sources of specialized organized knowledge such as the Medical Subject Headings (MeSH), the Unified Medical Language System (UMLS), and Wikipedia were investigated.

The supporting role of the NLP thesaurus search program was enhanced by incorporating spelling aid and a part-of-speech tagger to cope with misspellings in the queries and to determine the grammatical roles of the query words and identify nouns for special processing. To improve precision, multiple modes of searching were implemented including Boolean operators, and field-specific searches. Programs to convert a thesaurus or reference file into searchable support files can be deployed easily, and the resulting files are immediately searchable to produce relevance-ranked results with built-in spelling aid, morphological processing, and advanced search logic. Demonstration systems were built for several databases, including the DOE energy thesaurus.

Advanced Natural Language Processing Tools for Web Information Retrieval, Content Analysis, and Synthesis.

Comparison of actual accomplishments and goals of the project

Phase I made significant progress toward accomplishing the goals of the SBIR. The incorporation of a spelling checker in a search application was implemented by modularizing an existing spelling checker to facilitate its installation in any interactive user environment. A mechanism for combining a set of authoritative dictionaries with a dictionary derived from the target database made it possible to prevent suggestions for common, properly spelled words, while at the same time allowing the suggestion of spelling aid candidates from the specific database.

Interfacing a phrase parser in the query interface allowed the identification of parts of speech and significant phrases in both the query and the reference databases. The phrases identified by this procedure enhanced the clustering process used to organize search results. There was not enough time to evaluate the recall and precision performance of derivational morphology variants for search-result clustering enhancement.

A third goal of the SBIR was to enhance the retrieval of information obtained from reference semantic resources such as thesauri by broadening a search through the application of inflectional morphology transformations for pertinent search terms, and by enhancing the precision of the search using techniques such as Boolean logic, limiting results to exact matches, and specifying terms in specific fields of a reference document.

Summary of project activities during the period of funding

Background. The performance of this SBIR took advantage of previously existing technology to focus attention on new developments to advance the state of the art. Two existing metasearch engines, Polymeta and AllPlus, were used as the test platforms for integration of the Natural Language Processing support programs. In addition, separate test drivers were used to perform preliminary tests of the linguistic components prior to the integration tests. Existing databases and thesauri for the health and energy areas were used to build the reference knowledge bases used by the clustering metasearch engines. The health databases included the Medical Subject Heading (MeSH) file and the Unified Medical Language System (UMLS) from the National Library of Medicine. The energy database consisted of the ETDE/INIS Joint Reference Thesaurus provided by the Department of Energy.

Many of the activities summarized here were carried out concurrently. This list does not imply the sequence in which the activities were conducted.

Spelling Aid. Although spelling aid has been a component of office applications for almost three decades, the algorithms used for interactive searching are very closely guarded secrets due to the competitiveness of the field. The new work focused on development of techniques to provide spelling aid for unrestricted World Wide Web queries. Referencing a standard dictionary, as is the case in office applications, is not enough since the vocabulary of the web is virtually unlimited and it contains foreign words, jargon, and technical terms that are beyond the scope of ordinary dictionaries.

Advanced Natural Language Processing Tools for Web Information Retrieval, Content Analysis, and Synthesis.

The work for this SBIR had to discover new techniques for using multiple dictionaries to identify correctly spelled words, query terms similar to a word in the dictionary, and unrecognizable words. Algorithms had to be developed to optimize the suggestion process. It was necessary to maximize the percentage of correct spelling aid candidates, while avoiding the suggestion of incorrect candidates. To achieve this, subroutines were devised to identify words for which no candidates were suggested. Metrics to determine whether a suggestion was plausible involved the use of traditional string distance measures, phonetic similarities, and proprietary algorithms. Integration of spelling aid with the metasearch engine required developing Java interfaces for the subroutines using re-entrant code to support any number of users.

Text Analysis Procedures. Functions like part-of-speech tagging and phrase recognition require linguistic dictionaries and grammatical rules. In addition, since the text requires complex manipulation in different stages, it is necessary to maintain a data structure with the capability to store grammatical tags, phrase structure tags, and any other relevant information. The programs use a bi-directional list structure with associated attribute/value pairs that may contain any desired information..

The linguistic dictionaries share a common structure with the spelling aid dictionaries. In this way, the number of external files can be minimized for complex applications that require spelling aid and grammatical analysis. Access to specialized functions, like inflectional morphology, is controlled by the parameters provided to the interface routines.

Knowledge Base Searching. In order to implement retrieval of singular and plural variants, this SBIR integrated a phrase tagger interface within the search query analysis. Plurals (or singulars) for words that were tagged as nouns were added as supplemental search terms to the search query. For example, a query for "biomass fuel" which might not retrieve any entries from the energy thesaurus would be supplemented so that "biomass fuel" or "biomass fuels" could match. The supplemented query is equivalent to the Boolean query "biomass & (fuel | fuels)" except that term adjacency constraints are applied to the ranking of results from queries with supplemental terms, but not to those from queries with Boolean logic. The ranking of results is used to guarantee the retrieval of best component matches for complex queries with multiple topics.

In order to increase the precision of the terms retrieved from a knowledge database, an "exact match" mode and a "field specification" mode were developed. These retrieval modes make it possible to specify that the string "biomass fuels" should match exactly, or that a query like "biomass[lx] fuels[lx]" should only match when the words occur within a field with tag "lx". Multiple fields can be specified within the brackets, and it is possible to combine the exact match mode and the field specification mode with the option to match singular/plural.

Knowledge Base Preparation. Substantial consideration was given to the format of the database for the knowledge bases. Many widely supported databases such as MySQL are relatively easy to implement and maintain. However, their disadvantage is that it is not

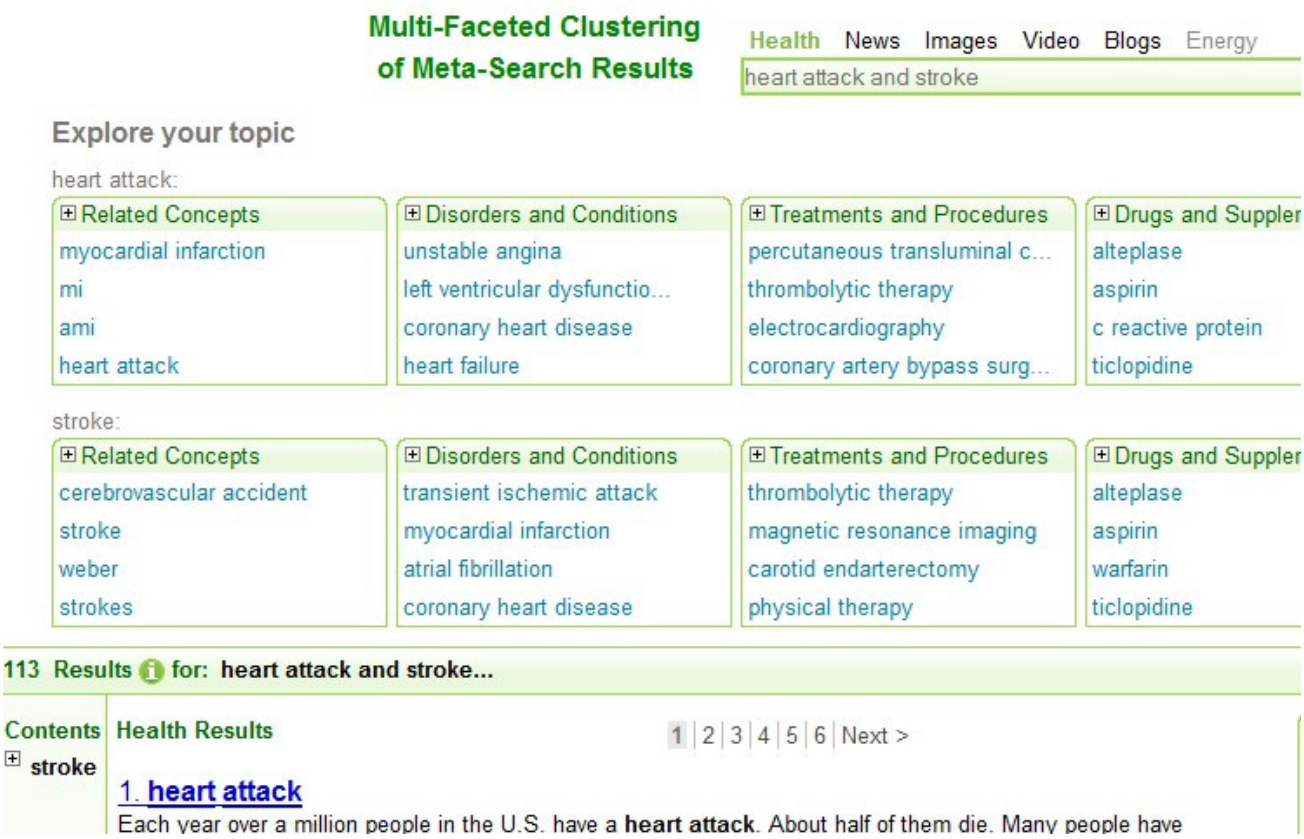
very easy to search any fields and rank the results in arbitrary ways. This SBIR developed a search program with great flexibility, but made it necessary to create customized databases. The database creation process is highly automated and makes it possible to convert any file format into a searchable database.

An initial step converts a source file, such as the MeSH index, or the Energy Thesaurus, into a SIL International Toolbox database format, which is a simple field-delimited file. The resulting file is processed by a tokenization program which creates an inverted index file for all the terms of the file. The resulting compressed index file contains all the information required to determine the physical proximity of the terms to each other. The search process is very fast because only the index is examined to determine co-occurrence of terms, number of terms in a document, and everything else that is needed to identify and rank the knowledge base entries that meet the specified search criteria. The process of establishing a new search database can be done relatively quickly once the initial source file is converted to the field-delimited format.

A demonstration of the functionality of these programs and of their integration with a clustering metasearch engine was presented to DOE on April 14, 2008.

Products developed and technology transfer activities

- Modular spelling checker that can be integrated in an interactive query interface.
- Protocol for automatically developing spelling aid dictionaries for any particular database.
- Software to generate inverted indexes to support high-performance searching of reference databases.
- Search program with customizable recall and precision using grammatical morphology, ranked output results, Boolean logic, and search field specification.
- Specialty Databases: MeSH, UMLS, Energy Thesaurus,



The figure above illustrates the application of the medical knowledge base to a query consisting of multiple concepts: "heart attack and stroke". The NLP components help to isolate the concepts and then retrieve related terms from the knowledge base. In this case, "heart attack" is associated with "myocardial infarction", "mi", and "coronary heart disease". The concept "stroke" is associated with "cerebrovascular accident" and "transient ischemic attack". These relationships were discovered from analysis of the query phrases when matched against the concepts in the knowledge base. The association process is fully automatic and takes advantage of decisions that have been made manually in the creation of the thesauri, as well as proprietary statistical correlations.