A KNOWLEDGE-BASED APPROACH TO INDEXING SCIENTIFIC TEXT

Lois Boggess and Julia Hodges, Principal Investigators

Department of Computer Science
Mississippi State University
Mississippi State, MS 39762

PROJECT GOALS

We are developing a system which scans articles from scientific literature for the purpose of indexing the text. That is, the system should assist in the rapid determination of the key topics and content of articles in a particular domain and in the production of brief phrases describing the content. The number of correct concepts generated should be 80% of the concepts present (a recall rate of 80%), as compared to the output of human document analysts processing the same material.

RECENT RESULTS

Our KUDZU system is an existing system designed for knowledge extraction from technical text, developed and tested primarily in the domain of veterinary medicine. The testbed for the current project consists of a large body of journal articles in the domain of physical chemistry, previously indexed by humans to relate to about 200 concepts taken from a hierarchy of more than 2000 concepts. Efforts for these first months have been in two directions: (1) understanding the relationships among the entities in the hierarchy provided by the domain experts; it is a tangled hierarchy — a graph rather than a tree; and (2) exploring the adaptation of clustering algorithms to generate data-driven hierarchies — hierarchies suggested by characteristics of the texts themselves.

The KUDZU system is designed to extract information from technical text in a single domain, yet many elements of the system are domain-independent. For example, although the tags with which it labels text include domain-dependent semantic components, most parts of the system are not affected by the actual values of those semantic components and do not have to be changed to accommodate a change of domain. Two exceptions are a) the portion of the system that attaches prepositional phrases to appropriate elements of sentences and b) the domain schema used by the knowledge analyzer of the system. The latter consists of a hierarchy of concepts (e.g., in the domain of veterinary medicine, heart is a body-part which is also a location), and a description of the relations deemed to be of interest, along with mandatory and optional concepts which fill roles in those

relations. We have access to a hierarchy of concepts used by domain experts to perform the indexing task. Part of the current effort is in coding the known required and optional elements of text used by the experts in determining whether a given index concept is present.

Until now we have used our own tag set, developed some years ago. We have just adopted a close correlate of the Penn Treebank tag set and have revised our tagger to work with bigrams and trigrams of any tag set.

PLANS FOR THE COMING YEAR

We will complete the computerization of (1) the hierarchy used by the domain experts to index text, (2) the mandatory and optional elements of text used by the experts to determine whether a given index concept is indeed present, and (3) a set of words known by the experts to map onto the index concepts. We will also be looking at automating the extraction of both kinds of information directly from a body of text of many hundreds of articles indexed by concept. Once the domain schema is modified, the KUDZU system will produce all relations from the text that the schema defines as being of interest. We expect at that time to analyze whether the set of concepts and the set of words mapping onto them are sufficient in the sense that they are the only ones whose presence in the text must be noted and given semantic labels. Additional analysis of the actual texts may be required to extend the set of concepts and words for the domain schema.

Once the revised system receives a body of text and produces a list of extracted relations from that text on the basis of prior descriptions of which relations are of interest, we will turn to the task of producing index phrases for that text. Most such phrases will probably not be identical to the index phrases that were actually generated by a document analyst. However, we hope within the year to have a set of generated index phrases that look to a non-expert as if most of them may be equivalent to the actual phrases that were produced by the human expert. At that point the output of the system will be given to the experts themselves for evaluation.