EXPLOITING CONCEPT SPACES FOR TEXT RETRIEVAL

Ellen M. Voorhees, Project Leader

Siemens Corporate Research, Inc.
Princeton, NJ 08540

PROJECT GOALS

The Learning Systems Department at Siemens Corporate Research is investigating the use of *concept spaces* to increase retrieval effectiveness. Similar to a semantic net, a concept space is a construct that defines the semantic relationships among ideas. The current focus of our research is to exploit the information in such a structure to ameliorate known shortcomings of statistical retrieval methods while maintaining the statistical methods' robustness. Our initial concept space is extracted from WordNet, a manually-constructed lexical database developed at Princeton University.

Our focus on statistical methods is a consequence of our goal to develop techniques that are applicable to matching texts in very large corpora. In particular, we impose two constraints on our research to meet this goal. First, we want to keep human intervention in the indexing and retrieval processes at a minimum, and so we use strictly automatic procedures. Second, since even automatic procedures need to be relatively efficient, and we believe this efficiency requirement precludes the use of deep analyses of document content for the foreseeable future, we restrict ourselves to shallow (statistical) processing of the text and concept space.

The specific problems we are addressing are the effects polysemy and synonymy have on retrieval performance. Polysemy depresses precision by causing false matches between texts, while synonymy depresses recall by causing true conceptual matches to be missed. We are investigating ways to index the content of text by the concepts of the concept space rather than the words that appear in the text, and thus avoid both polysemy and synonymy problems. With this approach, additional expense is incurred only during indexing — efficient concept-matching routines can be used for retrieval.

RECENT RESULTS

Our first experiments investigated the effectiveness of expanding a text's representation with words that are related to original text words in WordNet. The experimental evidence indicates that in the absence of a method to resolve word senses, expansion is almost always detrimental. For TREC-1, we were able to improve the performance of some queries using an expansion procedure that added only synonyms (as opposed to words related by other lexical relations) and required at least two original text words agree on the synonym (as a rudimentary check on the sense). However, this same procedure degrades the performance of other queries; overall performance is roughly comparable to the better statistical methods that do no special processing for synonyms.

Given the importance of sense resolution to workable expansion schemes, and the belief that polysemy is an important retrieval problem in its own right, we are currently studying automatic sense resolution procedures. In one approach, we use the nouns that co-occur within a text and the IS-A links within WordNet to select WordNet synonym sets as the senses for ambiguous nouns in the text. Retrieval performance degrades using this technique for two main reasons: the information inherent in the generalization/specialization hierarchy induced by the IS-A links is not sufficient to reliably select the correct sense of a noun from the set of fine distinctions in WordNet; and short query statements provide little context for disambiguation. In a separate approach, we are investigating the utility of classifiers that learn the contexts of the different senses of a given ambiguous word from training examples.

PLANS FOR THE COMING YEAR

Our research in the coming year will continue to focus on classifier-based sense resolution methods. We hope to improve the effectiveness of the classifiers by allowing them to learn syntactic templates that are indicative of a sense in addition to the more general context models they already learn. We must also integrate classifiers into the indexing phase of a retrieval system. Our intention is to annotate the synonym sets in the WordNet hierarchy with *sense vectors*, vectors that summarize the contexts in which members of the synonym set are likely to appear, and to select a sense based on the text's similarities to the sense vectors.