Session 5b. Information Retrieval

Donna Harman

National Institute of Standards and Technology
Gaithersburg, Md., 20899

1. Introduction

As this is the first time there has been a session on information retrieval at a DARPA Speech and Natural Language Workshop, it seems appropriate to provide a more detailed introduction to this topic than would normally appear. The term "information retrieval" refers to a particular application rather than a particular technique, with that application being the location of information in a (usually) large amount of (relatively) unstructured text. This could be done by constructing a filter to pull useful information from a continuous stream of text, such as in building an intelligent router or a library profiling system. Alternatively the text could be archived newspapers, online manuals, or electronic card catalogs, with the user constructing an ad-hoc query against this information. In both cases there needs to be accurate and complete location of information relevant to the ad-hoc query or filter, and efficient techniques capable of processing often huge amounts of incoming text or very large archives.

The currently-used Boolean retrieval systems grew out of the 100 or more year old practice of building cumulative indices, with early mechanical devices enabling people to join two index terms using AND's and OR's. This mechanism was adapted to computers and although today's commercial retrieval systems are much more sophisticated, they had not gone beyond the Boolean model. Boolean systems are difficult for naive or intermittent users to operate, and even skilled searchers find these systems limiting.

The widespread use of computers in the 1960's, and the availability of online text made possible some innovative and extensive research in new information retrieval techniques [5, 3]. This work has continued, with new models being proposed, many experimental techniques being tried, and some implementation and testing of these systems in real-world environments. For an excellent summary of various models and techniques, see [1], and for a discussion of implementation issues, see [2]. The major archival publications in the area of information retrieval are 1) Information Processing and Management, Pergamon Press; 2) Journal of the American Society for Information Science; and 3) the annual proceedings of the ACM SIGIR conference, available from ACM Press. text, with the goal being to match a user's query (or a filter) against the text in such a manner as to provide a ranked list of titles (or documents), with that rank based on the probability that a document is relevant to the query or filter. The use of statistical techniques rather than natural language techniques comes from the need to handle relatively large amounts of text, and the (supposed) lack-of-need to completely understand text in order to retrieve from it. For a survey of the use of natural language procedures in information retrieval, see [4].

The statistical techniques have proven successful in laboratories, and generally retrieve at least some relevant documents at high precision levels. The performance figure often quoted for these systems is 50% precision at 50% recall; roughly equivalent to the performance of Boolean systems used by skilled searchers. Unfortunately this performance has not seen major improvement recently, although improvements continue in related parts of information retrieval, such as interfaces, efficiency, etc. There are two explanations often given for this lack of improvement. The first is that the currently-available test collections are too small to allow proper performance of many of the proposed techniques, and second, that more sophisticated techniques are needed, including some natural language techniques.

The DARPA TIPSTER and TREC programs address both these issues, with a much larger test collection (4 gigabytes of text) being built, and a range of techniques, including sophisticated statistical techniques and efficient natural language techniques, being supported. Results from these projects will be reported in the future. The four papers in this session all apply natural language techniques to information retrieval, and illustrate some of the important ways that natural language processing can improve information retrieval.

2. Papers

The first paper, "Information Retrieval using Robust Natural Language Processing" by Tomek Strzalkowski of New York University, augments a basic statistical information retrieval system with various natural language components. One of these components is the replacement of the standard morphological stemmer with a dictionary-assisted stemmer, improving average precision by 6 to 8%, even in the small test collection being used. Additionally a very fast syntactic parser is used to derive certain types of phrases from the text. These phrases, in addition to the single terms, make for a richer representation of the text, and are also used to expand the queries. The query expansion involves finding similarity relationships between terms in these phrases, and then filtering these relationships to carefully select which terms to add to the query. This filtering (which adds only 1.5% of the possible relations) enables a performance improvement in average precision of over 13%, a significant result for this small test collection. The paper therefore addresses two of the major issues in information retrieval: improving accuracy (precision) using a better stemmer, and improving completeness (recall), without losing accuracy, by adding carefully selected terms to the query.

The second paper, "Feature Selection and Feature Extraction for Text Categorization" by David D. Lewis of the University of Chicago, deals with the problem of text categorization, or the assigning of texts to predefined categories using automated methods based on the text contents. Two particular areas are investigated. The first area involves finding appropriate statistical methods for assigning categories. Adaptions are made to a statistical model from text retrieval, and methods for determining actual category assignments rather than probability estimates are discussed. The second area of research examines various techniques for selecting the text features for use in this statistical method. Three types of features are tried: 1) single terms from the text, 2) simple noun phrases found using a stochastic class tagger and a simple noun phrase bracketing program, and 3) small clusters of features constructed using several methods. Additionally the effect of using smaller sets of all three types of features is investigated, and is shown to be more effective than using the full set. The problem of selecting which features of text to index is important in information retrieval, as often the terms in the queries are both inaccurate and insufficient for complete retrieval. By improving the indexing of the text, such as by adding selected phrases, clusters, or other features, these queries can be more successful. This work will continue with the larger test collections becoming available in the future.

The third paper, "Inferencing in Information Retrieval" by Alexa T. McCray of the National Library of Medicine, describes an information retrieval system being designed for the biomedical domain. This system takes advantage of the complex thesaurii built and maintained by the National Library of Medicine by making use of a metathesaurus and semantic network based on these thesaurii. The system uses a syntactic parser against the queries, related text, the metathesaurus, and an online dictionary to construct noun phrases that are grouped into concepts. It then attempts to match these concepts against documents that have not only some naturally-occurring text, but also manual indexing terms based on the thesaurii. The paper discusses the problems found in mapping the language of the queries to the language of the relevant documents, a major difficulty for all information retrieval systems. In this case, as opposed to the earlier papers, the features of the text that are indexed are fixed, and the issue is how to properly construct queries, or properly map natural language queries, into structures that will match the text features.

The fourth paper, "Classifying Texts using Relevancy Signatures" by Ellen Riloff and Wendy Lehnert of the University of Massachusetts, investigates feature selection for text classification, as did the second paper. The application here, however, is not how to route text into multiple predefined categories, but how to separate articles into only two sets: those relevant to a specific but complex topic, and those not relevant. This is used as a filtering or text skimming preprocessor to text extraction. The paper describes the design of an algorithm that will locate linguistic expressions that are reliable clues to document relevancy. These expressions are found by parsing the training set as input to the algorithm, and then automatically selecting the expressions or features that occur in the relevant and non-relevant documents. These features can then be used for later classification in new collections. As contrasted to the second paper, the techniques rely on analysis of the training collection to locate features, rather than on trying to identify more general methods of constructing features from the text.

References

1. Belkin N.J. and Croft W.B. (1987). Retrieval Techniques. In Williams, M. (Ed.), *Annual Review of Information Science and Technology* (pp. 109-145). New York, NY: Elsevier Science Publishers.

2. Harman D. and Candela G. (1990). Retrieving Records from a Gigabyte of Text on a Minicomputer using Statistical Ranking, *Journal of the American Society for Information Science*, 41(8), 581-589.

3. Salton G. (1971). *The SMART Retrieval System — Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, N.J.

4. Smeaton A.F. (1990). Natural Language Processing and Information Retrieval. Special Edition of *Information Processing and Management*, 26(1).

5. Stevens M.F. (1965). *Automatic Indexing: A State of the Art Report.* Monograph 91, National Bureau of Standards, Washington, D.C., March 1965.