Chapter 7

Document Processing

(Following section is taken from Chapter 7 “Document Processing”)
of the book: “Survey of the state of the art in human language technology”)

7.2 Document Retrieval

Donna Harman,[a] Peter Schäuble,[b] & Alan Smeaton[c]

[a] NIST, Gaithersburg, Maryland, USA
[b] ETH Zurich, Switzerland
[c] Dublin City University, Ireland, UK

Document retrieval is defined as the matching of some stated user query against useful parts of free-text records. These records could be any type of mainly unstructured text, such as bibliographic records, newspaper articles, or paragraphs in a manual. User queries could range from multi-sentence full descriptions of an information need to a few words, and the vast majority of retrieval systems currently in use range from simple Boolean systems through to systems using statistical or natural language processing. Figure 7.1 illustrates the manner in which documents are retrieved from various sources.

Several events have recently occurred that are having a major effect on research in this area. First, computer hardware is more capable of running sophisticated search algorithms against massive amounts of data, with acceptable response times. Second, Internet access, such as World Wide Web (WWW), brings new search requirements from untrained users who demand user-friendly, effective text searching systems. These two events have contributed to creating an interest in accelerating research to produce more effective search methodologies, including more use of natural language processing techniques.

There has been considerable research in the area of document retrieval for more than thirty years (Belkin & Croft, 1987), dominated by the use of statistical methods to automatically match natural language user queries against records. For almost as long, there has been interest in using natural language processing to enhance single term matching by adding phrases (Fagan, 1989), yet to date natural language processing techniques have not significantly improved

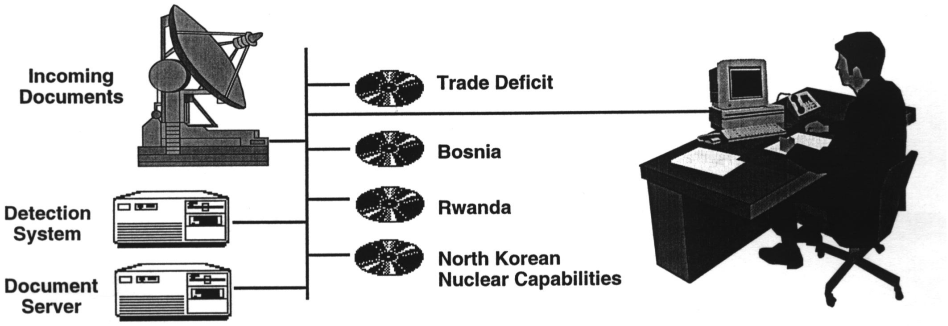

Figure 7.1: The document retrieval process.

performance of document retrieval, although much effort has been expended in various attempts. The motivation and drive for using natural language processing (NLP) in document retrieval is mostly intuitive; users decide on the relevance of documents by reading and analyzing them and if we can automate document analysis this should help in the process of deciding on document relevance.

Some of the research into document retrieval has taken place in the ARPA-sponsored TIPSTER project. One of the TIPSTER groups, the University of Massachusetts at Amherst, experimented with expansion of their state-of-the-art INQUERY retrieval system so that it was able to handle the three gigabyte test collection. This included research in the use of query structures, document structures, and extensive experimentation in the use of phrases (Broglio, Callan, et al., 1993). These phrases (usually noun phrases) were found using a part-of-speech tagger and were used either to improve query performance or to expand the query. In general, the use of phrases, as opposed to single terms, for retrieval did not significantly improve performance, although the use of noun phrases to expand a query shows much more promise. This group has found phrases to be useful in retrieval for smaller collections, or for collections in a narrow domain.

A second TIPSTER group using natural language processing techniques was Syracuse University. A new system, the DR-LINK system, based on automatically finding conceptual structures for both documents and queries, was developed using extensive natural language processing techniques such as document structure discovery, discourse analysis, subject classification, and complex nominal encapsulation. This very complex system was barely finished by the end of phase I (Liddy & Myaeng, 1993), but represents the most complex natural language processing system ever developed for document retrieval .

The TIPSTER project has progressed to a second phase that will involve even more collaboration between NLP researchers and experts. The plan is to develop an architecture that will allow standardized communication between document retrieval modules (usually statistically based) and natural language processing modules (usually linguistically based). The architecture will then be used to build several projects that require the use of both types of techniques. In addition to this theme, the TIPSTER phase II project will investigate more thoroughly the specific contributions of natural language processing to enhanced retrieval performance. Two different groups, the University of Massachusetts at Amherst group combined with a natural language group at BBN Inc., and a group from New York University will perform many experiments that are likely to uncover further evidence of the usefulness of natural language processing in document retrieval.

The same collection used for testing in the TIPSTER project has been utilized by a much larger worldwide community of researchers in the series of Text REtrieval Conference (TREC) evaluation tasks. Research groups representing very diverse approaches to document retrieval have taken part in this annual event and many have used NLP resources like lexicons, dictionaries, thesauri, proper name recognizers and databases, etc. One of these groups, New York University, investigated the gains for using more intensive natural language processing on top of a traditional statistical retrieval system (Strzalkowski, Carballo, et al., 1995). This group did a complete parse of the two Gbyte texts to locate content-carrying terms, discover relationships between these terms, and then use these terms to expand or modify the queries. This entire process is completely automatic, and major effort has been put into the efficiency of the natural language processing part of the system. A second group using natural language processing was the group from General Electric Research and Development Center (Jacobs, 1994). They used natural language processing techniques to extract information from (mostly) the training texts. This information was then used to create manual filters for the routing task part of TREC. Another group using natural language processing techniques in TREC was CLARITECH (Evans & Lefferts, 1994). This group used only noun phrases for retrieval and built dynamic thesauri for query expansion for each topic using noun phrases found in highly ranked documents. A group from Dublin City University derived tree structures from texts based on syntactic analysis and incorporated syntactic ambiguities into the trees (Smeaton, O'Donnell, et al., 1995). In this case document retrieval used a tree-matching algorithm to rank documents. Finally, a group from Siemens used the WordNet lexical database as a basis for query expansion (Voorhees, Gupta, et al., 1995) with mixed results.

The situation in the U.S. as outlined above is very similar to the situation in Europe. The European Commission's Linguistic Research and Engineering (LRE) sub-programme funds projects like CRISTAL, which is developing a multilingual interface to a database of French newspaper stories using NLP techniques, and RENOS, which is doing similar work in the legal domain. The E.U.-funded SIMPR project also used morpho-syntactic analysis to identify indexing phrases for text. Other European work using NLP is reported in Hess (1992); Ruge (1992); Schwarz and Thurmair (1986); Chiaramella and Nie (1990) and summarized in Smeaton (1992).

Most researchers in the information retrieval community believe that retrieval effectiveness is easier to improve by means of statistical methods than by NLP-based approaches and this is borne out by results, although there are exceptions. The fact that only a fraction of information retrieval research is based on extensive natural language processing techniques indicates that NLP techniques do not dominate the current thrust of information retrieval research as does something like the Vector Space Model. Yet NLP resources used in extracting information from text as described by Paul Jacobs in section 7.3, resources like thesauri, lexicons, dictionaries, proper name databases, are used regularly in information retrieval research. It seems, therefore, that NLP *resources* rather than NLP techniques are having more of an impact on document retrieval effectiveness at present. Part of the reason for this is that natural language processing techniques are generally not designed to handle large amounts of text from many different domains. This is reminiscent of the situation with respect to information extraction which likewise is not currently successful in broad domains. But information retrieval systems do need to work on broad domains in order to be useful, and the way NLP techniques are being used in information retrieval research is to attempt to integrate them with the dominant statistically-based approaches, almost piggy-backing them together. There is, however, an inherent granularity mismatch between the statistical techniques used in information retrieval and the linguistic techniques used in natural language processing. The statistical techniques attempt to match the rough statistical approximation of a record to a query. Further refinement of this process using fine-grained natural language processing techniques often adds only noise to the matching process, or fails because of the vagaries of language use. The proper integration of these two techniques is very difficult and may be years in coming. What is needed is the development of NLP techniques specifically for document retrieval and, vice versa, the development of document retrieval techniques specifically for taking advantage of NLP techniques.

Future Directions

The recommendations for further research are therefore to continue to pursue this integration but paying more attention to how to adapt the output of current natural language methods to improving information retrieval techniques. In addition, NLP techniques could be used directly to produce tools for information retrieval, such as creating knowledge bases or simple thesauri using data mining.

7.3 Chapter References

Adriaens, G. and Schreuers, D. (1992). From COGRAM to ALCOGRAM: Toward a controlled English grammar checker. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 595–601, Nantes, France. ACL.

AECMA (1995). *AECMA Simplified English: A Guide for the Preparation of Aircraft Maintenance Documentation in the International Aerosace Maintenance Language*. AECMA, Brussels.

Belkin, N. J. and Croft, W. B. (1987). Retrieval techniques. In Williams, M., editor, *Annual Review of Information Science and Technology*, volume 22, pages 109–145. Elsevier, New York.

Broglio, J., Callan, J., and Croft, W. (1993). The INQUERY system. In Merchant, R., editor, *Proceedings of the TIPSTER Text Program—Phase I*, San Mateo, California. Morgan Kaufmann.

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM*, 36(12).

Chiaramella, Y. and Nie, J. (1990). A retrieval model based on an extended modal logic and its applications to the RIME experimental approach. In Vidick, J.-L., editor, *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, pages 25–44, Brussels, Belgium. ACM.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In Zernik, U., editor, *Lexical Acquisition: Using On-Line Resources To Build A Lexicon*. Lawrence Earlbaum, Hillsdale, New Jersey.

Ciravegna, F., Campia, P., and Colognese, A. (1992). Knowledge extraction by SINTESI. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 1244–1248, Nantes, France. ACL.

COLING (1992). *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France. ACL.

David, P. A. (1991). *Technology and Productivity The Challenge for Economic Policy*, chapter Computer and Dynamo—The modern productivity paradox in a not-too-distant mirror. ODED, Paris.

DeJong, G. F. (1979). *Skimming stories in real time: an experiment in integrated understanding*. PhD thesis, Yale University.

Endres-Niggemeyer, B., Hobbs, J., and Sparck Jones, K. (1995). Summarizing text for intelligent communication. Technical Report Dagstuhl Seminar Report 79, 13.12-19.12.93 (9350), IBFI, Dagstuhl. http://www.bid.fh-hannover.de/SimSum/Abstract/ (Short and Full versions, the latter only available in electronic form).

Evans, D. and Lefferts, R. (1994). Design and evaluation of the CLARIT–TREC-2 system. In Harman, D., editor, *National Institute of Standards and Technology Special Publication No. 500-215 on the The Second Text REtrieval Conference (TREC-2)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Fagan, J. L. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40(2):115–132.

Hahn, U. (1990). Topic parsing: accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26(1):135–170.

Harman, D., editor (1994). *National Institute of Standards and Technology Special Publication No. 500-215 on the The Second Text REtrieval Conference (TREC-2)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Hess, M. (1992). An incrementally extensible document retrieval system based on linguistic and logical principles. In *Proceedings of the 15th SIGIR Conference*, pages 190–197, Copenhagen, Denmarkp.

Hoard, J., Wojcik, R., and Holzhauser, K. (1992). An automated grammar and style checker for writers of simplified English. In Holt, P. and Williams, N., editors, *Computers and Writing*. Kluwer Academic Publishers, Boston.

IPM (1995). Special issue on automatic summarizing. *Information Processing and Management*, 31(3).

Iwanska, L., Appelt, D., Ayuso, D., Dahlgren, K., Glover Stalls, B., Grishman, R., Krupka, G., Montgomery, C., and Riloff, E. (1991). Computational aspects of discourse in the context of MUC-3. In *Proceedings of the Third Message Understanding Conference*, San Diego, California. Morgan Kaufmann.

Jacobs, P. (1994). GE in TREC-2: Results of a Boolean approximation method for routing and retrieval. In Harman, D., editor, *National Institute of Standards and Technology Special Publication No. 500-215 on the The Second Text REtrieval Conference (TREC-2)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Jacobs, P., Krupka, G., Rau, L., Mauldin, M., Mitamura, T., Kitani, T., Sider, I., and Childs, L. (1993). The TIPSTER/SHOGUN project. In *Proceedings of the TIPSTER Phase I Final Meeting*, San Mateo, California. Morgan Kaufmann.

Liddy, E. D. et al. (1993). Development, implementation and testing of a discourse model for newspaper texts. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, pages 159–164, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Liddy, E. D. and Myaeng, S. H. (1993). DR–LINK: A system update for TREC-2. In Merchant, R., editor, *Proceedings of the TIPSTER Text Program—Phase I*, San Mateo, California. Morgan Kaufmann.

LIM (1993). The boeing simplified English checker. *Language Industry Monitor*, (13).

Marsh, E., Hamburger, H., and Grishman, R. (1984). A production rule system for message summarization. In *Proceedings of the National Conference on Artificial Intelligence*, pages 243–246. American Association for Artificial Intelligence.

Mellish, C. S. et al. (1995). The TIC message analyser. *Computational Linguistics*.

Paice, C. D. (1990). Constructing literature abstracts by computer. *Information Processing and Management*, 26(1):171–186.

Pereira, F. (1990). Finite-state approximations of grammars. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, pages 20–25, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Rau, L. F. (1988). Conceptual information extraction and information retrieval from natural language input. In *Proceedings of the Conference on User-Oriented, Content-Based, Text and Image Handling*, pages 424–437, Cambridge, Massachusetts.

Ruge (1992). Experiments in linguistically based term associations. *Information Processing and Management*, 28(3).

Schwarz and Thurmair, editors (1986). *Informationslinguistische texterschliessung*. Hildesheim: Georg Olms Verlag.

Smeaton, A. (1992). Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal*, 35(3).

Smeaton, A. F., O'Donnell, R., and Kelledy, F. (1995). Indexing structures derived from syntax in TREC-3: System description. In *National Institute of Standards and Technology Special Publication on the The Third Text REtrieval Conference (TREC-3)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Sparck Jones, K. (1993). What might be in a summary? In Knorz, G., Krause, J., and Womser-Hacker, C., editors, *Information retrieval '93: von der modellierung zur anwendung*, pages 9–26. Konstanz, Universitatsverlag Konstanz.

Strzalkowski, T., Carballo, J. P., and Marinescu, M. (1995). Natural language information retrieval: TREC-3 report. In *National Institute of Standards and Technology Special Publication on the The Third Text REtrieval Conference (TREC-3)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

TREC (1995). *National Institute of Standards and Technology Special Publication on the The Third Text REtrieval Conference (TREC-3)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Voorhees, E., Gupta, N. K., and Johnson-Laird, B. (1995). The collection fusion problem. In *National Institute of Standards and Technology Special Publication on the The Third Text REtrieval Conference (TREC-3)*, pages 95–104, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Wojcik, R., Harrison, P., and Bremer, J. (1993). Using bracketed parses to evaluate a grammar checking application. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 38–45, Columbus, Ohio. ACL.

Young, S. R. and Hayes, P. J. (1985). Automatic classification and summarization of banking telexes. In *Proceedings of the Second Conference on Artificial Intelligence Applications*, pages 402–408.