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")

7.3 Text Interpretation: Extracting Information

Paul Jacobs

SRA International, Arlington, Virginia, USA

The proliferation of on-line text motivates most current work in text interpretation. Although massive volumes of information are available at low cost in free text form, people cannot read and digest this information any faster than before; in fact, for the most part they can digest even less. Often, being able to make efficient use of information from text requires that the information be put in some sort of structured format, for example, in a relational database, or systematically indexed and linked. Currently, extracting the information required for a useful database or index is usually an expensive manual process; hence on-line text creates a need for automatic text processing methods to extract the information automatically (Figure 7.1).

Current methods and systems can digest and analyze significant volumes of text at rates of a few thousand words per minute. Using *text skimming*, often driven by finite-state recognizers (discussed in chapters 3 and 11 of this volume), current methods generally start by identifying key artifacts in the text, such as proper names, dates, times, and locations, and then use a combination of linguistic constraints and domain knowledge to identify the important content of each relevant text. For example, in news stories about joint ventures, a system can usually identify joint venture partners by locating names of companies, finding linguistic relations between company names and words that describe business tie-ups, and using certain domain knowledge, such as understanding that ventures generally involve at least two partners and result in the formation of a new company. Other applications are illustrated in Ciravegna, Campia, et al. (1992); Mellish et al. (1995). Although there has been independent work in this area and there are a number of systems in commercial use, much of the recent progress in this area has come from U.S. government-sponsored programs and evaluation conferences, including the TIPSTER Text Program and the

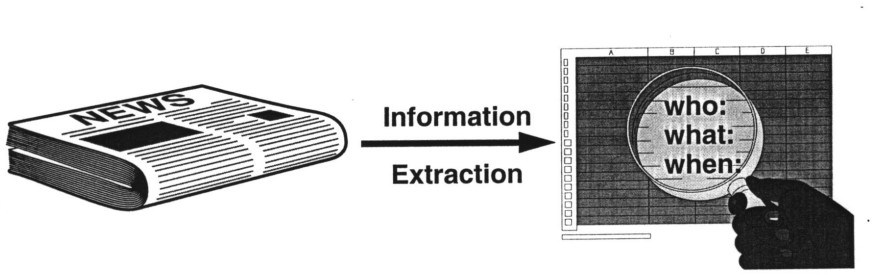

Figure 7.1: The problem of information extraction from text.

MUC and TREC evaluations described in chapter 13. In information extraction from text, the TIPSTER program, for example, fostered the development of systems that could extract many important details from news stories in English and Japanese. The scope of this task was much broader than in any previous project.

The current state of the art has produced rapid advances in the robustness and applicability of these methods. However, current systems are limited because they invariably rely, at least to some degree, on domain knowledge or other specialized models, which still demands time and effort (usually several person-months, even in limited domains). These problems are tempered somewhat by the availability of on-line resources, such as lexicons, corpora, lists of companies, gazetteers, and so forth, but the issue of how to develop a technology base that applies to many problems is still the major challenge.

In recent years, technology has progressed quite rapidly, from systems that could accurately process text in only very limited domains (for example, engine service reports) to programs that can perform useful information extraction from a very broad range of texts (for example, business news). The two main forces behind these advances are: (1) the development of robust text processing architectures, including finite state approximation and other shallow but effective sentence processing methods, and (2) the emergence of weak heuristic and statistical methods that help to overcome knowledge acquisition problems by making use of corpus and training data.

Finite-state approximation (Jacobs, Krupka, et al., 1993; Pereira, 1990) is a key element of current text interpretation methods. Finite-state recognizers generally admit a broader range of possible sentences than most parsers based on context-free grammars, and usually apply syntactic constraints in a weaker fashion. Although this means that finite-state recognizers will sometimes treat sentences as grammatical when they are not, the usual effect is that the finite state approximation is more efficient and fault tolerant than a context-free model.

The success of finite-state and other shallow recognizers, however, depends on the ability to express enough word knowledge and domain knowledge to control interpretation. While more powerful parsers tend to be controlled mainly by linguistic constraints, finite state recognizers usually depend on lexical constraints to select the best interpretation of an input. In limited domains, these constraints are part of the domain model; for example, when the phrase *unidentified assailant* appears in a sentence with *terrorist attack*, it is quite likely that the assailant is the perpetrator of the attack.

In broader domains, successful interpretation using shallow sentence processing requires lexical data rather than domain knowledge. Such data can often be obtained from a corpus using statistical methods (Church, Gale, et al., 1991). These statistical models have been of only limited help so far in information extraction systems, but they show promise for continuing to improve the coverage and accuracy of information extraction in the future.

Much of the key information in interpreting texts in these applications comes not from sentences but from larger discourse units, such as paragraphs and even complete documents. Interpreting words and phrases in the context of a complete discourse, and identifying the discourse structure of extended texts, are important components of text interpretation. At present, discourse models rely mostly on domain knowledge (Iwanska, Appelt, et al., 1991). Like the problem of controlling sentence parsing, obtaining more general discourse processing capabilities seems to depend on the ability to use discourse knowledge acquired from examples in place of detailed hand-crafted domain models.

Future Directions

We can expect that the future of information extraction will bring broader and more complete text interpretation capabilities; this will help systems to categorize, index, summarize, and generalize from texts from information sources such as newspapers and reference materials. Such progress depends now on the development of better architectures for handling information beyond the sentence level, and on continued progress in acquiring knowledge from corpus data.

7.4 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.