TAMAS E. DOSZKOCS

Computer Research Scientist
National Library of Medicine
Bethesda, Maryland

Natural Language User Interfaces in Information Retrieval

Introduction

This paper examines the role of natural language (NL) processing in information retrieval in the context of large operational information retrieval systems and services. State-of-the-art information retrieval systems combine the functional capabilities of the conventional inverted file—Boolean logic term adjacency approach—commonly employed by commercial search services, with statistical-combinatorial techniques pioneered in experimental information retrieval (IR) research, and formal natural language processing methods and tools borrowed from artificial intelligence (AI). The emergence and ever increasing importance of end-user searching provides challenging opportunities for the integration of sophisticated natural language analysis and processing techniques in user friendly interfaces.

IR systems achieve remarkable search speed and flexibility despite the virtual absence of formal language analysis procedures and meaning interpretation of the underlying text content.[1] State-of-the-art IR systems pragmatically blend the best features of diverse probabilistic-combinatorial and Boolean logic retrieval models and readily support free-form natural language user interfaces.[2] Yet, such direct natural language interfaces need to incorporate more sophisticated natural language processing and other Applied artificial intelligence techniques in order to cope intelligently with the inherent ambiguities of natural language queries and text, and to compensate for the inevitable semantic loss and confounding inherent in indexing and query matching.[3]

By the same token, AI approaches to natural language processing, particularly as applied to NL user interfaces and text searching, benefit from proven IR concepts and techniques in order to transcend still-prevalent domain-specificity and performance problems.[4] Thus natural language databases and user friendly online searching (UFOS) represent challenging and mutually supportive common problem areas for both IR and AI.

Information Retrieval

Although a variety of access methods have been developed for text retrieval,[5] operational IR systems are, almost without exception, characterized by the dominance of the inverted file, Boolean logic search paradigm.[6] In their basic form, IR systems are designed to manipulate fundamentally simple natural language text structures—such as bibliographic citations or full-text documents—although some IR software packages have been enhanced to incorporate generalized database management system (DBMS) access methods and special processing functions needed for the handling of integrated textual, numeric, graphics, and image data. The automatic "indexing rules" or algorithms are, by and large, lexically based procedures guided by delimiters and/or lists of nonindexed "stopwords," and the resulting inverted search keys are typically organized into B-tree structures for acceptable trade-off between speed of access and ease of updating. This approach is, for practical purposes, highly flexible and domain independent, although distinct indexing rules are needed in order to usefully fragment special textual fields, for instance, chemical names. In addition, in most IR systems, the automatically generated keywords are frequently augmented with human-assigned subject headings or thesaurus entries, mostly noun phrases, for added syntactic and semantic precision in searching.

Search queries, regardless of their form of input, must be ultimately transformed (typically, by trained intermediary searchers) to conform to the basic indexing scheme of the given IR system. Query search keys are matched against the inverted file search keys, and the corresponding ordered inverted lists are compared using exact-match Boolean AND, OR, AND NOT logic operators implemented as set operations (intersection, union, and complement, respectively) on the inverted lists. The process is exceedingly fast and efficient as implemented via currently available hardware and software architectures. The lack of linguistic and cognitive analysis procedures at indexing time and the resulting precision/recall problems are, to some extent, alleviated by the availability of powerful pattern-matching functions and metric (hierarchical/positional) operators, such as character masking, truncation, and adjacency. Trained intermediary searchers, in turn, provide the needed augmented "knowledge base," "inference engine," and "control strategy" for the proper configuration and sequencing of retrieval operations.

In the last few years, operational IR systems have been implemented that reflect the influence of G. Salton's seminal work and incorporate many useful ideas and results from theoretical and experimental IR research going back to the mid-1960s.[7]

IR research on the whole has been dominated by nonlinguistic, primarily lexical-statistical NLP methods as exemplified by Salton's SMART system[8] and the work of researchers such as Bookstein, Kraft, Rijsbergen,[9] and many others. Although experimental IR research prototypes have often suffered from problems of scale limitations in their technical approaches, IR research has nonetheless contributed a coherent conceptual framework and identified a core of desirable IR functions and evaluation methods that are of particular relevance to the design and implementation of end-user oriented information retrieval systems and services. The most important among these are the notions of (1) unrestricted natural language query input, (2) closest-match search strategy, (3) the ranking of the retrieval output according to expected relevance to the query, and (4) the dynamic utilization of user feedback in automatic query reformulation and search strategy modification.

Large-scale implementations of NL IR interfaces, such as CITE,[10] efficiently combine the best features of the Boolean and the probabilistic/combinatorial retrieval models with limited "intelligent" computational linguistic analysis and AI-type search heuristics. Such end-user oriented systems treat unrestricted natural language both in queries and text records as the least common denominator among different searchers, databases, and IR systems, offering the potential of true transportability and transparency among diverse users, information sources, and search systems. In order to successfully emulate the trained searcher's augmented retrieval "knowledge base," "inference engine," and "control strategy," however, UFOS must possess intelligent NLP capabilities of greater sophistication.[11]

The convergence of a number of hardware, software, and user trends necessitate the augmentation of conventional information retrieval and filtering capabilities with appropriate and efficiently implemented techniques adapted from AI application areas—such as natural language processing and understanding, expert systems, intelligent information management, and intelligent problem-solving. The trends include: full-text databases; very large databases; mixed information sources containing text, numerics, graphics, image, and other data; the increased diversity, depth, and breadth of information sources; hybrid technologies (e.g., compact disc-read-only memory, CD-ROM); special-purpose IR hardware; non-von-Neumann computer architectures; associative memories; distributed and parallel processing; and interactive end-user searching and personal, computer-enhanced "information metabolism."

Artificial Intelligence

Artificial intelligence, particularly applied artificial intelligence, is a growth industry of high expectations.[12] Notwithstanding the difficulty of defining intelligence, let alone artificial intelligence, and despite the many lingering doubts and the seeming tower of Babel situation, the AI field possesses a healthy basic research underpinning and the AI community has been successful in developing important concepts, tools, techniques, and applications of interest to other information-intensive disciplines.[13] Just as "conventional" programming languages and diverse data representation models in "classical" IR and DBMS systems serve as problem-solving tools for a wide variety of applications, AI knowledge representation models, such as production rules,[14] frames,[15] and semantic networks as well as AI programming languages, such as LISP (list processing), PROLOG (logic programming),[16] OPS5 (rule-based programming),[17] SMALLTALK (object-oriented programming), and other general purpose techniques and tools (e.g., heuristic search, ATN [automatic translation network] parsers, and rapid prototyping,[18]) represent versatile AI implements for a broad range of applications including IR.

Natural Language Processing—Problems

NLP and computational approaches to dealing with natural language were among the earliest objects of interest in AI research. The well-known ambitions and subsequent failures of machine translation research and development projects in the 1960s, as well as the recent success of more limited yet highly pragmatic computer-assisted language translation systems, epitomize both the great difficulties involved in coping with natural language and the degree of maturation and realistic goal setting of the field.

Understanding language, though it appears to humans to be naturally easy, is a difficult task that involves highly cognitive and not yet totally understood intellectual and psychological processes.[19] Schank argues that NLP is both a linguistic and a cognitive task, and it cannot occur without a knowledge base concerning the relevant subject area.[20] This in turn points to a synthesis of natural language processing and expert systems techniques. Typical areas of concern and investigation in AI NLP research involve automated lexical, syntactic, and semantic analysis; dealing with ill-formed or fragmentary input; ellipsis; conjunction and negation; pronoun/anaphora resolution; definite noun phrases, quantification; beliefs and intentions; fail-soft recovery; space and time and contextual understanding—to name just a few.

The role of NLP in information retrieval is less clear[21] given that the very nature of the textual documents traditionally dealt with in IR does not warrant elaborate analysis, and the fundamentally mechanistic techniques developed in IR have served to handle the passive query-to-document matching problem at an acceptable level of success. The advent of unmediated "user friendly" search interfaces on the one hand, and the considerably broader scope and depth of today's full-text databases on the other, however, necessitate a careful reexamination of this role.

Computational linguistic processing has reached fairly reliable stability and practical utility in morphological and syntactic analysis. While procedures for morphological analysis are decidedly nontrivial, they are generally more straightforward and efficient than syntactic analysis. Common strategies for morphological analysis employ some or all of the following: morphological rewriting rules, dictionary lookup, inflection generator, complexity testing, and idiom and compound recognition.[22] Prefixing and suffixing are the main concern in processing English. Often, in processing specialty languages—e.g., medical English—special attention must be paid to characteristic prefixing, suffixing, and morphosemantic problems.[23] Compare, for instance:

ALEXIA—DYSLEXIA
VITAMINS—AVITAMINOSIS
ANXIETY—ANTIANXIETY AGENTS
INFECT—DISINFECT
HYPERTENSION—HYPOTENSION
ADJUSTMENT—MALADJUSTMENT
ANTIBIOTICS—"-MYCINS"
INFLAMMATION—"-ITIS"

Reasonably—i.e. relatively—robust and efficient NL syntactic parsers have been developed and incorporated into many artificial intelligence applications. "Parsing efficiency is crucial when building practical natural language systems. This is especially the case for interactive systems, such as natural language database access, interfaces to expert systems and interactive machine translation."[24] Tomita's LR context-free parsing algorithm, for instance, takes advantage of the left-to-rightness of natural-language user input, and parsing starts as soon as the user types the first word, thus reducing apparent response time from the user's point of view.

Despite significant advances, a number of unsolved problems and limitations in AI NLP remain. Most importantly, none of the systems developed to date are fluent in the use of unrestricted natural language.[25]

For potential IR use of AI NLP techniques, it is important to remember that while most full sentences are unambiguous, their component parts are frequently ambiguous. Since IR systems utilize component or fragmentary lexical, syntactic, and semantic units as a result of indexing, and since indexing inevitably implies omission, the disambiguation problem in domain-independent IR systems is much less tractable than in narrow-domain artificial intelligence applications.

Acoustic-phonetic, lexical, syntactic, and pragmatic language complexity, as reflected in word-sense ambiguity, structural ambiguity, and the referential ambiguity of noun phrases, pose very challenging problems in user friendly artificial intelligence and information retrieval interfaces. Few if any existing NLP systems, for instance, are able to recognize and disambiguate compounds (compare database *v.* data base), acronyms (compare G-SUIT), or abbreviations (compare MD ==> Maryland ==> physician) in large textual databases.

Noun-phrase ambiguity may be further compounded in keyword indexing by ignoring word order, field, and other text boundaries, and by partial match search strategy in query matching, e.g.,:

MEDICAL LITERATURE ==> MEDICINE IN LITERATURE ==>
LITERATURE IN MEDICINE

SEXUAL PERVERSION ==> SEXUAL ABSTINENCE

There is a lot of humor in all of this, as in:

HUMOR IN MEDICINE ==> AQUAEOUS HUMOR IN MEDICINE

or, put differently, language understanding can be a joke. For proper perspective it is worth keeping in mind that there still does not exist any NLP computer program that can handle nearly all of English syntax. None can even come close to coping with the semantics of all of the English language and none is on the horizon. "Wretched and confusing prose can defeat even human comprehension."[27]

The impossibly large number of rules that would be necessary for the morphological, syntactic, and semantic disambiguation of natural language in real-life multidisciplinary knowledge domains precludes using the rule-based expert system approach as well. (No accurate estimation has been made of how many context-free rules are needed to cover English almost completely, but the number is very large.)[28] From the survey of the literature it appears that artificial intelligence is not quite ready to field any system flexible enough for mass use, save a few relatively small problem domains.[29]

Can information retrieval facilitate finding artificial intelligence solutions to practical natural language processing problems? Certainly so!

Domain-independent IR search techniques can serve as efficient filters for more refined in-depth natural language processing. As more and more powerful AI concepts and tools become available (compare AI PC toolkits) to more and more end users, the problems of scale and performance limitations that characterize contemporary AI systems will be gradually overcome. To the extent that the AI and IR communities will be able to learn from each other and gain mutual insights into the problems of language and searching, there will exist intelligent IR systems and effective domain-independent AI NL search systems. And perhaps—in theory at least—the distinctions between the two (mind) sets will be "fuzzy" at best.[30]

Natural Language and User Friendly Online Searching

Natural language interface technology represents a major breakthrough in "user friendly" computer systems.[31] Along with other end-user-oriented interface techniques (such as menus, windowing, graphics, icons, pointing, touching) commercial implementations of NL interfaces (such as Artificial Intelligence Corporation's INTELLECT or Texas Instrument's NLMenu DBMS front-end products) target the largest hitherto untapped segment of the information marketplace, namely end users. The same holds true for public access online catalogs in libraries and for user friendly IR interfaces in general,[32] and for natural language information retrieval interfaces in particular (e.g., the National Library of Medicine's CITE system for searching the world's largest medical literature databases, MEDLINE and CATLINE).[33]

Online systems and the personal computer revolution have made computer resources universally available. Now a similar revolution in software (i.e., user interfaces) is needed to make the computer universally usable.[34] Natural language interfaces in information retrieval and artificial intelligence are the scouts, shock troops, vanguards, and sometimes martyrs of the user interface revolution.

With occasional end users already outnumbering trained professional searchers in the user populations of online information utilities like The SOURCE and CompuServe, it becomes increasingly important to develop and refine analytical cognitive models to better assess the user's skills and understanding of the information stored, and to match the user's cognitive model to the system's model and knowledge representation.

Of course natural language is not always natural for a user interface,[35] but it is particularly well suited for IR and DBMS interfaces due to the very large number of potential users, high volume of query transactions, distribution of costs over large numbers of users, and the fundamentally linguistic nature of the user-system information exchange.[36]

Ease of learning, ease of use, and transportability are among the most attractive features of natural language front-ends. Natural language shifts the burden of understanding from the user to the system thus allowing the user to focus on the problem at hand.[37] At the current state-of-the-art, the high development cost and less than 100 percent reliability of knowledge-based domain-specific NL DBMS systems (compare EXPLORER, developed by Cognitive Systems, Inc. for oil exploration) appear to have palled their widespread development and commercial use. Considerably more commercial success has been achieved by domain-independent NL DBMS front-ends. INTELLECT, for example, substitutes knowledge of the physical and logical structure of the database and the topology of user interactions and system functions for domain-specific semantic knowledge, using the inverted DBMS index as lexical pointers to the system's rather small domain-specific and semantically augmented dictionary. Despite its shortcomings in ambiguity resolution,[38] INTELLECT succeeds in its practical utility, reliability, and operational performance.

It is interesting to note that while a DBMS NL front-end like INTELLECT has to cope with the full spectrum of linguistic processing problems of a "habitable subset" of the English language—namely the limited domain of the DBMS command language and a relatively small number of query interaction paradigms—it does not have to deal with language ambiguity and matching problems at the level of the database content, due to the fact that DBMS systems deal, for the most part, with discrete and finite nontextual data.

By contrast, a natural language information retrieval textual database interface, like NLM's CITE system[39] must focus on language ambiguity and matching problems for all of English. At the same time, CITE need not invest a great deal of effort in full-scale linguistic analysis in query-to-command language translation due to the relatively small number of IR commands available and the simplicity of the underlying database structure. (It would be perfectly feasible and appropriate to use INTELLECT-like NLP techniques, instead of menu choices, in CITE to "understand" the type of query at hand, identify its topical component, as well as any implied or explicitly stated limitations as to type of material desired, language restrictions, currency of material.)

Natural Language Queries in NL Databases

Natural language information retrieval interfaces must then deal with the problem of language ambiguity at the level of both the query and the database content and must resolve the matching problem between free-form queries and the database in a manner acceptable to end users who typically approach a natural language system with high expectations of (artificial machine) intelligence. The basic problem areas of matching can be conveniently divided into lexical, syntactic, semantic, and special concerns.

Lexical Problems

The inverted keyword index of a NL database of 1 million records is likely to contain in excess of a quarter million distinct lexical words.[40] The frequency distribution of these lexical entries will typically conform to the characteristics of the Zipf distribution. This empirical fact has the following practical implications:

1. There will be a few dozen to a few hundred highly posted (i.e., high database frequency) index entries, many of which will be natural candidates for exclusion from the index (compare "stopwords").
2. Approximately half of all the index entries will have a database frequency of one, and as many as half or more of these may be misspellings. This suggests the incorporation of efficient spelling error detection/correction algorithms and dictionaries in the NL user interface and in indexing, with the dictionary preferably derived from and dynamically updated in conjunction with the database itself.
3. Knowledge of the frequency distribution of the lexical entries suggests implicit heuristics for automatic search strategy formulation. The NL interface must be capable of recognizing and properly dealing with frequent acronyms, abbreviations, numerals, chemical names, names of people/syndromes:

X RAY ==> X-RAY; CAT SCAN; US ==> USA ==> U.S. ==> U.S.A.; AI ==> ARTIFICIAL INTELLIGENCE; VITAMIN B 1 ==> VITAMIN B1 ==> THIAMIN; TYPE 1 ==> TYPE I; FACTOR V; 2,4,5-T ==> TRICHLOROPHENOXYACETIC ACID ==> AGENT ORANGE; EPSTEIN-BARR VIRUS; BARR BODIES; BARRE-LIEOU SYNDROME; GUILLAN-BARRE SYNDROME

orthographic and phonetic transcribing and transliteration:

TUMOUR ==> TUMOR;
GYNAECOLOGY ==> GYNEKOLOGIA ==> GYNECOLOGY

idioms and clichés:

OFF COLOR; CHANGE OF HEART; OUT OF SIGHT

slang, lingo, jargon, and lore:

POT; SPEED; ANGEL DUST; CRACK

stopwords that are noun-phrase components:

VITAMIN A; HEPATITIS A; "TO BE OR NOT TO BE";
THE "ME" GENERATION

compound words:

DATABASE ==> DATA BASE; ONLINE ==> ON-LINE ==>
ON LINE; BACKACHE ==> BACK ACHE ==> BACK PAIN

Morphological Analysis

Morphological analysis (stemming or "conflation") warrants special attention in large-scale natural language information retrieval interfaces. The automatic identification of lexical roots in inverted file based operational IR systems is the first step in the process of matching query words and inverted index entries. The latter may have been derived from auxiliary vocabularies that serve as semantic database navigational tools, or from the database records themselves. The following examples from the MEDLINE and MEDICAL SUBJECT HEADINGS inverted indexes illustrate ever-present and familiar lexical ambiguities and semantic noise introduced as a result of using word roots with variable-length masking operations in matching:

ACCESS ==> ACCESSORY
ASPIRIN ==> ASPIRATION
AUDIT ==> AUDITORY
BATTERED ==> BATTERY
COMMUNICABLE ==> COMMUNICATIONS
CREATINE ==> CREATIVENESS
DIGITAL ==> DIGITALIS
EXTREME ==> EXTREMITIES
EXPECTATION ==> EXPECTORANT
INFANT ==> INFANTILISM
INFORM ==> INFORMAL ==> INFORMER
LABOR ==> LABORATORY
MEDIA ==> MEDIAN ==> MEDIATED
METHOD ==> METHODIST
MIGRAINE ==> MIGRANT
NURSERY ==> NURSES ==> NURSING
RECEPTION ==> RECEPTORS
SHORT ==> SHORTAGE ==> SHORTHAND
TREAT ==> TREATMENT ==> TREATISE ==> TREATY

Short words require even more caution:

AID ==> AIDS
ANAL ==> ANALYSIS
APE ==> APES ==> APEX
ARM ==> ARMY
CARE ==> CAREER
FAIR ==> FAIRS
HEAR ==> HEARING ==> HEART ==> HEARTWATER ==> HEARTWORM

Many thousands of other such examples could be found.

Syntax-Related Problems

Natural language topical surrogates—e.g. book and journal titles, headlines, table of contents, back-of-the-book indexes—are usually expressed via larger syntactic units, mostly noun phrases. Noun phrases are frequently ambiguous in and of themselves—e.g., SHELLFISH POISONING ==> POISONING OF SHELLFISH ==> POISONING BY SHELLFISH. The use of Boolean operators, metric operators, and *m* out of *n* weighted-logic closest-match search strategy—in order to compensate for the lack of linguistic analysis in indexing—further compounds the text-matching problem. Consider, for example:

ABUSE ==> CHILD ABUSE ==> DRUG ABUSE ==> ELDER ABUSE ==> SPOUSE ABUSE

CRISIS MANAGEMENT ==> MANAGEMENT CRISIS ==> MANAGEMENT BY CRISIS ==> CRISIS BY MANAGEMENT

KIDNEY ==> KIDNEY BEAN LECTINS ==> KIDNEY DISEASES

SEXUAL ABSTINENCE ==> SEXUAL PERVERSION

SHORT TERM EFFECTS ==> SHORT TERM MEMORY ==> SHORT TERM PSYCHOTHERAPY

Since literally hundreds of similar lexical and syntactic matching problems are encountered daily in a large operational NL IR system, it is evident that automatic query analysis and matching can substantially benefit from morphological and syntactic analysis in order to lend additional precision to the available truncation, character masking, Boolean, metric, weighted-logic, or generalized pattern-matching strategies. Consider for instance the automatic generation of Boolean search statements from NL queries.[41]

Semantic Problems

A great many formal semantic aids and ad hoc heuristics are used by trained searchers when interacting with information retrieval systems. Some examples are controlled vocabularies, "hedges," "preexplodes," multidatabase cross indexes, stored search strategies, and the like. Systems that rely on controlled vocabularies often lack in currency, database warrant, or conceptual exhaustivity. For instance, *Medical Subject Headings* does not currently (1986 edition) have a subject heading for BIOTECHNOLOGY, and it uses PMS as a cross reference to an older subject heading PREGNANT MARE SERUM (GONADOTROPINS, EQUINE), but at the same time PMS is not linked to PREMENSTRUAL SYNDROME, nor is AI linked to ARTIFICIAL INTELLIGENCE.

As was noted earlier, systems without automated semantic aids shift the full burden of query understanding and matching on the searcher. NL IR interfaces must minimally rely on and must intelligently utilize existing machine-readable semantic search aids. The existing aids, however, need to be augmented by additional semantic mapping tools such as statistical term associations, switching vocabularies, enriched "fuzzy" thesauri, "scriptal" micro-lexicons, production rules, and/or heuristics elicited from expert searchers.

The natural language information retrieval interface must also be designed to deal with special problems such as multiple languages, specialty languages, dialects, and professional jargon.[42] To a considerable extent, the CITE experimental R & D system and its operational versions have attempted to address many of the linguistic problem areas outlined in this paper.[43]

CITE represents a domain-independent NL IR interface approach that combines conventional inverted file, Boolean logic with term frequency-based weighted logic, closest-match search strategy and efficient NLP techniques involving "intelligent" stemming,[44] partial syntax analysis, automatic query-to-controlled-vocabulary mapping, look ahead search ambiguity resolution and filtering of combinatorial controlled vocabulary term displays, automatic user feedback processing, and other techniques adapted from applied AI such as domain-specific semantic navigational tools, refined textual pattern matching, and ad hoc expert searcher heuristics. The appendix illustrates several NL user interactions on the CITE system. The last example serves to put things in humbling perspective: The search query NATURAL LANGUAGE PROCESSING automatically picks up the subject heading NATURAL DISASTERS!

Other AI Applications of Direct Relevance to IR

In addition to natural language processing techniques in general and NL DBMS front-ends in particular, the following artificial intelligence areas are perceived by this author to be of direct relevance to IR:

1. *Expert systems.* To the extent that rule-based expert systems are instances of very high level programming tools that allow the expression of order-independent rules instead of ad hoc pieces of order-dependent conventional program code, they can be of benefit in NL IR interface development in capturing trained searcher expertise as well as codifying broad linguistic processing rules. Efficient microcomputer implementations of rule-based expert systems are becoming increasingly available.[45]
2. *Intelligent information management.* Intelligent information management involves the analysis of the interrelationships among multiple databases, information sources, user behavior, including observation of past actions and codified procedures, in order to develop rules for enhanced data retrieval and management.[46] Elements of this approach have been utilized in information retrieval—e.g., in the PAPERCHASE system[47] and the Syracuse University SUPARS Project. The systematic development and utilization of intelligent information management techniques should benefit IR systems in the future.
3. *AI knowledge representation techniques.* In general, AI researchers have found that amassing large amounts of knowledge rather than sophisticated reasoning techniques are responsible for the power of expert systems.[48] The knowledge encoded in conventional controlled vocabularies can be potentially augmented via rule-based relationships, "ISA" knowledge-representation constructs,[49] or predicate calculus statements and fuzzy logic.
4. *Integrating IR, DBMS, AI, and other technologies.* To date, relatively little work has been done in such integrative R & D. Videotex[50] and CD-ROM systems are perhaps the most promising applications in this category. The latest videotex systems—e.g. The SOURCE and CompuServe—combine sophisticated IR, DBMS, electronic mail, and other technologies, and typically offer full-text inversion and Boolean search, distributed database management as well as menu-driven, tree-structured, user friendly access. CD-ROM database publishing and special-purpose knowledge-base applications similarly combine state-of-the-art IR, DBMS, AI, and overall information-management technologies. The integration of efficient NLP techniques and intelligent computer-assisted instruction[51] capabilities will enable videotex and CD-ROM users, and users of diverse information utilities to augment their own intellectual power by machine intelligence that is perhaps

going to be able to grasp—if not understand—database information content, discover new relationships, synthesize new knowledge, and postulate new hypotheses.

REFERENCES

1. Salton, Gerard, and McGill, Michael. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983; and Sparck-Jones, K., and Kay, M. *Linguistics and Information Science*. New York: Academic Press, 1972.

2. Doszkocs, Tamas E., and Rapp, Barbara A. "Searching MEDLINE in English: A Prototype User Interface with Natural Language Query, Ranked Output, and Relevance Feedback." In *Information Choices and Policies* (Proceedings of the ASIS 42nd Annual Meeting, Minneapolis, Minn., 14-15 Oct. 1979), edited by Roy D. Tally and Ronald R. Deultgen, pp. 131-39. White Plains, N.Y.: Knowledge Industry Publications, 1979; Koll, M., et al. "Enhanced Retrieval Techniques on a Microcomputer." In *The National Online Meeting* (Proceedings of the 5th National Online Meeting, New York, 10-12 April 1984), compiled by Martha E. Williams and Thomas H. Hogan. Medford, N.J.: Learned Information, 1984; and Berstein, L.M., and Williamson, R.E. "Testing of a National Language Retrieval System for a Full Text Knowledge Base." *JASIS* 35(July 1984):235-47.

3. Doszkocs, Tamas E. "Natural Language Processing in Intelligent Information Retrieval." In *Proceedings of the ACM Annual Meeting*, edited by S. Ron Oliver, pp. 356-59. New York: Association for Computing Machinery, 1985.

4. Andriole, Stephen, J., ed. *Applications in Artificial Intelligence*. Princeton, N.J.: Petrocelli Books, 1985.

5. Faloutsos, Christos. "Access Methods for Text." *Computing Surveys* 17(March 1985):49-74.

6. Salton, and McGill, *Introduction to Modern Information Retrieval*.

7. Doszkocs, "Searching MEDLINE in English," pp. 131-39; Koll, et al., "Entrance Retrieval Techniques," pp. 165-70; and Bernstein, and Williamson, "Testing of a Natural Language Retrieval System," pp. 235-47.

8. Salton, Gerard, ed. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

9. Bookstein, A. "Implications of Boolean Structure for Probabilistic Retrieval." In *Proceedings of the 8th Annual International ACM SIGIR Conference* (Montreal, Canada, 5-7 June 1985), edited by S. Ron Oliver, pp. 11-17. New York: Association for Computing Machinery, 1985.

10. Doszkocs, Tamas E. "CITE NLM: Natural-Language Searching in an Online Catalog." *Information Technology and Libraries* 2(Dec. 1983):364-80.

11. ________, "Natural Language Processing," pp. 356-59.

12. Andriole, *Applications in Artificial Intelligence*.

13. Lunin, L., and Smith, Linda, eds. "Perspectives on Artificial Intelligence: Concepts, Techniques, Applications, Promise." *JASIS* 35(Sept. 1984):277-319; Grishman, R. "Natural Language Processing." *JASIS* 35(Sept. 1984):291-96; Cooper, W.S. "Bridging the Gap Between AI and IR." In *Research and Development in Information Retrieval* (Proceedings of the 3d Joint BCS and ACM Symposium), edited by C.J. van Rijsbergen, pp. 259-65. Cambridge: Cambridge University Press, 1984; Sparck-Jones, K. "Natural Language Access to Databases: Some Questions and a Specific Approach." *Journal of Information Science* 4(March 1982):41-48; and Kolodner, J. "Indexing and Retrieval Strategies for Natural Language Retrieval." *ACM Transactions of Database Systems* 8(1983):434-64.

14. Hayes-Roth, Frederick. "Rule-Based Systems." *Communications of the ACM* 28(Sept. 1985):921-32.

15. Fikes, Richard, and Kehler, Thomas. "The Role of Frame-Based Representation in Reasoning." *Communications of the ACM* 28(Sept. 1985):904-20.

16. Politt, A.S. "A 'front-end' System: An Expert System as an Online Search Intermediary." *Aslib Proceedings* 36(May 1984):229-34.

17. Brownston, Lee, et al. *Programming Expert Systems in OPS5.* Reading, Mass.: Addison-Wesley, 1985.

18. Schutzer, Daniel. "Artificial Intelligence-Based Very Large Data Base Organization and Management." In *Applications in Artificial Intelligence*, pp. 251-78.

19. Hendrix, Gary G., and Sacerdoti, Earl D. "Natural Language Processing: The Field in Perspective." In *Applications in Artificial Intelligence*, pp. 149-92.

20. Schank, Roger, and Schwartz, Steven P. "The Role of Knowledge Engineering in Natural Language Systems." In *Applications in Artificial Intelligence*, pp. 193-212.

21. Sparck-Jones, and Kay, *Linguistics and Information Science.*

22. Kay, Martin. "Morphological and Syntactic Analysis." In *Linguistic Structures Processing*, edited by A. Zampolli, pp. 131-234. New York: North-Holland, 1977.

23. Pacak, M.G., and Dunham, G.S. "Computers and Medical Language." *Medical Informatics* 4(1979):13-27.

24. Tomita, Masaru. *Efficient Parsing for Natural Language.* Hingham, Mass.: Kluwer Academic Publisher, 1986, p. xvii.

25. Hendrix, and Sacerdoti, "Natural Language Processing," pp. 149-92.

26. Golden, F.L. *Jest What the Doctor Ordered. A Recording of Medical Humor.* New York: Frederick Fell Publishers, 1949; and Cowan, L. and Cowan, M. *The Wit of Medicine.* London: Frewin, 1972.

27. Charniak, Eugene, and McDermott, Drew. *Introduction to Artificial Intelligence.* Reading, Mass.: Addison-Wesley, 1985.

28. Tomita, *Efficient Parsing.*

29. Hice, Gerald F., and Andriole, Stephen J. "Artificially Intelligent Videotex." In *Applications in Artificial Intelligence*, pp. 295-312.

30. Schmucker, Kurt J. *Fuzzy Sets, Natural Language Computations, and Risk Analysis.* Rockville, Md.: Computer Science Press, 1984.

31. Schank, and Schwartz, "The Role of Knowledge Engineering," pp. 193-212.

32. Cowan, and Cowan, *The Wit of Medicine.*

33. Doszkocs, Tamas E. "From Research to Application: The CITE Natural Language Information Retrieval System." In *Proceedings of the Fifth BCS and ACA SIGIR Conference.* Berlin, Germany: Springer-Verlag, 1983, pp. 251-62.

34. Carbonell, Jaime G. "The Role of User Modeling in Natural Language." In *Applications in Artificial Intelligence*, pp. 213-26.

35. Rich, E. "Natural Language Interfaces." *Computer* (Sept. 1984):39-47.

36. Petrick, S.R. "On Natural Language Based Computer Systems." In *Linguistic Structures Processing*, pp. 313-40; and Schank, and Schwartz, "The Role of Knowledge Engineering," pp. 192-212.

37. Carbonell, "The Role of User Modeling," pp. 213-26.

38. Schank, and Schwartz, "The Role of Knowledge Engineering," pp. 193-212.

39. Doszkocs, "CITE NLM," pp. 364-80.

40. Doszkocs, Tamas E. "AID—An Associative Interactive Dictionary for Online Searching." *Online Review* 2(June 1978):163-73; and Doszkocs, Tamas E., et al. "Analysis of Term Distribution in the TOXLINE Inverted File." *Journal of Chemical Information and Computer Sciences* 16(1976):131-35.

41. Salton, Gerard, et al. "Automatic Query Formulations in Information Retrieval." *JASIS* 34(July 1983):262-80.

42. Cowan, and Cowan, *The Wit of Medicine*; Bove, A., et al. "Hellenic Influence in Medical English." *Med Clin* 83(1984):209-13; and Burnum, J.F. "Dialect is Diagnostic." *Annals of Internal Medicine* 100(June 1984):899-901.

43. Doszkocs, "Natural Language Processing," pp. 356-59.

44. Ulmschneider, John E., and Doszkocs, Tamas E. "A Practical Stemming Algorithm for Online Search Assistance." *Online Review* 7(1983):301-18.

45. Lehner, and Barth, "Expert Systems on Microcomputers," pp. 109-24.

46. Schutzer, "Artificial Intelligence-Based Very Large Data Based Organization and Management," pp. 251-78.

47. Horowitz, G.L., and Bleich, H.L. "Paperchase: A Computer Program to Search the Medical Literature." *New England Journal of Medicine* 305(15 Oct. 1981):924-30.

48. Gevartner, William B. "Expert Systems: Limited but Powerful." In *Applications in Artificial Intelligence*, pp. 125-42.

49. Rada, Roy, et al. "A Medical Informatics Thesaurus." In *Proceedings of the MEDINFO '86 Conference* (26-30 October 1986, Washington, D.C.), pp. 1164-1172. North Holland: Amsterdam, Holland.

50. Hice, and Andriole, "Artificially Intelligent Videotex," pp. 295-312.

51. Fletcher, J. Dexter. "Intelligent Instructional Systems in Training." In *Applications in Artificial Intelligence*, pp. 427-52.