From: AAAI-87 Proceedings. Copyright ©1987, AAAI (www.aaai.org). All rights reserved.

Information Retrieval from Never-ending Stories

Lisa F. Rau
Artificial Intelligence Branch
GE Company, Corporate R&D
Schenectady, NY 12301 USA

Abstract

The System for Conceptual Information Summarization, Organization, and Retrieval (SCISOR) is a research system that consists of a set of programs to parse short newspaper texts in the domain of corporate takeovers and finance. The conceptual information extracted from these stories may then be accessed through a natural language interface.

Events in the world of corporate takeovers unfold slowly over time. As a result of this, the input to SCISOR consists of multiple short articles, most of which add a new piece of information to an ongoing story. This motivates a natural language, knowledge-based approach to information retrieval, as traditional methods of document retrieval are inappropriate for retrieving multiple short articles describing events that take place over time. A natural language, knowledge-based approach facilitates obtaining both concise answers to straightforward questions and summaries or updates of the events that take place. The predictable events that take place in the domain make expectation-driven, partial parsing feasible.

I. Introduction

As more and more information is stored electronically, the need for intelligent methods of accessing this information becomes increasingly critical. In the distant future, we can imagine the computer not as a static repository of information, but as a librarian who not only knows where all the information in the library is kept, but also has read and understood everything in that library. This librarian could instantly absorb all the latest texts, and keep in mind the special interests and needs of a vast number of users.

This work rests on the belief that a truly helpful information system must, in some sense, *understand* what the user is looking for. In order to understand this, the system must minimally *understand* what it knows about. Such a system becomes no longer a storage and retrieval device, but an intelligent and knowledgeable assistant, available to answer questions and provide digested summaries of topics. The work described here is one small step on the way to making that assistant exist.

The System for Conceptual Information Summarization, Organization, and Retrieval (SCISOR) is a research system being developed at GE's Corporate Research and Development Center, that operates in the domain of corporate takeovers and finance. SCISOR is designed to read short articles from newspapers, and answers questions about the domain. This paper describes the preliminary system implementation and objectives of SCISOR, but focuses on some of the features of the domain that motivate the computational approach taken, as well as the system objectives themselves.

SCISOR is designed to increase the usefulness of an automated system to extract information from naturally occurring texts. Below is an interaction a user might eventually have with the SCISOR system:

1. **User:** Did ACE hardware take over the ACME food company?

System: *Yes it did. The ACE hardware company completed the takeover last Friday.*

2. **User:** What were the events in the ACE-ACME takeover deal?

System: *Rumors that ACME was to be taken over started May 13, 1985. The stock rose $5 a share to $65. On May 16, the ACE company announced that it had made an offer to ACME valued at $66 a share. The offer was a cash offer to purchase all 2 million outstanding shares of ACME. After a month of deliberation involving some potential conflict of interest matters, the ACE-ACME deal was completed last Friday.*

3. **User:** What happened yesterday in the HAL Corporation bid for BIG?

System: *HAL Corporation increased its bid for the BIG company yesterday, up $4 to $70 a share. BIG has not commented on the new offer.*

These three exchanges illustrate three capabilities difficult to obtain from traditional information or document retrieval technology [Salton and McGill, 1983]. The first illustrates the ease of obtaining answers to straightforward questions by asking the question in natural language. The second illustrates a summary of an event sequence whose description spanned multiple articles occurring over time. In the third exchange, a simple update of an ongoing story, presumably being followed by the user, is obtained.

In this paper, some problems with the paradigm of full-text retrieval are isolated to motivate the natural language, knowledge-based approach to information retrieval taken in SCISOR. These problems manifest themselves especially strongly when certain domain characteristics are present. The domain of corporate takeovers and how it exhibits these characteristics are described. This is followed by a discussion of the implementation of the SCISOR system and its current system status.

A. Problems with Full Text Retrieval

The most widely used methods of storing and retrieving information are by storing the full document, and retrieving via either automatically or manually constructed keywords, or full-text search [Salton and McGill, 1983]. Current full-text retrieval systems have three problems. The first problem is with the accuracy of retrieval, which can stand to be improved. Recent studies have shown that one-fifth of relevant articles are retrieved, and only three-fourths of those retrieved are judged by users to be relevant [Salton, 1986]. This problem may disappear with the advent of massively parallel machines to perform document retrieval, where some promising results have already been obtained [Stanfill and Kahle, 1986].

The second problem is that full-text retrieval systems are designed to have as output a document, when a user might really desire certain kinds of information *from* the document. For example, full-text retrieval is a poor method of extracting simple, factual information from text. This is because users must isolate a document or set of documents that contains an answer to a question through the construction of a potentially complex combination of keywords. Then the relevant passage or passages must be read before the sought-after information is obtained. A much more natural and time-efficient technique is to give the user the option of posing questions to the system in natural language, and having the system respond, not with the original text, but with an answer to the question posed, as illustrated in the first exchange above.

Full-text retrieval systems are also incapable of relating articles to one another. Thus, it is impossible to ask a system for a *summary* of a situation that unfolds over a period of time, potentially involving multiple documents, as shown in the second exchange given previously. The user must retrieve the entire series of articles and read each one to obtain an understanding of all that has gone on. Given that most articles consist of background information potentially known to the reader, simply restricting the response to new information would be very helpful. The best scenario, however, is to give the user the ability to retrieve a preprocessed summary of events in any given situation.

Three features of the articles that appear in the corporate takeover domain in particular make it an appropriate medium for replacing traditional IR techniques with a natural language, knowledge-based approach. Note that these points apply regardless of the method of document retrieval used. That is, the problems still exist whether document retrieval is performed extremely quickly with a highly parallel machine as described in Stanfill and Kahle [Stanfill and Kahle, 1986], with automatic or manually constructed topic indices, or with keyword or free-text search for lexical items. The problems are the following:

1. Most of the content of input articles is NOT new information, but a rehashing of old information. Thus although articles that are relevant may be retrieved with IR technology, the user will be interested in only a small part of that retrieved information.
2. Related to the above point, the frequent rehashing of past events that occur in news stories makes retrieval of multiple articles dealing with the same events likely. This redundancy in retrieval can increase the time a user spends finding the information desired.
3. Events that take place in the domain are not self-contained in singular articles as in the earthquake or terrorist domain used in IPP [Lebowitz, 1983], or the banking telex domain in TESS [Young and Hayes, 1985], for example. Rather, stories continue over long periods of time. Even assuming an IR system that was totally accurate in retrieving the entire series of small articles updating and modifying the events that occurred, the user would still have to read all the articles in the correct order to obtain an understanding of what had transpired.

B. AI information retrieval

Recently, there have been some limited successes in the development of AI systems to parse partially and to understand short texts in constrained domains [DeJong, 1979, Lebowitz, 1983, Kolodner, 1984, Young and Hayes, 1985]. However, work in effectively *accessing* these knowledge bases has not been as successful. For example in Young and Hayes [Young and Hayes, 1985], the information cannot be accessed after it has been understood except through a traditional database front-end, or by direct examination of the conceptual, frame-like representation. In CYRUS, a natural language question-answering component allows queries of the knowledge base. However, it is not always guarenteed to answer correctly due to the reconstructive nature of its retrieval. Although SCISOR has not been tested with large numbers of documents and questions, it is hoped that it will demonstrate some of the uses that can be made with automatically extracted conceptual information from text when that conceptual information is combined with a powerful and robust method of spontaneous retrieval.

Summary:

Full-text retrieval: Full-text systems are inappropriate for certain tasks in certain domains such as the corporate takeover domain. Answers to simple questions and summaries of a series of events are two examples of such tasks.

AI approaches: AI approaches have successfully demonstrated the feasibility of partially understanding free text in constrained domains. No system yet, however, has demonstrated both reliability of information retrieved and accessibility of stored information.

II. The Domain

SCISOR operates in the world of corporate takeovers and takeover attempts: at present, relevant articles are taken directly from news sources and manually "fed" to the system. Typically articles in the domain that appear in the business section of newspapers or the Wall Street Journal are between one and three paragraphs long. Much of the information in the articles is frequently a rehash of the events that may have taken place previously in the current takeover deal. The events in the domain are generally quite predictable, but typically take place over long periods of time.

For example, after a company has made an offer to take over another company, it may be months before the situation is resolved. During this time, a number of predictable developments may arise, such as legal complications, other suitors entering the bidding, or an increase or withdrawal of the initial offer. The predictable nature of the events in the domain, along with the long intervals between initiation and resolution of events, makes the corporate takeover world a rich, but constrained, domain to experiment in.

The following is a typical input to the SCISOR system:

> Group Offers to Sweeten Warnaco Bid
>
> April 8 - An investor group said yesterday that it is prepared to raise its cash bid for Warnaco, Inc. from $40 a share to at least $42.50, or $433.5 million, if it can reach a merger agreement with the apparel maker.
>
> The California-based group, called W Acquisition Corp., already had sweetened its hostile tender offer to $40 a share from the $36 a share offered when the group launched its bid in mid-March.

Notice how only the first half of this story gives new information to the ongoing sequence of events. Both the initial bid and the first increase of the offer in the second half are references to previous events. In order to deal with these references correctly, an information system must retrieve the recorded events and recognize that the recorded events are the same as the references to them. Then a user may obtain an update of the story containing only the new information.

III. Implementation

Figure 1 illustrates the architecture of the SCISOR system. Each of the boxes in Figure 1 will be briefly discussed. First, newspaper stories, or questions about the stories that deal with corporate takeovers, such as the above, are interpreted using the TRUMP (TRansportable Understanding Mechanism Package) parser and semantic interpreter [Jacobs, 1986]. Questions asked by users are parsed with the same understanding mechanism as is used for the input stories. They are stored along with the stories, for future user modeling, and to enable the system to answer a user's question when the answer comes along, if it was not known to the system at the time it was posed. After answers to input questions have been retrieved, they are passed to the KING (Knowledge INtensive Generator) [Jacobs, 1985] natural language generator for expression.

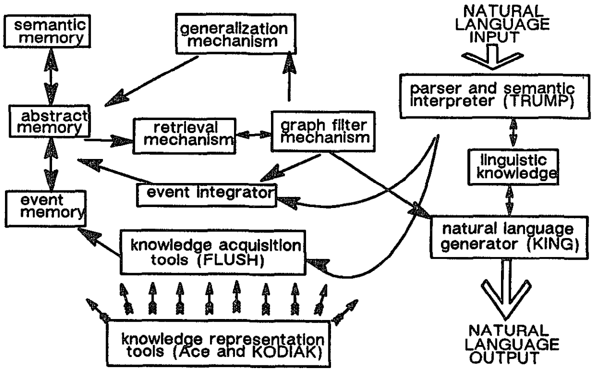

Figure 1: SCISOR System Architecture

These stories are represented in the KODIAK [Wilensky, 1986] knowledge representation language. KODIAK has been augmented with some scriptal knowledge [Schank and Abelson, 1977] of typical sequences of events in the domain. TRUMP and KING were designed to access the same linguistic knowledge base. Linguistic information is represented with KODIAK and the Ace [Jacobs and Rau, 1985] knowledge representation framework.

The story event integration mechanism "fills in" new information in stories with the story summary obtained thus far. It also unifies references of past events in new stories with the previously stored representation of those events, if necessary.

The retrieval mechanism retrieves answers to user's questions or the history of a story being continued in a new input story. For example, upon input of the first sentence of the example story given previously, the history of the initial W Acquisition bid and increased bid for Warnaco are retrieved. The retrieval mechanism operates by using a form of constrained marker-passing [Charniak, 1983]. In the following discussion, "episodes" will refer to events in stories or questions users may have asked, both of which are stored in the system.

Retrieval occurs as a by-product of the understanding process. As new concepts are instantiated in the system, other instantiated concepts related to one another through the semantic category information in the knowledge base are marked. For example, consider the representation of the question "How much has a company offered to take over an apparel company?" given in Figure 2. Previously

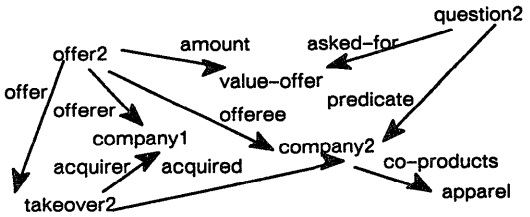

Figure 2: How much are offers for apparel companies in takeovers?

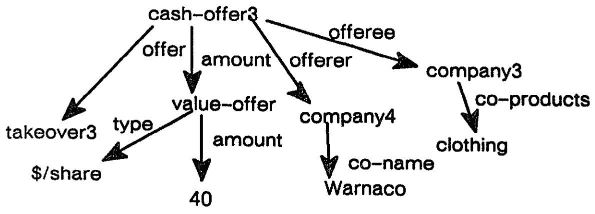

Figure 3: Part of Story Episode

stored in the system is part of a story episode that should be retrieved to give an answer to the question, shown in Figure 3. The instantiation of the concepts in the question (OFFER2, COMPANY1, COMPANY2, QUESTION2, TAKEOVER2, APPAREL2) causes markers to be passed to related concepts, such as CASH-OFFER3, CLOTHING3 and TAKEOVER3. Note how the instantiation of an OFFER causes a more specific kind of offer; a CASH-OFFER, to get marked. Similarly, CLOTHING is marked even though the input question only specified an APPAREL company. This marker-passing to related concepts allows SCISOR to find answers to questions even when the questions ask for information at a different level of conceptual generality than is present in the story.

Intersections of marked concepts occur when a subset of concepts in an episode are marked, and the episodes with the most marks are put into a short-term memory buffer. A filtering process, represented in Figure 1 as the "graph filter mechanism" is then run on these candidates to determine the nature of the match between input question and the answer in a story, for example. All concepts are unmarked periodically.

This two-step retrieval process is very efficient, in that only likely candidates are examined closely. Also, it is very tolerant of erroneous, incomplete or partial input information. This is important in the corporate takeover domain to ensure retrieval of previous events even when a new state of affairs may contradict the previous events. For example, SCISOR must find that Warnaco was trying to take over W Acquisition if today W Acquisition announced that it was trying to take over Warnaco. A more complete description of the retrieval mechanism may be found in Rau [Rau, 1987].

The generalization mechanism is being designed to notice new trends in the domain, through automatic detection of multiple cases of similar situations not previously seen.

A. Question Answering

In SCISOR, the processes that find the approximate location of an answer to a user's question and the processes that determine what the answer should be are separate. A great deal of work has been done on the second problem, most notably by Lehnert [Lehnert, 1978]. Determining an appropriate answer to a user's question, given that the context in which the user's question was posed is already known, is a separate process from the initial retrieval of a context in which to search for an answer. This initial retrieval of a context is the event retrieval this paper briefly describes.

SCISOR also is limited in the kinds of questions the system can answer. Currently the SCISOR system is capable of answering only questions about information explicitly stored in the knowledge base. Any information that potentially could be *reconstructed* or *inferred* (in the sense of Kolodner [Kolodner, 1984]) from information stored in the knowledge base is not available. The line between what is explicit in a story and what can be deduced from that story is not sharp, because some amount of "figuring" must go on to obtain any reasonable understanding of the story. To obtain this understanding, SCISOR computes something similar to a maximally complete inference set [Cullingford, 1986] as the set of information present explicitly in articles and inferred from the context and other world knowledge. Anything in that understanding can be directly retrieved.

For example, SCISOR is able to answer the question "What company was sold for $3 billion?" without preindexing a story containing that information by AMOUNT-OF-SALE. However the system cannot answer the question "Which companies have been taken over more times than they have taken over other companies?" for example, because an answer would require counting all the times a company has been taken over and has taken over other companies, comparing these two numbers, and repeating the process for every other company in the knowledge base. Such procesing capabilities may be added in the future.

B. System Status

SCISOR is implemented in Common Lisp, and it is used on VAX computers and Symbolics and SUN workstations. The TRUMP parser and semantic interpreter has not yet been tested with a large grammar or vocabulary, but in these early stages it has been relatively easy to customize. On the SUN-3 it processes input at the rate of a few seconds per sentence, including the selection of candidate parse and semantic interpretation. The KING natural language generator was implemented in Franz Lisp, and at this writing has not yet been converted to Common Lisp to run with TRUMP.

The system has a dozen or so stories stored in the knowledge base. Hundreds of semantic concepts and domain vocabulary are also present. About a dozen questions are answered by the system. It has not yet been tested on a large number of documents. The tests that have been performed so far, however, are quite promising. Before any definitive claims can be made about the ultimate usefulness of this type of system, the system must be tested with a very large sample of documents in real information retrieval tasks. The next stage of the project will include such tests.

IV. Summary and Conclusions

In certain domains and for certain tasks, the traditional output of document or full-text retrieval systems (i.e., documents) is inappropriate for the task. Obtaining a piece of information, a summary, or an update are examples of such tasks. In these cases, providing the information desired may be more helpful than providing the original full-text source. This is especially true when the full-text sources that contain the information desired span multiple documents.

The information desired may span multiple documents when the events in the domain take place over time, as they do in the world of corporate takeovers. To obtain a summary of a typical takeover using document retrieval techniques, one would have to be able to retrieve all the articles dealing with each event that took place, put them in the correct order and read them all. Also, when a domain deals with newspaper articles in particular, writers frequently include in new articles descriptions of events that have been described in previous articles. Thus, a search for information about any given event in the domain will find all articles that refer to that event. A user would have to peruse all these articles before being satisifed that nothing relevant had been missed.

SCISOR is an experiment in the utility of *understanding* short inputs to increase the usefulness and accuracy of an information retrieval system. The corporate takeover domain has proven to be well constrained. The event sequences that occur are highly predictable, making understanding of the stories in context feasible. Moreover, the unfolding of the stories over time makes the natural language, knowledge-based approach particularily well motivated.

References

[Charniak, 1983] E. Charniak. Passing markers: a theory of contextual influence in language comprehension. *Cognitive Science*, 7(3):171–190, 1983.

[Cullingford, 1986] R. E. Cullingford. *Natural Language Processing: A Knowledge-Engineering Approach.* Rowman and Littlefield, Totowa, NJ, 1986.

[DeJong, 1979] G. DeJong. *Skimming Stories in Real Time: An Experiment in Integrated Understanding.* Research Report 158, Department of Computer Science, Yale University, 1979.

[Jacobs, 1985] P. Jacobs. *A knowledge-based approach to language production.* PhD thesis, University of California, Berkeley, 1985. Computer Science Division Report UCB/CSD86/254.

[Jacobs, 1986] P. Jacobs. Language analysis in not-so-limited domains. In *Proceedings of the Fall Joint Computer Conference*, pages 247–252, IEEE Computer Society Press, Washington, DC, November 1986.

[Jacobs and Rau, 1985] P. Jacobs and L. Rau. Ace: associating language with meaning. In T. O'Shea, editor, *Advances in Artificial Intelligence*, pages 295–304, North Holland, Amsterdam, 1985.

[Kolodner, 1984] J. Kolodner. *Retrieval and Organizational Strategies in Conceptual Memory: A Computer Model.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1984.

[Lebowitz, 1983] M. Lebowitz. Generalization from natural language text. *Cognitive Science*, 7(1):1–40, 1983.

[Lehnert, 1978] W. G. Lehnert. *The Process of Question Answering: Computer Simulation of Cognition.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1978.

[Rau, 1987] L.F. Rau. Knowledge organization and access in a conceptual information system. *Information Processing and Management, Special Issue on Artificial Intelligence for Information Retrieval*, Forthcoming(Summer), 1987.

[Salton, 1986] G. Salton. Another look at automatic text-retrieval systems. *Communications of the Association for Computing Machinery*, 29(7):648–656, 1986.

[Salton and McGill, 1983] G. Salton and M. McGill. *An Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

[Schank and Abelson, 1977] R. C. Schank and R. P. Abelson. *Scripts, Plans, Goals, and Understanding.* Lawrence Erlbaum Associates, Halsted, NJ, 1977.

[Stanfill and Kahle, 1986] C. Stanfill and B. Kahle. Parallel free-text search on the connection machine system. *Communications of the Association for Computing Machinery*, 29(12):1229–1239, 1986.

[Wilensky, 1986] R. Wilensky. Knowledge Representation - A Critique and a Proposal. In J. Kolodner and C. Riesbeck, editors, *Experience, Memory, and Reasoning*, pages 15–28, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Young and Hayes, 1985] S. Young and P. Hayes. Automatic classification and summarization of banking telexes. In *The Second Conference on Artificial Intelligence Applications*, pages 402–208, IEEE Press, 1985.