Creating Segmented Databases From Free Text for Text Retrieval

Lisa F. Rau and Paul S. Jacobs
Artificial Intelligence Laboratory
GE Research and Development Center
Schenectady, NY 12301 USA

Abstract

Indexing text for accurate retrieval is a difficult and important problem. On-line information services generally depend on "keyword" indices rather than other methods of retrieval, because of the practical features of keywords for storage, dissemination, and browsing as well as for retrieval. However, these methods of indexing have two major drawbacks: First, they must be laboriously assigned by human indexers. Second, they are inaccurate, because of mistakes made by these indexers as well as the difficulties users have in choosing keywords for their queries, and the ambiguity a keyword may have.

Current natural language text processing (NLP) methods help to overcome these problems. Such methods can provide automatic indexing and keyword assignment capabilities that are at least as accurate as human indexers in many applications. In addition, NLP systems can increase the information contained in keyword fields by separating keywords into segments, or distinct fields that capture certain discriminating content or relations among keywords.

This paper reports on a system that uses natural language text processing to derive keywords from free text news stories, separate these keywords into segments, and automatically build a segmented database. The system is used as part of a commercial news "clipping" and retrieval product. Preliminary results show improved accuracy, as well as reduced cost, resulting from these automated techniques.

1 Introduction

Natural language technology for text processing has advanced rapidly in recent years, to the point where text processing programs can accurately extract structured information from free text in constrained domains [1]. In the long term, this capability will lead to the commercial realization of systems like SCISOR [2] that retrieve conceptual information *in place of* text in response to a natural language query. In the shorter term, natural language techniques can improve conventional information dissemination and retrieval systems by automatically extracting key information from free text.

This paper reports on a system, known as NLDB for Natural Language DataBase, that significantly improves on a common model of text retrieval—the use of "keyword" indices to search a text database. The advance has widespread applications, particularly in on-line information services, and is part of a commercial product offering. Yet it uses some of the latest in natural language technology, including context-based lexical analysis of text [3], a semantically-driven pattern matcher [4], and a system for recognizing proper names in free text [5].

The technical approach aims at *incrementally* improving upon existing text retrieval products. We leave untouched the current retrieval methods, which we have found to be quite difficult to replace in practice because of the significant investment in existing text databases, software applications, and retrieval products. Instead, we have been focusing on improving the *content* of the databases used. These improvements come from two major areas. First, natural language text processing methods allow for more consistent, and probably more accurate, assignment of "keywords" or content indicators to news stories. Second, these methods allow the differentiation of keywords into "segments", for example, distinguishing proper names from other content indicators, companies mentioned in passing from those actively involved in mergers or other events, and locations of companies from locations of stories.

The results of the approach are being incorporated into a commercial news "clipping" and retrieval product, and evaluation of the improvements is underway. The first phase of the evaluation has shown that the automated indexing methods are more complete and accurate than human indexers. The second phase, which depends on a more difficult comparison of user requests in both the old and new systems, should produce results in time for the SIGIR conference.

This report will contrast the original approach to the new indexing and retrieval strategy, describe the text analysis methods used, and analyze the performance data obtained so far.

2 The Old Way

Many traditional text retrieval systems, including the one we sought to improve, operate by using keywords to index the texts. These keywords may be subject codes, words from the text, words that describe the content of the text, or names. The creator of these keywords (the "indexer") is generally free to include any word that might be of use in the subsequent retrieval of the source text. Figure 1 illustrates a simple example of a text and its associated keywords.

In addition to the expense, inconsistency, and inaccuracy of human indexers, this approach has the problem that keywords lose their context when isolated. For example, the word PRIME can be a company name, as in Prime Computer, or a subject indicator, as in prime rate[1], or a content word, as in prime beef[2].

Another disadvantage of this approach is that users cannot be very specific about the kind of texts they are interested in. For example, a user may want to determine the target of a corporate takeover by a company American Exploration Company, but be forced to specify:

```
(American Exploration Company   OR   AEC)
                 AND
(takeover   OR   acquire   OR  purchase)
```

Along with potentially relevant texts, they may find texts about AEC acquiring a new phone system. They may miss a text that refers to the acquisition if the indexer has failed to identify the text with the *takeover* keyword.

3 The New Way

Using a set of natural language text processing tools [6], we built a news categorization system, NLDB, that automatically assigns keywords, addressing many of the problems with the current system. The evaluation and extension of NLDB is ongoing, but the current version addresses problems with retrieval accuracy through four basic methods:

1. **Segmented Databases:** The keywords are divided into segments that create *conceptual categories* of keywords. For example, a company-name segment will contain Prime, meaning **Prime Computer**, where as a
text-subject segment will contain **prime** meaning **prime rate**.

2. **Text Category Browsing:** Texts are divided into a hierarchy of topic areas for use in assigning words to their proper segments. These topic areas form a conceptual hierarchy that the end user may browse, either to look at representative texts in an area or to bring up a "query template" that reflects specific information contained about texts in that topic area.

3. **Query by Relationship:** Information is extracted from the text using natural language, so that *relationships* between keywords can be used.

4. **Special Name Handling:** Names are automatically extracted from the text, and all possible ways of referring to these names are created, standardizing and making complete the name segments.

This paper briefly describes three technical aspects of this project: the representation of the lexicon used for processing the texts (Section 3.2.1), company name handling mechanism (Section 3.2.2), and the pattern-matching mechanism to perform story categorization (Section 3.2.3). An example of the text category browsing mechanism is given in Section 4 and illustrated in Figure 4. The "query-by-relationship" capability involves quite complex natural language text processing techniques beyond the scope of this paper; details can be found in [7].

We will now contrast the original database with the enhanced, segmented NLDB version.

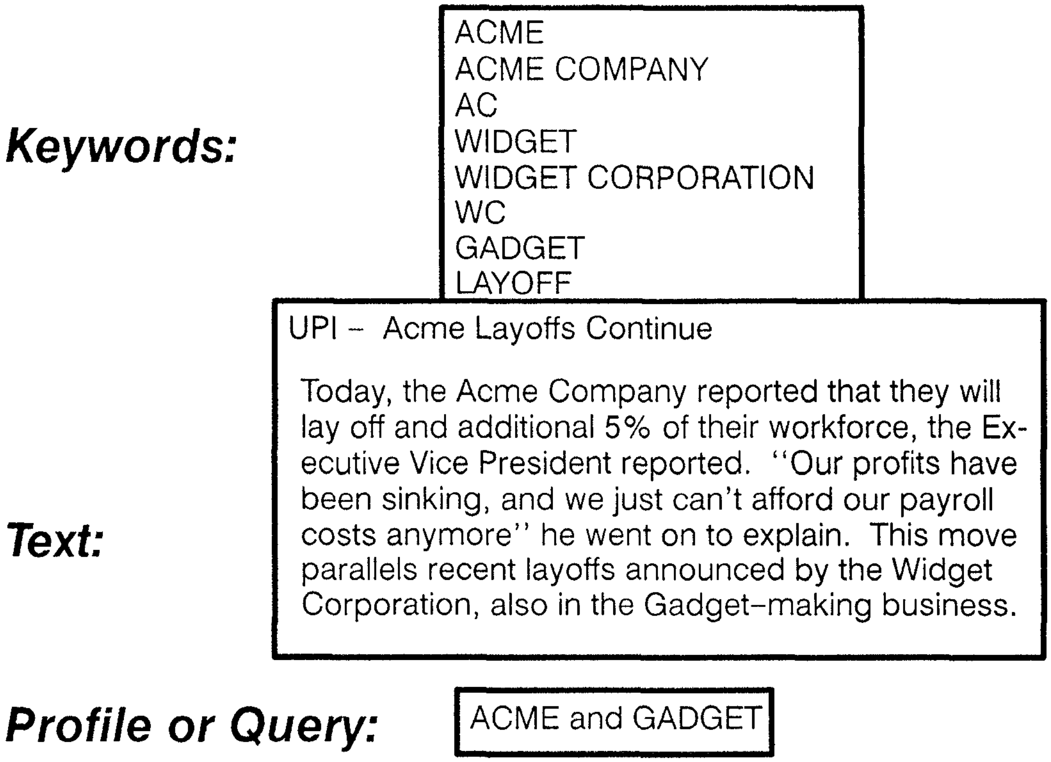
3.1 Segmented Databases

Although most commercial text databases have separately searchable segments that can restrict search queries, these are not typically used to their best effect. In library systems, such segments often distinguish the title and authors of an item from content indicators, dates, and abstracts. In on-line information services, this facility is seldom used, other than to qualify a query by a certain time period. The main reason for this limited use is that training, compensating, and monitoring human indexers in constructing additional fields requires a large investment. On the other hand, this additional information can only help retrieval, because users can still search all segments if they care to.

In engineering NLDB, we separated common keywords from a text database of business and financial news into five initial segments: company-names, proper-names, topic-areas, industry-segment, and geographic-location. This relatively small change

[1] The prime rate is lowest official rate that US commercial banks charge on business loans.

[2] "PRIME" is a US Dept. of Agriculture rating of the quality of commercial meats.



Today, the Acme Company reported that they will lay off and additional 5% of their workforce, the Executive Vice President reported. "Our profits have been sinking, and we just can't afford our payroll costs anymore" he went on to explain. This move parallels recent layoffs announced by the Widget Corporation, also in the Gadget-making business.

Figure 1: The Application Environment - Before

eliminated many confusions, such as the false positives due to the word `Prime` in `Prime Computer` *vs.* `prime rate`. Automating the population of these segments requires some natural language processing, to recognize the differences between the two uses of the word `prime`.

Figure 2 conceptually illustrates the changes we made to the information retrieval environment *without replacing the existing text database or text retrieval product or interface*.

Figure 3 shows the keyword indices used in two of the (harder) database segments, the topic and industry indicators.

Additional knowledge engineering produced the hierarchy of topic areas of stories, for example `corporate-takeover`, `company-sales`, `joint-agreements`, `management-changes`, and `filings`. We used a pattern matching mechanism [4] to perform the story categorization. The matcher, while simple, has access to a rich morphology, lexicon, and concept hierarchy.

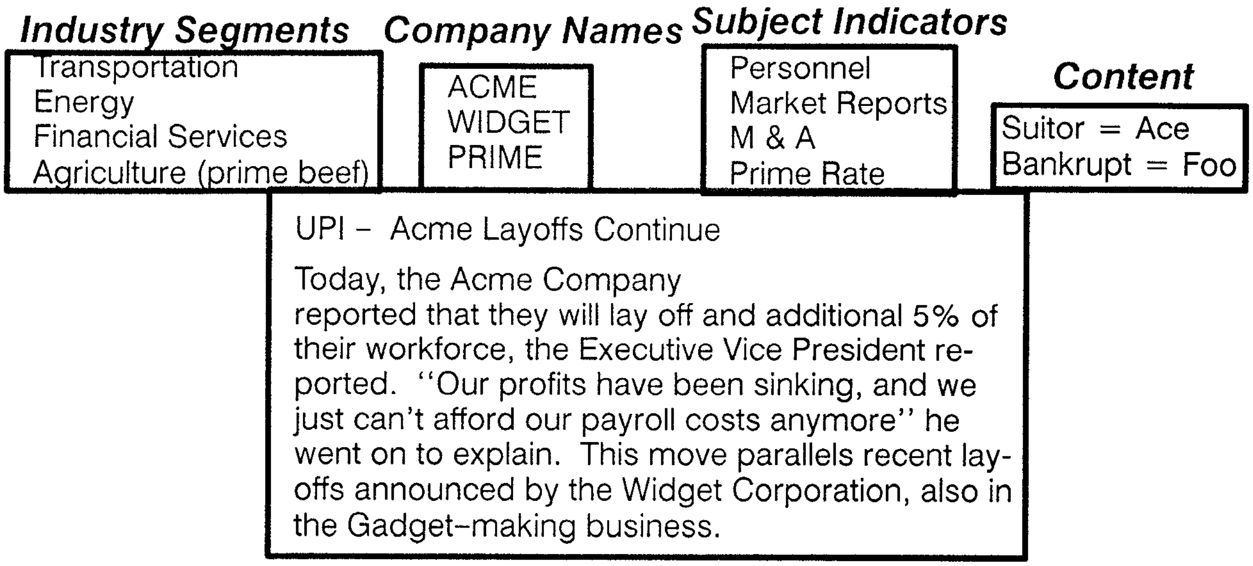
This approach resulted quickly in a richer, more accurate database than the original; however, exploiting the content of this new database also demanded an easy method for users to build queries that use the new segments. The key to this was to use the hierarchy of topic categories to "pop up" query templates, allowing users to select segments that they wanted to search.

3.2 How it Works

The NLDB categorization system uses natural language techniques to recognize important words in each text, discriminate these words when they have different meanings in different contexts, identify proper names and important combinations of words, and perform topic analysis. These tasks do not require the most powerful, computational aspects of natural language processing, but they do use the full power of our lexicon and semantic hierarchy. This section will briefly describe the important aspects of a knowledge base for this type of analysis, along with the algorithms for topic analysis using pattern matching and name recognition.

3.2.1 The Lexicon and Core Hierarchy

The most important function of a lexicon for information retrieval is to capture the similarities among words with similar meanings without being overwhelmed by the ambiguity of words with multiple senses. Our lexicon contains word roots, word senses, sense-based morphology, and syntactic and semantic information for each sense. The current lexicon contains about 8000 unique roots, with an associated hierarchy of about 1000 "core" concepts, 80 affixes (prefixes and suffixes) used in derivations, and several hundred common combinations.



UPI - Acme Layoffs Continue

Today, the Acme Company reported that they will lay off and additional 5% of their workforce, the Executive Vice President reported. "Our profits have been sinking, and we just can't afford our payroll costs anymore" he went on to explain. This move parallels recent layoffs announced by the Widget Corporation, also in the Gadget-making business.

Figure 2: The Improved Application Environment

Industry Segments

advertising
aerospace
agriculture
autos
aviation
banking
beverages
biotechnology
broadcasting
building + material
business + services
chemicals
computers
construction
consumer + products
defense + contracting
educational + services
electronic + publishing
electronics
entertainment
environmental + services
financial + services
food
forestry + products
freight
health + care
industrial + products

insurance
machinery
metals
mining
nuclear + energy
office + equipment
personal + care + products
petroleum + products
pharmaceuticals
photography
plastics
precious + metals
publishing
railroads
real + estate
restaurants
retail
rubber
ship building
telecommunications
textiles
tobacco
toys
travel services
trucks
utilities

Subject Indicators

air + force
antitrust
appointment
bankruptcy
boycott
budget
business
cabinet
capitol
career
chg-naq
commodity
congress
contract
corporate
coup
crime
debt
deficit
democracy

depression
divestiture
dividend
earnings
economy
election
executive change
expansion
export
government
import
inflation
insider + trading
joint venture
labor
lawsuit
layoff
legislation
market
merger

military
money
nasd halt
nasd resume
navy
new product
news
newsbrief
prime + rate
public offering
recession
refinancing
resignation
restructuring
socialism
space
strike
taxes
trade
unemployment

Figure 3: Keywords for Industry and Topic Segments

Knowledge about words centers around the individual senses, which include sense-specific syntactic information and pointers into the conceptual hierarchy For example, the following are the lexical entries for the word *issue*:

```
( issue
   :POS noun
   :SENSES
   (( issue1
      :EXAMPLE (address important issues)
      :TYPE p
      :PAR (c-concern)
)
    ( issue2
      :EXAMPLE (is that the october issue?)
      :TYPE s
      :PAR (c-published-document)
)))
( issue
   :POS verb
   :G-DERIV nil
   :SENSES
   (( issue1
      :SYNTAX (one-obj io-rec)
      :EXAMPLE (the stockroom issues supplies)
      :TYPE p
      :PAR (c-giving)
      :S-DERIV ((-able adj tr_ability)
                (-ance noun tr_act)
                (-er noun tr_actor))   )
    ( issue2
      :SYNTAX (one-obj io-rec)
      :EXAMPLE (I issued instructions)
      :TYPE p
      :PAR (c-informing)
      :S-DERIV ((-ance noun tr_act))   )
    ( issue3
      :SYNTAX (one-obj no-obj)
      :EXAMPLE (good smells issue from the cake)
      :TYPE s
      :PAR (c-passive-moving)   )))
```

The lexicon, by design, includes only the coarsest distinctions among word senses; thus the financial sense of *issue* (e.g., a new security) falls under the same core sense as the latest *issue* of a magazine. This means that some tasks require some additional processing or inference to augment the core lexical knowledge, but avoids many of the problems with considering many nuances of meaning or low-frequency senses. For example, the *progeny* sense of issue, as well as the *exit* sense, are omitted from our lexicon.

Each entry has a part of speech :POS and a set of core :SENSES. Each core sense has a :TYPE field that indicates p for all primary senses and s for secondary senses While we are in the process of enriching the information contained in this field, the general rule is that the interpreter should not consider secondary senses without specific contextual information. For example, the word *yard* can mean an enclosed area, a workplace, or a unit of measure, but only the enclosed area sense is considered in the zero-context.

The :PAR field links each word sense to its immediate parent in the semantic hierarchy. The basic semantic hierarchy acts as a sense-disambiguated thesaurus [8], under the assumption that in the absence of more specific knowledge, word senses will tend to share semantic constraints with the most closely related words. Derivative lexical entries, such as **week-ly**, do "double duty" in the lexicon, so that an application program can use the derivation (i.e., that the root is *week*) as well as the semantics of the derivative form (i.e. that a *weekly* is a type of published document).

The :G-DERIV and :S-DERIV fields mark morphological derivations. G-DERIV (NIL in this case to indicate no derivations) encodes these derivations at the word root level, while S-DERIV encodes derivations at the sense preference level. We have been gradually moving more of the derivations to the sense level on the basis of corpus analysis. For example, the S-DERIV constraint allows *issuance* to derive from either of the first two senses of the verb, with *issuer* and *issuable* deriving only from the *giving* sense.

This application uses only the portions of the lexicon that apply without real parsing, such as the derivational morphology, word combinations, and preferred senses. But this information helps to discriminate different senses of words that might contribute to problems in indexing. In addition, the use of a conceptual hierarchy groups together words with similar meanings, so that these groupings can be used in more general analysis.

3.2.2 Proper Name Identification

Most of the unique words in individual text samples, and by far most of the important words for automatic indexing, are proper names. Recognizing these names is essential both for assigning the names themselves as indices and for using the names in topic analysis. For example, companies buying companies represents a different topic from companies buying other objects; hence the accurate recognition of company names is especially important.

With mixed-case input, a program can easily extract company names by looking backward from a company name indicator (i.e., Incorporated, Corporation, etc.) to the first non-capitalized word. However, we have found that this simple heuristic fails to identify correctly approximately 10% of real company names, and fails entirely with upper-case input. Consider, for example:

1 ... that Structural Research and Analysis Corp., developers of the ...

2. ... process steam to an adjacent Packaging Corporation of America paper mill.
3. The company's investment in Semi-Tech Microelectronics (Far East) Ltd. (STMFE) was reduced from 51 percent to 41 percent.
4. The first qualified institutional buyers to be fully approved for PORTAL are Aetna Life & Casualty; Grantham, Mayo, van Otterloo & Co.; The New England; and Standish, Ayer & Wood Inc.
5. AMERICAN INT'L MEDICAL SIGNS LETTER OF INTENT WITH VIRGINIA-BASED COMPANY
6. UPI - CORPORATE DIVIDENDS Company Period Amount Payable Record

Extracting company names from text is one problem; recognizing subsequent references to a company is another. Because companies appear, disappear, and change their names, accurate identification requires recognizing new names as well as references to known companies. NLDB uses a well-tested module that performs both tasks.

Our solution to name recognition is heuristic–i.e., rather than defining a grammar for names *per se*, it depends on a variety of rules that reflect the nature of references to companies over many millions of words of news stories. The approach requires linguistic knowledge, intelligent pre-processing, a broad lexicon, and tools for handling special cases and exceptions.

Rather than detail the complete algorithm for name recognition (which is described in [5]), here we will discuss some of the difficult cases of handling company names—cases in all-capital input (which is still common in many news sources as well as in headlines for mixed-case stories), and when the name contains an embedded conjunction or preposition.

In the case of input texts that are all upper-case, the problem is determining where a company name begins and ends. We have developed a complex stop condition that uses a combination of an empirically derived stoplist, company word length restrictions, lexical lookup and natural language text segmentation to determine the start of the company name. Our stoplist contains a list of the most common words that immediately precede the start of a company name, determined through exhaustive corpus analysis. This list changes slightly from one corpus to another.

We use our natural language programs [2] to perform some syntactic segmentation of the text. This is necessary to understand certain sentences. For example, without segmentation, the program may extract a company name of `Chile, North Lily Mining`, instead of `North Lily Mining`, as in the following example:

```
As operator of the Taltal Joint Venture in
    Chile, North Lily Mining Co. is pleased to
    announce the receipt of final ore reserve
    estimates on Yolanda Pampa.
```

Some additional details on the method used to perform segmentation of text can be found in [9].

After the program determines the boundaries of a company name, it creates a list of potential referring expressions (called *variations* or *alternations*); alternate ways of referring to that company. Abbreviations for words within the company name are resolved at name recognition time, so all possible abbreviations for words in a company are not specifically included in the representation of the variation. For example, `International Business Machines` will be recognized if it appears as `INTL BUSINESS MACH` in the text, as long as these abbreviations have been defined in the lexicon. Also, overgeneration of variants does not hinder the subsequent natural language analysis or text retrieval.

The results of applying the lexicon and proper name recognition systems is a set of *tokens* corresponding to words or word combinations in the text, with a set of properties, collections of possible senses in the case of core words or unique identifyers with other derived properties (e.g. human or company) for proper names. This stream of tokens forms the input to the pattern matching mechanism, which assigns topic keywords to each text.

3.2.3 Pattern Matching

The pattern matcher uses sequences of words, names, punctuation, or conceptual categories to spot structures in texts that can determine, among other things, the topic of a text. The use of this sort of superficial analysis is more robust and efficient than full linguistic analysis, and has been shown to have high accuracy in other applications [10]. The distinguishing characteristic of our pattern matcher is the ability to take full advantage of the lexicon and name recognition components of the system

Patterns for topical keywords and phrases help to perform topic analysis as well as to "clean up" the input to allow other patterns to match. For example, stories containing patterns like < *company* >...<acquire>...< *company* > are likely to be about corporate mergers. The accuracy of such matching, however, depends on the ability to discard certain parenthetical information from the text, as in *GE, whose acquisition of RCA led to increased earnings, ...*

Patterns in the system are associated with action rules, including, for example, the topic linked to a particular pattern or the action to collapse a "chunk" of information such as an appositive or parenthetical phrase. When the system loads the pattern-activation rules, it indexes each pattern by the lexical features (i.e. the words, lexical categories, roots and concepts) of each of its constituents, distinguishing those that require lexical analysis from the word-only rules. At run-time, the pattern matcher performs the following four operations:

1. It examines each input token (only) once for any features that index pattern tests.
2. Each satisfied pattern test "triggers" its enveloping rule. The satisfied pattern tests are cached so subsequent occurrences of the same input token avoid the feature examination.
3. After all input tokens have been examined, the program matches all triggered rules against the input The matching uses a best-first search algorithm, where the "best" match is one that uses the most pattern constituents and the most input tokens.
4. The system executes the actions of all matched rules.

NLDB currently includes several hundred patterns for 60 topic categories in a hierarchy. These range from very general categories (e. g. *world news* or *diplomacy*) to fairly specific discriminators such as earnings and losses. The patterns cover the complete range of topic keywords ordinarily assigned by human indexers

4 The NLDB System

This approach to automatic database segmentation is implemented as part of a prototype news retrieval system using commercial database technology The same keyword assignments and segments are being used for news routing in a custom clipping product. Because of the strict demands on commercial databases, the natural language components require no changes to the actual database or retrieval methods. Rather, the text processing and segmentation mechanisms run as a "text server" that automatically builds the database from the incoming stream of texts.

The text processing components run on a SUN workstation, creating database update files for a hierarchical database system which runs on a large global mainframe service. The database interface, which also runs on a SUN [3], communicates between the database hierarchy, stored locally, and the central database on the mainframe. Exploiting the increased number of database segments required the construction of additional query "templates", which allow the user to constrain which segments to search for particular information This architecture minimizes the load on the mainframe database as well as insulating the user from the actual query language.

Figures 4 and 5 show "screen dumps" from NLDB. In Figure 4, the user has selected a topic area `management-appointment` to search, and within that topic area, a specific industry group `telecommunications`. This request will retrieve all texts about management changes (hiring, firing, transfer, promotion, and retirement) of personnel in the telecommunications industry.

Figure 5 illustrates the use of natural language extraction of information from text to allow for a query about a relationship between text components. In this case, the user has specified an interest in the corporate takeover of the `Hershey Oil Company` by the `American Exploration Company`. Note the user need not (and often cannot) specify the full name of the company. The company name extraction software has created the acronym `AEC` so this story is retrieved even though `AEC` is not mentioned verbatim.

The screen dumps show portions of the topic hierarchy in the left hand portion, where the user can select a portion of the hierarchy, to look through sample stories, zero in on a specific topic area, constrain the search, or bring up sample query templates to narrow the search to particular segments.

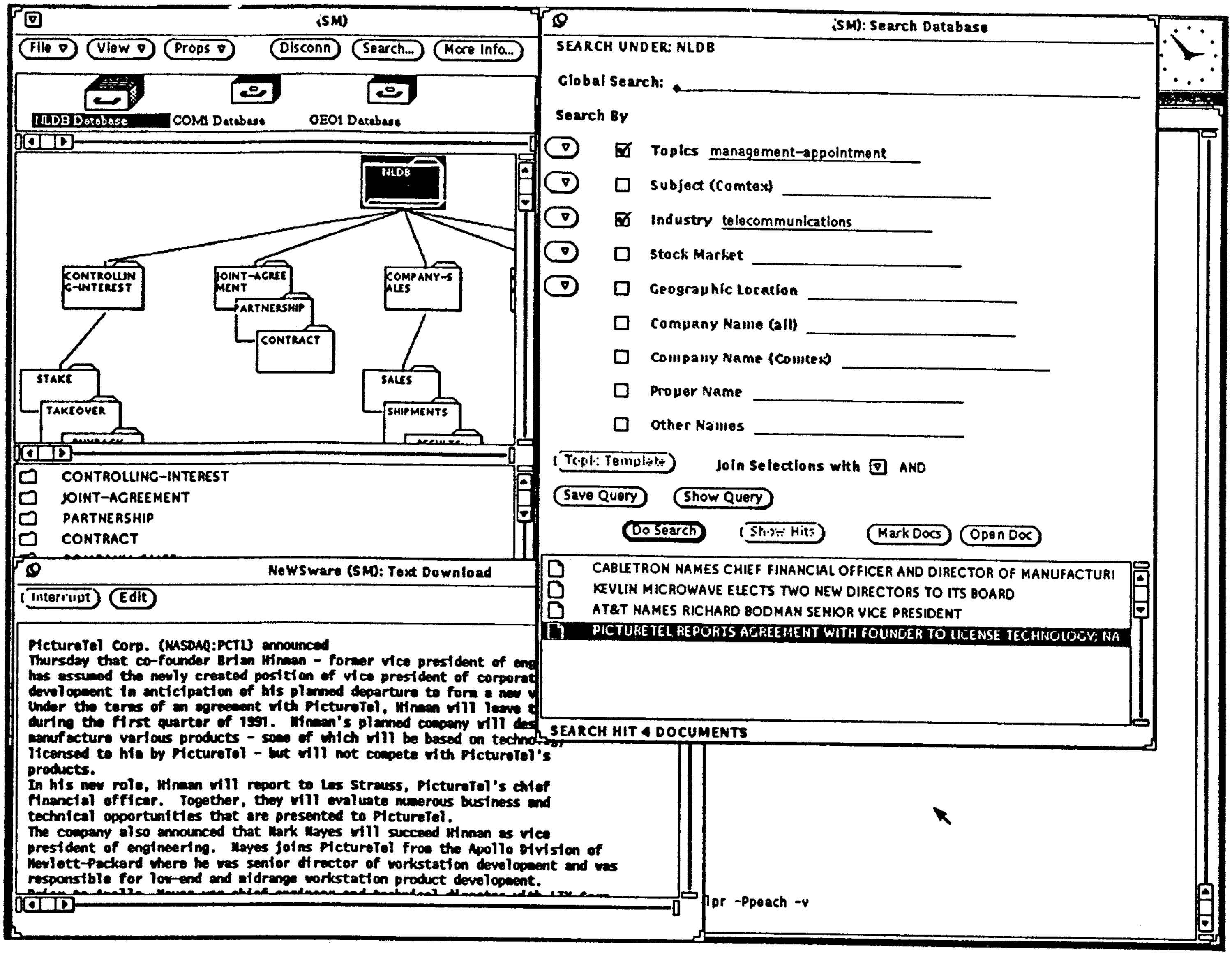
While the commercial success of this sort of system depends ultimately on its ease of use, the quality of the text, and user's perceptions of performance, the contribution to IR research depends on the ability to measure the contribution of key technologies to system accuracy. The next section will discuss a number of performance metrics for the system.

5 Performance Evaluation

The performance of this system seems to depend on four major factors: (1) accuracy of name recognition, (2) accuracy of topic analysis, (3) quality of interface, and (4) the accuracy of retrieval over a representative sample of user goals. We have results for the first two measures, will avoid attempting a quantitative measure of (3), and plan to complete a test of measure (4) in time for the conference. While IR research tends to concentrate on measure (4), it is (1) and (2) that are essential for commercial viability, as they measure the performance of the system against the paid human indexers.

The company name recognition subsystem has been tested on over 10 million words of naturally occurring financial news, with accuracy results for over 1 million words (several thousand stories) with manually-assigned company name indices. The program extracted thousands of company names with over 95% precision compared to a human, and succeeded in extracting 25% more companies than the human, clearly exhibiting better-than-human performance. In addition, the program is consistent in assigning variations, such as

[3] A PC Windows interface is also available, with most, but not all, of the features described here.



SEARCH UNDER: NLDB

Global Search:

Search By

☑ Topics management-appointment

□ Subject (Comtex)

☑ Industry telecommunications

□ Stock Market

□ Geographic Location

□ Company Name (all)

□ Company Name (Comtex)

□ Proper Name

□ Other Names

Join Selections with ▽ AND

- CABLETRON NAMES CHIEF FINANCIAL OFFICER AND DIRECTOR OF MANUFACTURI
- KEVLIN MICROWAVE ELECTS TWO NEW DIRECTORS TO ITS BOARD
- AT&T NAMES RICHARD BODMAN SENIOR VICE PRESIDENT
- PICTURETEL REPORTS AGREEMENT WITH FOUNDER TO LICENSE TECHNOLOGY; NA

SEARCH HIT 4 DOCUMENTS

```
PictureTel Corp. (NASDAQ:PCTL) announced
Thursday that co-founder Brian Hinman - former vice president of eng
has assumed the newly created position of vice president of corporat
development in anticipation of his planned departure to form a new v
Under the terms of an agreement with PictureTel, Hinman will leave t
during the first quarter of 1991.  Hinman's planned company will des
manufacture various products - some of which will be based on techno
licensed to him by PictureTel - but will not compete with PictureTel's
products.
In his new role, Hinman will report to Les Strauss, PictureTel's chief
financial officer.  Together, they will evaluate numerous business and
technical opportunities that are presented to PictureTel.
The company also announced that Mark Mayes will succeed Hinman as vice
president of engineering.  Mayes joins PictureTel from the Apollo Division of
Hewlett-Packard where he was senior director of workstation development and was
responsible for low-end and midrange workstation product development.
```

Figure 4: Sample screen: Segmented Database

acronyms, while human indexers make fairly frequent errors.

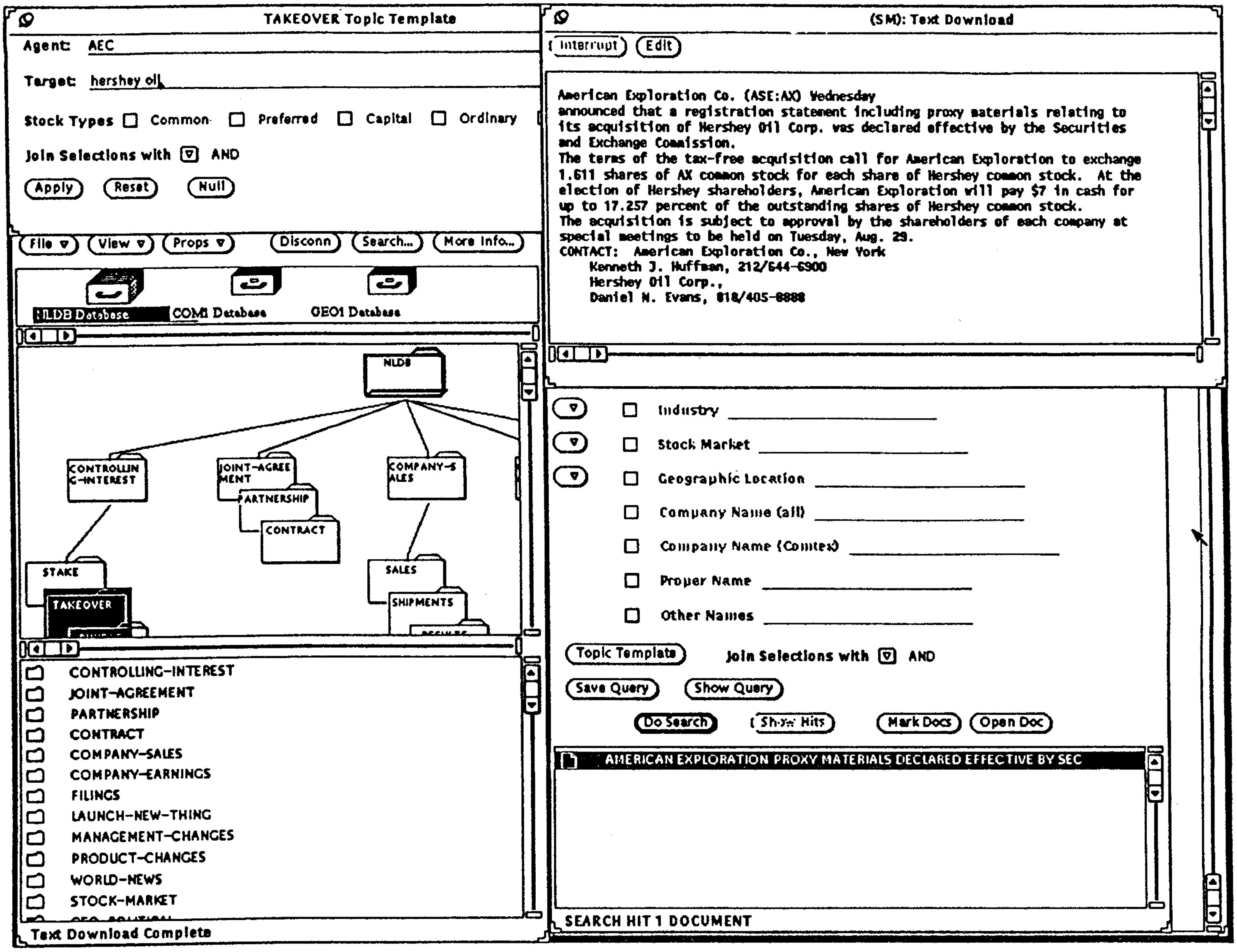
The topic analyzer, run on the same test sample, currently produces slightly less than 90% coverage of human-assigned topics, with over 90% precision. These results tend to improve over time as the patterns are improved with respect to a large corpus. Some of the gaps in coverage result from problems with the name recognition component, but the most common lapses seem to be from very confusing stories and from texts where human indexers apparently erred.

In both of the above measures, as with other similar tests, it is a major problem to determine the accuracy of the test data (i. e. the indices that have been assigned by human beings to the test sample). Since the figures assume 100% accuracy of the test data, they always understate the system performance. While it is clear that 90% is comparable to human accuracy, we do not yet know the variance or upper limit to performance.

The overall performance on all indexing tasks, including topics, industries (which are the hardest to measure), geography, and names, is almost 90% accuracy with about 35% more terms assigned by the program.

We intend to evaluate retrieval performance by having subjects use each version of the program for a fixed set of retrieval goals, measuring both relevance judgements and elapsed time.

In fairness, we expect that retrieval accuracy will have no significant impact on the market success of this sort of system—after all, users do not seem to be demand-



TAKEOVER Topic Template

Agent: AEC

Target: hershey oil

Stock Types □ Common □ Preferred □ Capital □ Ordinary

Join Selections with AND

(SM): Text Download

American Exploration Co. (ASE:AX) Wednesday
announced that a registration statement including proxy materials relating to
its acquisition of Hershey Oil Corp. was declared effective by the Securities
and Exchange Commission.
The terms of the tax-free acquisition call for American Exploration to exchange
1.611 shares of AX common stock for each share of Hershey common stock. At the
election of Hershey shareholders, American Exploration will pay $7 in cash for
up to 17.257 percent of the outstanding shares of Hershey common stock.
The acquisition is subject to approval by the shareholders of each company at
special meetings to be held on Tuesday, Aug. 29.
CONTACT: American Exploration Co., New York
    Kenneth J. Huffman, 212/644-6900
    Hershey Oil Corp.,
    Daniel N. Evans, 818/405-8888

□ Industry
□ Stock Market
□ Geographic Location
□ Company Name (all)
□ Company Name (Context)
□ Proper Name
□ Other Names

Join Selections with AND

AMERICAN EXPLORATION PROXY MATERIALS DECLARED EFFECTIVE BY SEC

SEARCH HIT 1 DOCUMENT

Text Download Complete

Figure 5: Sample screen: Extracting relationships

ing higher accuracy at this point, and it certainly could have been provided by other means. The major benefits of this method are clearly the cost savings and potential improvements in accuracy over human indexers. However, as a scientific endeavor, we feel we should attempt to measure relative accuracy in retrieval as well and to continue to report comparative improvements.

IR researchers often expect comparisons between system performance on this sort of retrieval task and "traditional" IR benchmarks, and are often suspicious (sometimes justifiably) of the high recall and precision results that news categorization and retrieval systems often report. While automated indexing methods [11, 12] show promise even for broader retrieval tasks, the apparent success of more knowledge-based approaches (cf. [13]) in on-line news depends on some distinguishing characteristics of the news retrieval task. Most importantly, on-line news generally requires the particular style of indexing (e.g. names, industries, and topics) described here, because retrieval is only one of many functions of news services. Further, news is simply more restrictive, both in style and content, than general retrieval tasks. One can therefore expect much higher accuracy from knowledge-based approaches in news processing and should not attempt to compare results across tasks.

6 Conclusion

This paper describes a new approach to indexing news items for retrieval. A fully automated system assigns keywords to texts, separates them into database segments, and creates a user-browsable topic hierarchy. This approach is aimed at providing cheaper, cleaner, more accurate indexing of news stories along with greater functionality for end users.

State-of-the-art natural language processing software applies lexical and semantic knowledge, along with a heuristic method of identifying proper names and pattern matching for categorization. The result is improved accuracy in indexing, broader utility, and a means of inserting further advances in natural language processing into commercial text database products.

References

[1] Beth Sundheim. Second message understanding conference (MUCK-II) test report. Technical Report 1328, Naval Ocean Systems Center, San Diego, CA, 1990.

[2] Paul Jacobs and Lisa Rau. SCISOR: Extracting information from on-line news. *Communications of the Association for Computing Machinery*, 33(11):88–97, November 1990.

[3] Susan McRoy. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, (forthcoming), 1991.

[4] Paul S. Jacobs, George R. Krupka, and Lisa F. Rau Lexico-semantic pattern matching as a companion to parsing in text understanding In *Fourth DARPA Speech and Natural Language Workshop*, San Mateo, CA, February 1991 Morgan-Kaufmann.

[5] Lisa F. Rau. Extracting company names from text. In Tim Finin, editor, *Sixth IEEE Conference on Artificial Intelligence Applications.* IEEE Computer Society Press, Miami Beach, Florida, February 1991.

[6] Paul S. Jacobs and Lisa F. Rau. The GE NL-Toolset: A software foundation for intelligent text processing. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, volume 3, pages 373–377, Helsinki, Finland, 1990.

[7] Paul S. Jacobs and Lisa F. Rau. Innovations in text interpretation. *Artificial Intelligence*, 46, In Submission 1991.

[8] E. Fox, J. Nutter, T. Ahlswede, M. Evens, and J. Markowitz. Building a large thesaurus for information retrieval. In *Proceedings of Second Conference on Applied Natural Language Processing.* Association for Computational Linguistics, February 1988.

[9] Paul Jacobs. To parse or not to parse: Relation-driven text skimming. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, pages 194–198, Helsinki, Finland, 1990.

[10] 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.

[11] Karen Sparck Jones and J. I. Tait. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66, March 1984.

[12] D. D. Lewis, W. B. Croft, and N. Bhandaru. Language-oriented information retrieval. *International Journal of Intelligent Systems*, 4:285–318, 1989.

[13] Paul S. Jacobs and Lisa F. Rau. Natural language techniques for intelligent information retrieval. In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval*, pages 85–99. Presses Universitaires de Grenoble, June 1988.