Construe-TIS: A System for Content-Based Indexing of a Database of News Stories

Philip J. Hayes and Steven P. Weinstein

The Construe news story categorization system assigns indexing terms to news stories according to their content using knowledge-based techniques. An initial deployment of Construe in Reuters Ltd. topic identification system (TIS) has replaced human indexing for Reuters Country Reports, an online information service based on news stories indexed by country and type of news. TIS indexing is comparable to human indexing in overall accuracy but costs much less, is more consistent, and is available much more rapidly. TIS can be justified in terms of cost savings alone, but Reuters also expects the speed and consistency of TIS to provide significant competitive advantage and, hence, an increased market share for Country Reports and other products from Reuters Historical Information Products Division.

The Problem

The Historical Information Products Division at Reuters offers a range of textual and numeric database products. Four of the textual products—Reuters Country Reports, Reuters Textline, Reuters Company

Newsyear, and Reuters Newsbank—have among their selling points indexing schemes that aid subscribers in researching and retrieving information (they primarily contain news stories) from them. The indexes allow the subscriber to find information of interest without attempting the near-impossible task (Furnas et al. 1987) of thinking of all the words that could be used to express this information, a task that would be necessary in a retrieval system based purely on a Boolean key-word, full-text search. For example, a subscriber interested in corporate acquisitions could specify the corresponding indexing term and get all the stories about corporate acquisitions without having to worry about whether the news stories retrieved talked about takeovers; mergers; buyouts; or, simply, one company buying or purchasing another. This variability in the words and phrases that can be used to express an idea is the primary reason for the relative inaccuracy of information-retrieval systems based on Boolean key-word techniques (Blair and Maron 1985).

Although valuable to Reuters subscribers, indexing of this kind is expensive because it is highly labor intensive. Human indexing also slows the insertion of the latest news into the textual databases. Given the high volume of news from Reuters busy newsrooms (figure 1, page 48), delays of several days are not uncommon. Moreover, quality targets of consistency and accuracy in the indexes produced are difficult to achieve using a group of human editors. Finally, the work tends to be boring, and as a consequence, staff turnover is undesirably high, with additional unwelcome effects on expense.

Given this background, Reuters became interested in automating the indexing function. It established the following minimum criteria for success: (1) cost reduction (by reduction in the human indexing staff), (2) more rapid availability of indexed news (in minutes or seconds rather than days), and (3) quality comparable to human indexing (an automated system might make some stupid mistakes because it does not understand its input as completely as human indexers, but these errors should be balanced by an overall increase in consistency). Beyond this minimum level of success, Reuters was also interested in the potential competitive advantage and, hence, increased market share that could result from rapid availability of consistently indexed news.

With these goals in mind, Reuters engaged Carnegie Group to develop a news story categorization system. The process was a lengthy one from first discussions in early 1986 to initial deployment in late 1989, but the result was a system that meets all Reuters basic criteria for success: cost reduction, rapidly available indexes, and accuracy comparable to humans. There are also good prospects for translating these advantages into the second level of success: competitive advantage and increased market share.

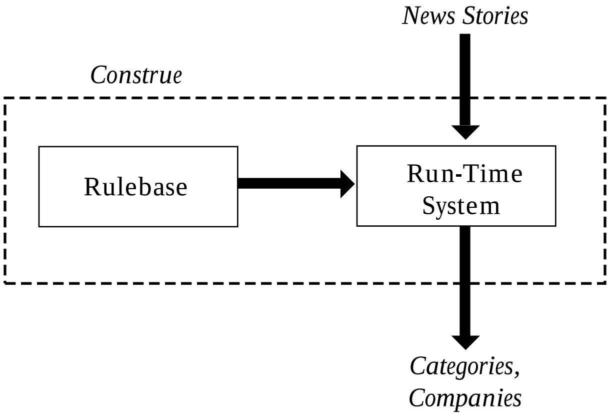

Figure 2. Construe.

The following sections describe Construe, the system initially delivered by Carnegie Group to Reuters in 1988; the technical approach used in Construe; and the eventual deployment of Construe as TIS.

Construe

Reuters first approached Carnegie Group concerning the development of a news story categorization system in early 1986. After strongly encouraging results from a small-scale pilot system developed in the second half of 1986 (Hayes, Knecht, and Celio 1988), Reuters contracted with Carnegie Group for the development of a full-scale news story categorization system, which we called Construe (categorization of news stories, rapidly, uniformly, and extensibly). Development began in April 1987.

The specific goals for Construe were all met or exceeded. They were to

- Accept a broad stream of Reuters news stories, including economic, financial, and general news
- Categorize each story into zero, one, or several of 150 distinct categories (674 distinct categories were delivered)
- Recognize mentions of companies from a database of 10,000 company names (over 17,000 names were delivered).
- Process stories in an average of five seconds (delivered with an average of 4.36 seconds)
- Achieve an average accuracy level of 85 percent (89-percent accuracy level was delivered)

- Be easily maintainable in terms of new and revised category definitions (accomplished by specifying the categorizations performed by Construe through an explicit rule base [figure 2], providing a support environment for rule base development, and training Reuters personnel in rule base development on an apprenticeship basis)
- Operate on standard commercial hardware with the capability for robust 24-hour operation (accomplished through implementation on multiple DEC VaxStations each running Construe and each accepting and returning categories for one story at a time over DECNet from an existing Reuters system that holds a story queue. See figure 5 for more details of this arrangement.).

Construe was delivered to Reuters in April 1988. It is implemented in Lucid Common Lisp.

Between 4 and 8 Carnegie Group personnel worked on Construe, expending a total effort of approximately 6.5 person-years. Of this time, approximately 2.5 person-years were spent on rule development. To supplement the Carnegie Group effort and maximize technology transfer, Reuters also stationed personnel at Carnegie Group. Two experienced journalists spent approximately half a person-year hand categorizing about 20,000 news stories for rule- development and accuracy-testing purposes, and three apprentice rule base developers spent close to one person-year assisting in the rule-development effort. Two of these apprentices went on to maintain the delivered system for Reuters. The total effort was thus about 8 person-years for Construe. Adding 1.5 person-years spent on pilot system development gives a total of 9.5 person-years.

Subsequent to its delivery, Carnegie Group generalized Construe into a package of reusable technology, which is being marketed under the name of text categorization shell (TCS) (Hayes et al. 1990). TCS has the same basic structure and functions as Construe. The major functional difference is that TCS is not tied to news stories with their structure of headline, dateline, body, and so on, but can deal with texts of arbitrary structure. Based on the rate of rule development toward the end of the Construe project, we estimate that recreation of the Construe rule base would take a little under one person-year. Adding another person-year for communication and integration code specific to the Reuters environment means that with TCS, Construe could be produced in about 2 person-years compared to the 8 person-years it took from scratch.

The measurement of accuracy for Construe uses the standard information-retrieval measures of recall and precision (Salton and McGill 1986). *Recall* is the percentage of times that a particular category

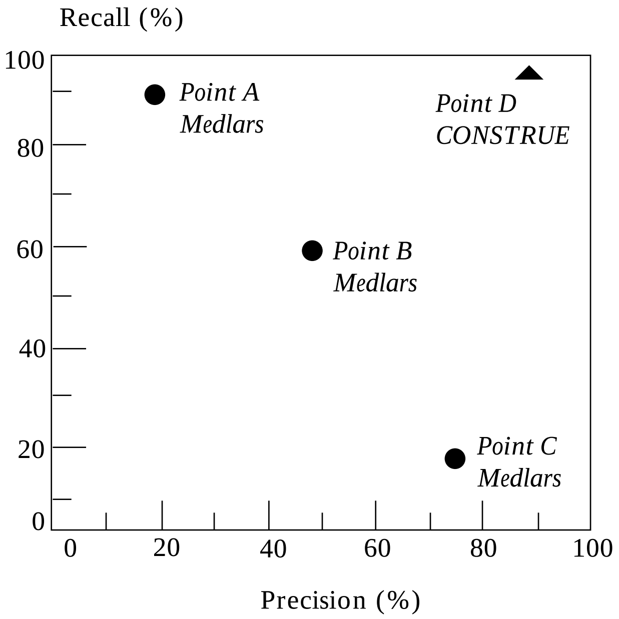

Figure 3. Comparative Recall and Precision Figures for Medlars and Construe.

should have been assigned to a story and was, in fact, assigned; *precision* is the percentage of category assignments actually made that were correct. High recall means few false negatives, and high precision means few false positives. The single accuracy figures mentioned previously represent an average of recall and precision. Recall and precision for Construe were 94 percent and 84 percent, respectively, measured on a set of 723 stories not previously processed by the system and not previously examined by Construe developers.

As figure 3 shows, typical results for key-word Boolean retrieval systems, as represented by the Medlars Information Retrieval System (Salton and McGill 1986), do not simultaneously allow recall and precision to rise much above 50 percent (point B). However, one measure can be traded against the other: A loosely specified search (point A) will find most of what is required (high recall) together with a lot of irrelevant material (low precision); however, a tightly specified search (point C) will miss most of the relevant material (low recall) and find little that is irrelevant (high precision). There are two basic reasons for this poor accuracy: First, a user might not think of the words that were actually used by the author of a relevant text. For example, a user might specify “takeover” to find texts about one company buying another, but the author of a relevant text might have only used “acquisition.” (See Furnas et al. [1987] for an indication of the pervasiveness and difficulty of this problem.) Second, the method does not take into account the way words relate to each other or are contextually modified. “Acquisition” could just as easily refer to the purchase of a piece of capital equipment as the takeover of another company.

The results produced by Construe, point D, represent a much higher level of accuracy than is obtainable by Boolean key-word techniques. The task of Construe differs from Medlars: In Medlars, the users spontaneously generate queries, but in Construe, the system recognizes a fixed and predetermined set of categories. However, the comparison is still appropriate. The graph in figure 3 represents the kind of performance a user would experience if trying to emulate one of Construe’s categories through a spontaneous query. Construe’s higher level of accuracy reflects the knowledge-based framework it uses, as described in Technical Approach.

We do not have specific measurements for the performance of the human indexers employed by Reuters, but informal observations suggest that recall is in the low 90-percent range, with precision in the upper 90-percent range (the human indexers rarely put a story into the wrong category but sometimes omit categories from stories that should be in several). Thus, the performance of Construe is comparable to existing human performance.

The set of categories handled by Construe includes 135 economic categories (mergers and acquisitions, corporate earnings, money–foreign exchange, interest rates, various commodities, various currencies, and others) and 539 proper name categories (people, countries, international organizations, and stock exchanges). Reuters devoted a considerable amount of time to creating this categorization scheme, in deciding both what the category labels would be and which types of stories each should be assigned to. The automatic approach allowed the categorization scheme to be more fine grained than Reuters had been able to achieve through human categorization and was targeted to the needs of specific user groups. For example, the currency categories were intended not merely to identify stories that referred to particular currencies but also to pick stories of interest to currency traders.

The distinctions that needed to be made to put stories in the correct category were often subtle. For example, Reuters decided to assign a currency category when the currency in question was the one fluctuating and not when it was the standard against which another was being measured. Thus, a story about the price of the dollar against the yen would get the dollar category label but not the yen label. Moreover, a currency category was only to be assigned when the discussion of the currency rates was the main point of the story and not if the performance of a currency was cited as background, say, as a reason for a change in a leading economic indicator.

Proper name categories also posed some interesting challenges. It was straightforward to identify all cases in which proper name categories might be relevant based on whether the proper name in question was mentioned, which led to a high average recall score of 98 percent. However, average precision for these 539 categories was only 84 percent. The lower-precision scores occurred because Reuters only wanted proper name categories assigned to stories in which the person, place, organization, or exchange was the focus rather than just a mention. The 135 economic categories, however, tended to present greater problems for recall (average 89 percent) than for precision (average 92 percent) because of the typically large variety of phrases that could be used to express them.

Technical Approach

The need for high accuracy made the Boolean key-word techniques previously discussed inappropriate for Construe. The need to deliver a practical system that met stringent execution-time requirements and was straightforward to maintain made deep natural language understanding techniques equally inappropriate.

Construe instead follows a middle course by using shallow, knowledge-based techniques. We saw the trade-off between speed and accuracy mediated by the depth of processing as the central design issue for Construe and, indeed, for all automatic processing of the extremely large volumes of text that occur in real-world, text-based applications. We adopted a minimalist strategy with respect to the depth of processing and were pleasantly surprised by how much can be accomplished by simple semantic techniques. These techniques allow Construe to provide much higher accuracy than key-word techniques, but by stopping short of deep understanding techniques, they can still provide enough speed and sufficiently low knowledge base creation and maintenance effort to be practical for commercial applications. In this regard, Construe has goals that are similar to systems such as Frump (Dejong 1982) and Scisor (Jacobs and Rau 1988) but is able to remain shallower than these systems because its task—categorization—is simpler than theirs—summarization. The technical approach adopted in Construe to achieve this balance is summarized in figure 4. After an initial translation of news stories into an internal format, there are two main processing steps: concept recognition and categorization rules.

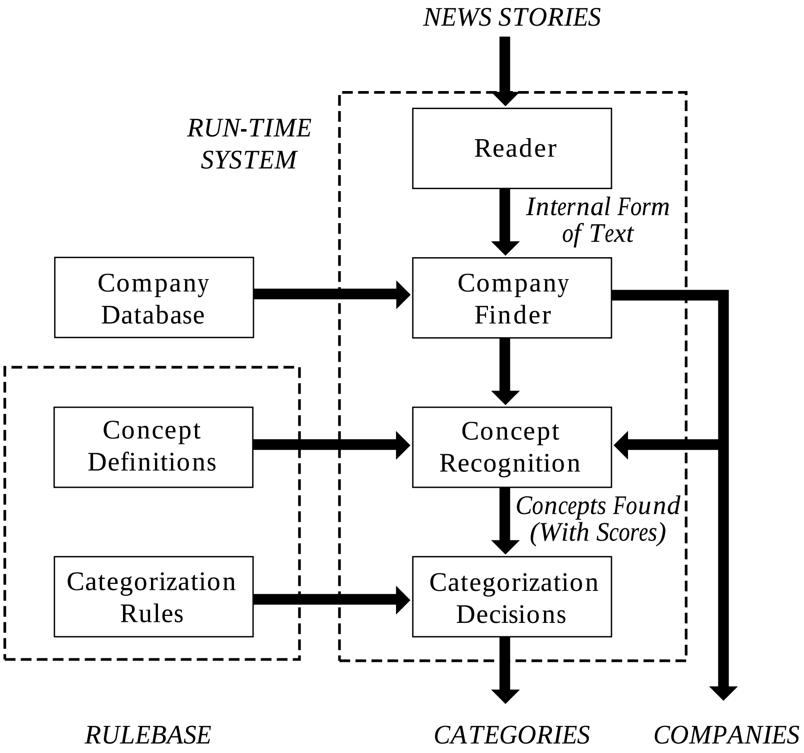

Figure 4. Construe's Flow of Control.

Construe recognizes concepts or ideas in stories through a *concept definition*, a set of words and phrases that are entered into its rule base as indicative of this concept. The language used for concept definitions resembles the most advanced of the Boolean key-word languages but goes further in scope. This resemblance means that Construe shares the robustness of key-word techniques in the face of a variety of languages without respect to grammaticality or style.

Construe's concept recognition also has the following three important characteristics: First, Construe's concept definitions provide a framework in which all words and phrases indicative of a concept can be grouped together, thus making it easy to deal with wide variations in the way a concept is expressed. Second, the Construe concept-definition language allows matches to be contextually restricted (for example, this word so long as it is (or is not) followed by this other phrase within six words); this arrangement can be used to cut down on spurious matches and to express some aspects of grammatical relations between words (such as the relationship between verb and subject).

Third, the phrases in a Construe concept definition can be individually weighted according to how indicative they are of the concept. Construe also tracks the number of matches of each phrase and comes up with an overall score for the strength of the appearance of the concept in the story. These differences hold the potential for much greater accuracy than that typically obtained through key-word techniques.

Construe makes its final categorization decisions on the basis of if-then rules rather than a fixed scheme for combining the weights of concepts recognized (figure 4), such as the one adopted by the Topic system (McCune et al. 1985). This control scheme allows the rule developer considerable flexibility in combining evidence from different concepts and also allows for encoding specialized knowledge related to particular categories. We found these facilities essential in encoding the distinctions necessary to achieve the high accuracy we obtained with Construe.

For example, Construe deals with the potential confusion caused by several different countries all calling their currency dollar by using a rule that assigns the Australian dollar category if (1) the Australian dollar concept matches or (2) a generic dollar concept not specific to any country matches, **and** the Australia concept matches, **and** the U.S. dollar concept and the Singapore dollar concept do **no**t match.

In addition to a run-time system that categorizes stories, Construe incorporates an interactive workbench for rule base editing, experimentation with specific rules against specific stories, recall-precision calculations, and rule base management. The workbench has proven invaluable in keeping the time and effort required for rule base development and maintenance within economically feasible bounds. Specific facilities include (1) *story analysis tools* for collecting and analyzing statistical information about the words and phrases contained in a set of stories to help the rule developer identify words and phrases that are indicative of specific categories; (2) *editing tools* for editing concept definitions and categorization rules; (3) rule base management tools for managing the files and directories that contain the rule base; (4) a tryout facility, an interactive menu-driven tool, for testing concept definitions and rules and generating recall and precision figures for stories that have been precategorized by human experts; (5) detailed output for providing rule developers with a full understanding of how each story was processed; and (6) batch facilities for noninteractive testing of a rule base on large files of stories (these facilities were essential in the latter stages of the project when the iterative process of rule base refinement frequently involved processing batches of hundreds of stories; the similar workbench facilities of TCS are described in Hayes et al. [1990]).

Concept Recognition

Construe categorization is based on the recognition of concepts in a story. A powerful pattern language allows rule developers to define a *concept* as patterns of words and phrases in context and to weight these patterns according to how indicative of the concept they are.

The Construe concept-definition language permits a clear, concise representation for complex word patterns. A single pattern can match many different words and phrases. Patterns can specify phrases with gaps for a specified number of arbitrary words, so the system can locate key expressions even when it is impossible to predict exactly what words they will contain. This function, combined with the word-order relationships expressed in a pattern, allows Construe to approximate the identification of grammatical relationships such as subject-verb and subject-object without the computational expense of syntactic parsing capabilities.

The Construe pattern language enables applications to filter matched patterns and reject patterns that contain certain key words but don't actually reflect the concept. For example, Construe detects the word gold to recognize the concept of gold as a commodity. However, phrases such as gold reserve, gold medal, or gold jewelry do not indicate the gold commodity concept. In the following pattern, &n indicates that the pattern should not match if the words following &n are present; so the pattern

(gold (&n (reserve ! medal ! jewelry)))

will detect the word gold but will not match the phrases gold reserve, gold medal, and gold jewelry. However, gold will match in these phrases: gold, gold mines, gold mining operations, and gold production. Because the &n operator can require the absence of words and phrases in the specific context of other phrases, rather than in the document as a whole, Construe rule developers have more precise control over filtering unwanted matches than do users of most key-word languages.

Each pattern in a concept definition is weighted to indicate how strongly it suggests the presence of this concept. For example, in Construe, the definition of the foreign exchange concept includes words and phrases about deficits or surpluses in the money markets, central bank actions, treasury repurchase agreements, any kind of interbank activity, currency speculation, and any kind of short-term fund. Any of these subjects could indicate the presence of the foreign exchange concept, but a phrase such as "the Fed intervened in the money market" is a much stronger indicator than a phrase such as "currency speculation." Thus, the pattern for the former phrase is weighted more heavily than that for the latter.

Categorization Rules

Categorization decisions are controlled by procedures written in the Construe rule language, which is organized around if-then rules. These rules permit application developers to base categorization decisions on Boolean combinations of concepts that appear in a story, the strength of the appearance of these concepts, and the location of a concept in a story. Thus, the Construe rule language is more flexible than the Boolean key-word languages. Combining concepts in categorization rules allows Construe (1) to infer the correct category assignment from among a set of similar categories and (2) to support weak evidence for an assignment with information about related concepts. As an example of the first point, consider that many countries use a currency called dollar, and a currency is not always specified by its full name in a news story. In a sentence such as “Australia announced today that it would devalue the dollar,” Construe can infer that the Australian dollar, rather than the U.S. or Singapore dollar, is under discussion. The rule for the category Australian dollar looks as follows:

```
(if
test: (or        [australian-dollar-concept]
                 (and    [dollar-concept]
                         [australia-concept]
                         (not [us-dollar-concept])
                         (not [singapore-dollar-concept])
                         ...))
action: (assign australian-dollar-category)...)
```

This rule says to assign the Australian dollar category if (1) the Australian dollar concept matches or (2) the dollar concept with no country specified matches, **and** the Australia concept matches, **and** the U.S. dollar concept and the Singapore dollar concept do **not** match.

Construe keeps score of the patterns matched in each concept and the weights assigned to them, so a weak indication of a concept can be strengthened by the presence of related concepts. For example, a concept definition for a commodity in Construe typically mentions the name of the commodity. One commodity handled by Construe is lead. Unfortunately, the word lead is ambiguous, and its presence might not provide sufficient evidence for assigning the category lead, especially if it only appears once in a story. The lead category can be correctly assigned to the following story because of the combined presence of the following concepts: lead, indicated by the word lead; metals, indicated by mining, mine, and ore; and general commodities, indicated by metric tons.

> **DOWA MINING TO PRODUCE GOLD FROM APRIL**
> TOKYO, March 16 - Dowa Mining Co Ltd said it will start commercial production of gold and lead from its Nurukawa Mine in northern Japan in April. A company spokesman said the mine's monthly output is expected to consist of 1,300 metric tons of gold ore and 3,700 of black ore. A company survey shows the gold ore contains up to 13.3 grams of gold per metric ton, he said. Proven gold ore reserves amount to 50,000 metric tons while estimated reserves of gold and black ores total one mln metric tons, he added.

Construe rule developers can capture knowledge about the organization of stories by specifying the part of a story in which to search for a concept. For example, a concept that is mentioned in the headline of a news story is generally a stronger indicator of its topic than the same concept appearing elsewhere in the story. The following rule is based on this fact: To assign the category gold, the gold concept definition should match once in the headline and once in the body, but it must match four times if it is only in the body:

```
(if
test: (or       (and    [gold-concept :scope headline 1]
                        [gold-concept :scope body 1])
                [gold-concept :scope body 4])
action: (assign gold-category)...)
```

Rules can also specify that a categorization decision should be made sensitive to the length of a story. For example, one match of a concept in a 50-word story might be sufficient evidence for the assignment of an associated category, but in a 250-word story, additional evidence might be required before the category is assigned.

Deployment Experience

Soon after delivery, Reuters transferred the development of Construe from New York to London and spent slightly over a year integrating the system into a new environment. During this time, it became known within Reuters as TIS. TIS differs from Construe only in the rule bases used with it and the Reuters systems it is connected to; the underlying categorization engine and development environment are unchanged. The first deployment of TIS to provide value to Reuters subscribers was with Reuters Country Reports in November 1989. The second was with Reuters Textline. To meet the specific indexing needs of existing subscriber groups, both of these deployments use rule bases that are modified subsets of the 674 category rule base delivered with Construe. Reuters anticipates that most of the as-yet-undeployed categories from Construe will be incorporated into future TIS systems.

Country Reports offers information, including news, on all 196 of the world's countries. Before the introduction of TIS, human editors indexed the information by country and type of news (general, economic, political, sport) using a total of 200 indexing terms. This indexing was slow and often meant that news on this service was days behind the actual events. The same indexing is now provided by TIS within moments of the time the stories are reported on the live news services. The rule base used for this deployment was developed by Reuters personnel from the rule base supplied with the Construe delivery with approximately six person-months of effort. It was measured on a sample of 700 stories with 98-percent recall, 99-percent precision for the country categories and 94-percent recall, 96-percent precision for the four general categories.

Textline is a textual database of business news from more than 1000 publications, including those from Reuters. It is currently indexed according to several hundred terms assigned by human editors. The large number of index terms available makes it hard for subscribers to determine which term to use for retrieval. Therefore, Reuters is in the process of replacing this scheme with a much smaller set of 88 terms to be generated by TIS. The TIS system for Textline is in its final test and was deployed in December 1990. The development of the revised Textline category scheme and the modification of the Construe rule base to correspond to it took about 1 person-year of effort.

Both of these TIS systems run together on a dedicated DEC VaxStation 3100 with 32 megabytes (MB) of main memory and 312 MB of hard disk. As shown in figure 5, the system configuration includes several such VaxStations, each running both systems, connected through DECnet to a central Reuters news editing system running on a Vax cluster. The news editing system queues stories for processing by the TIS systems. The TIS systems independently take single stories from the queues and return the corresponding index terms. The news editing systems then route the stories with index terms to the appropriate database server machines. This architecture allows for straightforward enhancement of throughput capacity by simply adding more TIS machines. It also means that machine hang-ups can disrupt the processing of only one story and machine downtime need only reduce throughput rather than completely halt processing. The integration of TIS into this environment required an additional person-year of effort by Reuters personnel.

Reuters estimates that the use of TIS will reduce costs by $752,000 in fiscal 1990, the first full year of TIS deployment. Savings in 1991 are predicted to be $1,264,000. Indexing staff members displaced by TIS have been moved to other projects and products that make better use

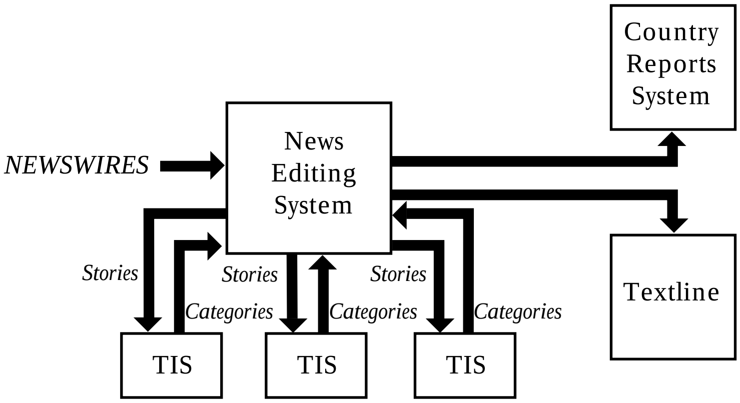

Figure 5. TIS Deployment Architecture.

of their experience in, and understanding of, business news. TIS also provides better service to subscribers than human indexers because of its more rapidly available and more consistent indexing. Reuters expects that this service enhancement will improve its competitive advantage and, hence, increase its market share. TIS has not, however, been in service long enough to permit such an increase to be measured or quantified in financial terms.

Summary

It is already clear that Construe, now deployed as TIS, will handsomely repay Reuters for its initial investment. Construe-TIS has shown that an automated knowledge-based, text- categorization system can provide indexing that is comparable to human indexing in accuracy on a commercially important indexing task but at a lower cost and more rapid pace. Thus, Construe-TIS meets Reuters minimum criteria for a successful application.

Moreover, Construe-TIS goes beyond the simple replacement of function already provided by people. Its greater consistency and speed actually improve the service to subscribers. It is too soon to know for sure what impact these service enhancements will have from a market perspective, but Reuters expects to gain a significant competitive advantage and, hence, an increased market share, meeting Reuters criteria for a second, higher level of success.

We believe that the success of Construe-TIS was critically dependent on the kind of shallow, semantic techniques we adopted. The techniques are shallow so that they can operate fast enough to process large volumes of text but can allow the incorporation of enough domain knowledge to provide high accuracy on the kind of text-categorization tasks we have pursued. We see the trade-off between speed and accuracy mediated by the depth of processing as a central design issue in the automatic processing of the extremely large volumes of text that occur in real-world applications. We were pleasantly surprised by how much could be accomplished by simple semantic techniques.

There are many other potential applications of this approach to the handling of real-world text, including routing texts to appropriate people, retrieving texts from a database, and creating indexes for large documents. Carnegie Group anticipates that TCS, its generalized version of Construe, will be applied to many of these problems by Carnegie Group and its clients, with benefits similar to Construe-TIS in terms of cost savings and competitive advantage from enhanced service.

Acknowledgments

The success of Construe-TIS results from the efforts of many people in addition to the authors (who directed the development effort for their respective organizations). The Construe project was initiated for Reuters by Geoff Chapman. The Carnegie Group team was coordinated by Laura Knecht, with additional key contributions from Peggy Andersen, Monica Cellio, Dawn Haberlach, Eric Lehmann, Peter Neuss, and Irene Nirenburg. The Reuters team for the development of Construe included Sally Bergen, Sam Dobbins, Kristin Holmes, Joe Polifroni, Michael Topliss, and Glenna Young, with TIS deployment handled by David Francis, Peter Sillet, Jonathan Richards, Glenna Young, and Sara Young.

References

Blair, D. C., and Maron, M. E. 1985. An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Communications of the ACM* 28(3): 289–299.

Dejong, G. 1982. An Overview of the Frump System. In *Strategies for Natural Language Processing*, eds. W. G. Lehnert and M. H. Ringle, 149–176. Hillsdale, N.J.: Lawrence Erlbaum.

Furnas, G. W.; Landauer, T. K.; Gomez, L. M.; and Dumais, S. T. 1987. The Vocabulary Problem in Human-System Communication. *Communications of the ACM* 30(11): 964–971.

Hayes, P. J.; Knecht, L. E.; and Cellio, M. J. 1988. A News Story Categorization System. In Proceedings of the Second Conference on Applied

Natural Language Processing, 9–17. Cambridge, Mass.: Association for Computational Linguistics.

Hayes, P. J.; Andersen, P. M.; Nirenburg, I. B.; and Schmandt, L. M. 1990. TCS: A Shell for Content-Based Text Categorization. In Proceedings of the Sixth IEEE AI Applications Conference, 320–326. Los Alamitos, Calif.: IEEE Computer Society Press.

Jacobs, P. S., and Rau, L. F. 1988. A Friendly Merger of Conceptual Analysis and Linguistic Processing in a Text Processing System. In Proceedings of the Fourth IEEE AI Applications Conference, 351–356. Los Alamitos, Calif.: IEEE Computer Society Press.

McCune, B. P.; Tong, R. M.; Dean, J. S.; and Shapiro, D. G. Rubric: A System for Rule-Based Information Retrieval. *IEEE Transactions on Software Engineering* SE-11(9): 939–945.

Salton, G. 1986. Another Look at Automatic Text Retrieval Systems. *Communications of the ACM* 29(7): 648–656.

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.