Prism: A Case-Based Telex Classifier

Marc Goodman

Banks receive a wide variety of telex communications. Bank operators must sort these telexes and route them to the appropriate department for processing. Although certain messages can easily be classified and routed, others require a more thorough understanding of the telex content and bank organization to determine the proper destination. The large number of messages that must be reviewed each day, the urgency of these messages, and the difficulty of maintaining a staff of sufficiently skilled operators all indicate the advantages of automating this task. Thus, Prism is a system that combines case-based and inductive techniques to classify and route bank telexes. Developed by Cognitive Systems, Inc. (CSI), the Prism system has been in continual daily operation at Chase Manhattan Bank's Letter of Credit Department (Chase L/C) since October 1989 and has been customized for installation at Manufacturer's Hanover Trust (MHT) and the American Express Bank, Limited (AEBL).

The Telex Classification Domain

Most large banks are attached to an international telex communications network. These banks receive interbank telex communications from a variety of foreign and domestic correspondent banks 24 hours a day. The average number of telexes received each day varies from site to site, ranging from several hundred to several thousand each day. These telexes cover a variety of topics, ranging from accounts receivable to volume banking. CSI has identified about 109 content-based classifications for these telexes.

Each bank has a number of different departments, ranging from 6 to 30, to handle subsets of this traffic. The structure and organization of these departments vary greatly from site to site. For example, one bank might send all telexes concerning payments to a specific department, whereas another bank might distinguish between foreign and domestic payments, payment investigations, payment amendments, and so on. Because a significant portion of these telexes deal with the transfer of large amounts of money from one bank or account to another, any delay in routing these telexes to the appropriate department can have costly consequences.

AEBL was using a manual process to route these telexes. Telexes were printed as they were received off the wire. An operator would read the telex and place the hard copy into an addressed mail envelope. This envelope would be hand delivered to the appropriate department, where processing would occur. This process could take as long as several minutes for each telex. Because telexes were received overnight, each morning would start with a backlog of hundreds of telexes, many of which had been queued for hours. As with any manual record processing, these records could be lost or misplaced. Given the sensitivity of these documents, such a risk is unacceptable.

Both Chase L/C and MHT had developed semiautomatic methods of dealing with telex traffic. At both sites, incoming telexes were appended to a queue. An operator sitting at a terminal would view the telex and attach an appropriate route code. The telex was then automatically forwarded to the appropriate department over the computer system. Although the time required to process a telex was reduced to one or two minutes for each telex, training costs increased somewhat because the operator needed training on the routing system. Also, the problem of telexes that were batched overnight remained.

There is a high employee turnover in telex operator positions. When combined with the wealth of specialized knowledge required to perform this job (as much as six months of training is typical), banks end up spending a large amount of money on human resources.

Prism Design Requirements

We envisioned Prism as a system that could be installed at a large number of banking sites, with a minimum amount of work required to customize the system at each site. Because the departmental structure varied greatly from bank to bank, we decided that Prism should be composed of two modules, a content-based classifier and a router. The *classifier* would determine the generic content of the telex from about 109 telex types, and the *router* would determine the appropriate customer department for a telex with this classification. Because most of the knowledge about classification would be generic, the total amount of knowledge engineering needed to customize Prism for a particular site would be reduced. Prism also needed to be aware of a large amount of customer-specific information to perform routing, including the names of employees and departments referred to in attention lines of the telex, the name and location of common correspondent banks, and the structure and content of telex reference and account numbers. Each of these pieces of information can have an impact on the final routing code, independent of the content of the telex. To reduce knowledge engineering time, this customer-specific information would be stored in external databases, and Prism would have facilities for looking up and using this information.

There were several other requirements for installing Prism. Prism needed to be easily portable to a wide variety of existing hardware platforms. The total amount of time it took to classify and route a telex needed to be less than the amount of time required in a bank's existing system. This meant that Prism needed to handle a telex in under one minute. Prism's accuracy needed to be at least as good as the average accuracy of a human operator. Fortunately for Prism, this number turned out to be only 75-percent accuracy because of the complexity of the task. Finally, as a business consideration, the total cost required to customize Prism for a client site needed to be less than the amount that could be charged for a Prism sale.

Rule-Based Prism

The first version of Prism combined a shallow, demon-based parsing of a telex with a forward-chaining rule system. The parser, based on Dypar (Dyer 1982), attempted to use semantic and syntactic information to perform an expectation-based parsing of the telex. This parsing involved creating several thousand lexical and pattern definitions. A large body of domain knowledge modeling the banking domain was also created, using a structured inheritance network (Brachman 1978). During the course of the parsing, information extracted from the telex was posted to a blackboard, and a forward-chaining rule system similar to OPS5 (Brownston et al. 1985) was used to classify and route the telex. Approximately 700 rules were created for the classification and routing. A schematic overview of rule-based prism is shown in figure 1.

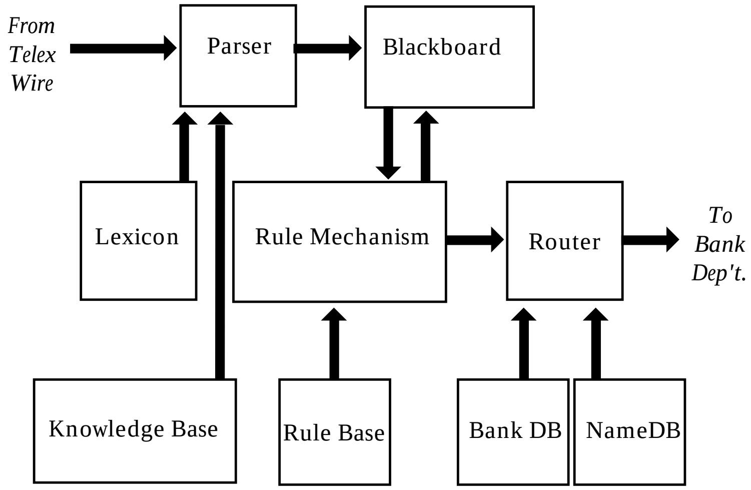


Figure 1. Rule-Based Prism.

Two systems were built using this architecture. Both systems satisfied the time requirement (a telex could be classified and routed in 44 seconds on the average) and the accuracy requirement (both systems had a 76-percent accuracy rate).The first system went live at MCI for Chase Manhattan Bank in July 1988, and the second system went live at Societe Generale de Banque in November 1989.

Unfortunately, the difficulty in customizing and maintaining such large, knowledge- intensive systems caused the cost of knowledge engineering new systems to be impractical. Each system required approximately one person-year to knowledge engineer, and prospects for enhancing system performance were grim. The complexities of rule chaining in the system quickly led to a situation where any modifications to the rule base to fix misclassifications resulted in an entirely new set of problems. Fixing this new set of problems would, in turn, cause new problems to develop. It was clear that no substantive system improvements could be made without a major redesign of the knowledge base. In an effort to reduce this knowledge engineering time, CSI turned to case-based and inductive approaches.

The History of Case-Based Prism

Case-based Prism began as a series of experiments in January 1989. CSI was beginning the second year of a three-year contract with the De-

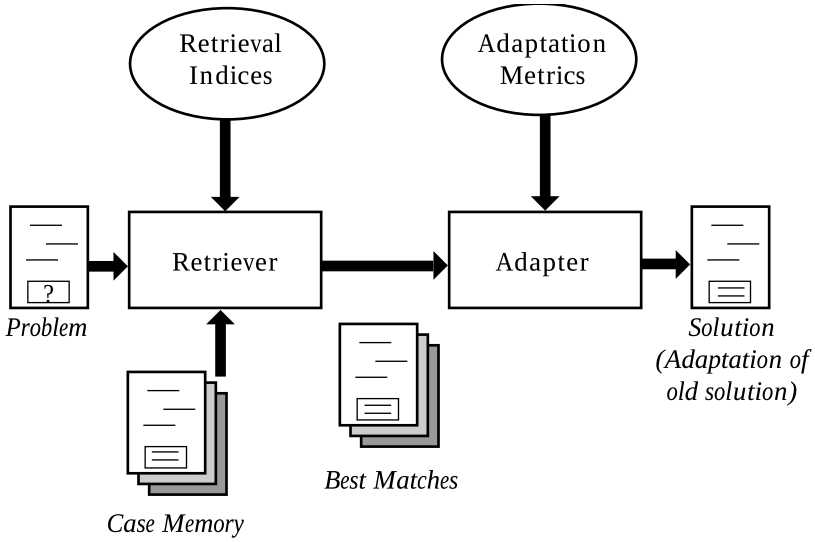


Figure 2. Case-Based Reasoning Process Model.

fense Advanced Research Projects Agency (Darpa) to develop a case-based reasoning tool (the CBR Tool) (Riesbeck 1988) and had experienced results indicating that significant reductions in knowledge engineering time were possible with this technology (Goodman 1988, 1989). A process model for case-based reasoning is shown in figure 2 (Riesbeck 1989). In this model, an initial problem description is passed into a case retriever. This retriever uses a set of retrieval indexes to select best matches from a case library. The retrieved cases are passed into a case adapter, which uses a set of adaptation metrics to compare the problem description with the retrieved cases to tweak solutions indexed on the retrieved cases to account for any remaining differences from the problem description. The result of this adaptation is a new solution. In contrast to the rule-based Prism approach, where the goal was to use detailed domain knowledge to reason from the content of a message to its classification, the goal of a case-based approach is to retrieve cases similar to an incoming telex from its case library and to use the classification of these telexes as the basis for a classification of the new telex.

The construction of a CBR system involves several steps. An adequate case representation must be defined, including the type and nature of features used to describe each case. Cases must be gathered and stored in the case library. Some notion of what makes cases similar must be created. CSI’s CBR Tool uses inductive techniques to build a directed acyclic graph, where each node in the graph is a binary discrimination on the presence or absence of a feature in the case. These features, which serve as indexes for case retrieval, are selected using a credit-assignment algorithm that evaluates the correlation between each feature and the variance in case outcome.

In the first series of experiments, a proof of concept was developed. CSI's Macintosh II–based CBR Tool was used to construct a case library of 4000 telexes. Initial classifications were provided by the rule-based Prism. A telex was considered to be composed of a set of individual words in isolation, and indexes were generated that corresponded to words that appeared to account for variance in classification. After two weeks of initial development, a system was created that achieved over 90-percent accuracy in predicting the classification portion of rule-based Prism. Further, because of the simplicity of the features used for classification, this first version of a case-based classifier was able to process the average telex in . 2 seconds. These initial results were promising enough that work proceeded to a second series of experiments.

In the second series of experiments, these techniques were used to build classifiers that would determine what the language of a telex was (from six languages: French, English, German, Dutch, Italian and Spanish) and whether the telex was a funds transfer telex. This second series of experiments required extension of the case library (one person-week) and new index generation (one person-week).At the end of these experiments, CSI had developed a language determiner that was 98-percent accurate and a funds transfer separator that was 90-percent accurate on test data.

By January 1989, CSI had negotiated a new installation with Chase L/C. Although the complexity of this task was reduced from Prism with 109 classifications, because Chase L/C was only interested in distinguishing between various kinds of letters of credit, the subtleties of classification made this effort a major one. A decision had to be made about whether CSI would use its established (but expensive) rule-based approach or chance using its newer case-based reasoning technology. After much internal discussion, the decision was made to try case-based reasoning, mainly because of the personnel requirements.

The first step toward building the new system was the solicitation of 3000 new letter-of credit telexes. These telexes were added to the case library, and a domain expert used the CBR Tool's interactive data entry to classify the new and existing letter-of-credit telexes. This process took approximately three person-weeks. Because a major source of error in CSI's previous case-based classifiers was the tool's inability to recognize synonymous words and phrases during index generation, CSI changed the internal representation of the telex. Instead of scanning a telex into a set of words that would serve as the basis for index generation, CSI instead borrowed its existing lexical pattern matcher from its parser and used it to scan the telex. This pattern matcher incorporated facilities for spelling correction along with the ability to define groups of synonymous words, patterns with optional components, and abbreviations. The output of the lexical pattern matcher was a set of symbolic values that could hierarchically be organized in the CBR Tool. Adapting the lexical pattern matcher to the case-based reasoning environment took one person-week, developing the lexicon took two person-weeks (starting from the base of the existing rule-based Prism lexicon), and creating the symbolic hierarchy took one person-week. Generating the indexes for case retrieval took an additional person-week using inductive techniques. Figures 3 through 10 illustrate the steps in constructing the case library.

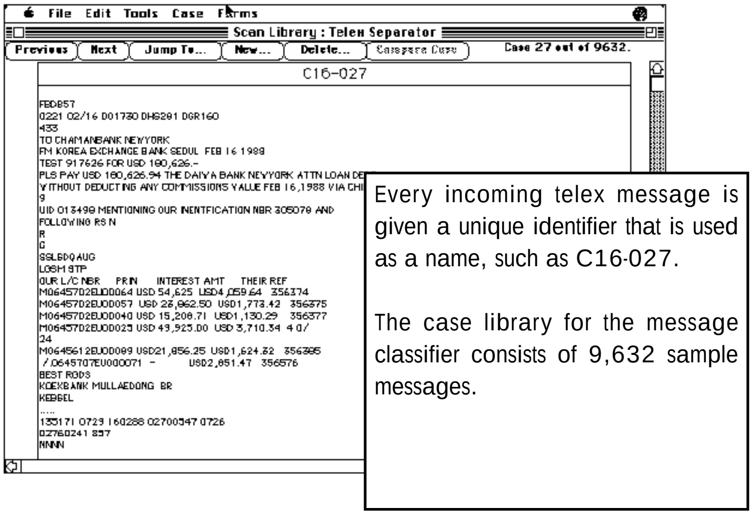
After these eight person-weeks, a system was developed that was 90-percent accurate at determining the classification for a letter-of-credit telex. Although the time required to run a telex through the lexical pattern matcher added five seconds to the total processing time, the time to classify a telex was still less than six seconds. This system has been in continuous, daily operation at Chase L/C since October 1989 and processes several hundred telexes each day.

By June 1989, CSI had negotiated two new installations of Prism, one for MHT and one for AEBL. Given CSI's previous success with case-based Prism, it seemed clear that this approach should be used.

Case-Based Prism Design

The final Prism design consists of three modules. The first module is the lexical pattern matcher, borrowed from CSI's natural language parser. The input to this module is a bank telex, and the output is a set of hierarchically organized symbolic values (such as Pay, Value-Date, Sender, Debit). This set of symbols defines the case representation to the second module, the CBR module, which classifies the telex into one of about 109 different content- based classifications. The result of this classification is passed to the third module, a rule-based router (all that remains of the rule-based Prism), which contains customer-specific rules for extracting additional information from the telex and deciding on the final routing code. The router is also responsible for recognizing attention lines and extracting reference numbers and telex party references with a shallow parser. A schematic overview of case-based Prism appears in figure 11.

Because of the small number of rules required by the router (22 rules

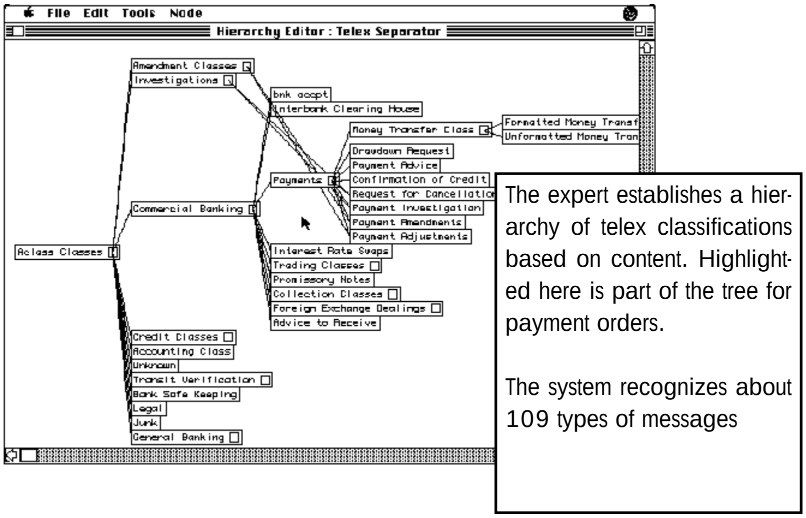


```
FEDB57
0221 02/16 D01730 DHS281 DGR160
433
TO CHAMANBANK NEWYORK
FM KOREA EXCHANGE BANK SEOUL  FEB 16 1989
TEST 917626 FOR USD 180,626.-
PLS PAY USD 180,626.94 THE DAIWA BANK NEWYORK ATTN LOAN DE
WITHOUT DEDUCTING ANY COMMISSIONS VALUE FEB 16,1988 VIA CHI
9
UID 013498 MENTIONING OUR INENTFICATION NBR 305078 AND
FOLLOWING RS N
R
G
SSLBDQAUG
LOSH STP
OUR L/C NBR    PRIN     INTEREST AMT    THEIR REF
M06457D2EUD0064 USD 54,625  USD4,059.64   356374
M06457D2EUD0057  USD 28,862.50  USD1,773.42   356375
M06457D2EUD0040 USD 15,208.71  USD1,130.29   356377
M06457D2EUD0025 USD 49,925.00  USD 3,710.34  4 0/
24
M064561 2EUD0089 USD21,856.25  USD1,624.82   356385
 /.0645707EU000071  -        USD2,851.47   356576
BEST RGDS
KOEKBANK MULLAEDONG  BR
KEBSEL
.....
135171 0729 160288 02700347 0726
02760241 857
NNNN
```

Every incoming telex message is given a unique identifier that is used as a name, such as C16-027.

The case library for the message classifier consists of 9,632 sample messages.

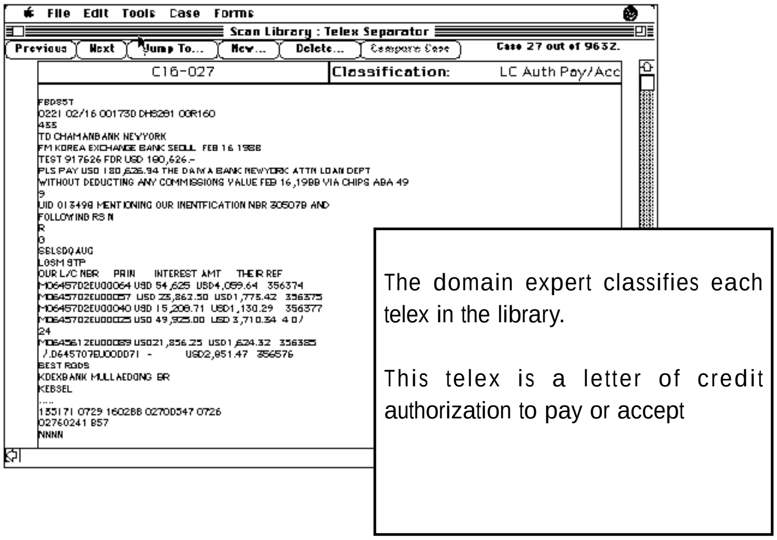
Figure 3. Collecting Messages.



The expert establishes a hierarchy of telex classifications based on content. Highlighted here is part of the tree for payment orders.

The system recognizes about 109 types of messages

Figure 4. Classification Hierarchy.

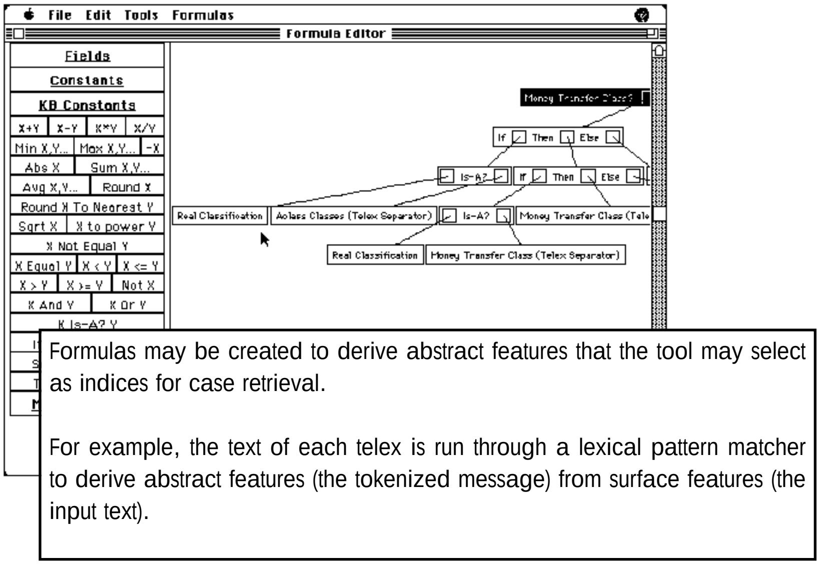


```
FBD85T
0221 02/16 001730 DHS281 OOR160
435
TO CHAMANBANK NEWYORK
FM KOREA EXCHANGE BANK SEOUL  FEB 16 1988
TEST 917626 FOR USD 180,626.-
PLS PAY USD 180,626.94 THE DAIWA BANK NEWYORK ATTN LOAN DEPT
WITHOUT DEDUCTING ANY COMMISSIONS VALUE FEB 16,1988 VIA CHIPS ABA 49
9
UID 013498 MENTIONING OUR INENTFICATION NBR 305078 AND
FOLLOWING RS N
R
G
SBLSDQ AUG
LGSM STP
OUR L/C NBR    PRIN      INTEREST AMT    THEIR REF
MD6457D2EU00064 USD 54,625  USD4,059.64   356374
MD6457D2EU00057 USD 25,863.50  USD1,775.42   356375
MD6457D2EU00040 USD 15,208.71  USD1,130.29   356377
MD6457D2EU00025 USD 49,925.00  USD 3,710.34  4 0/
24
MD64561 2EU00089 USD21,856.25  USD1,624.32   356385
 /.D645707EU00071  -        USD2,851.47   356576
BEST RGDS
KOEXBANK MULLAEDONG BR
KEBSEL
.....
135171 0729 160288 02700547 0726
02760241 B57
NNNN
```

The domain expert classifies each telex in the library.

This telex is a letter of credit authorization to pay or accept

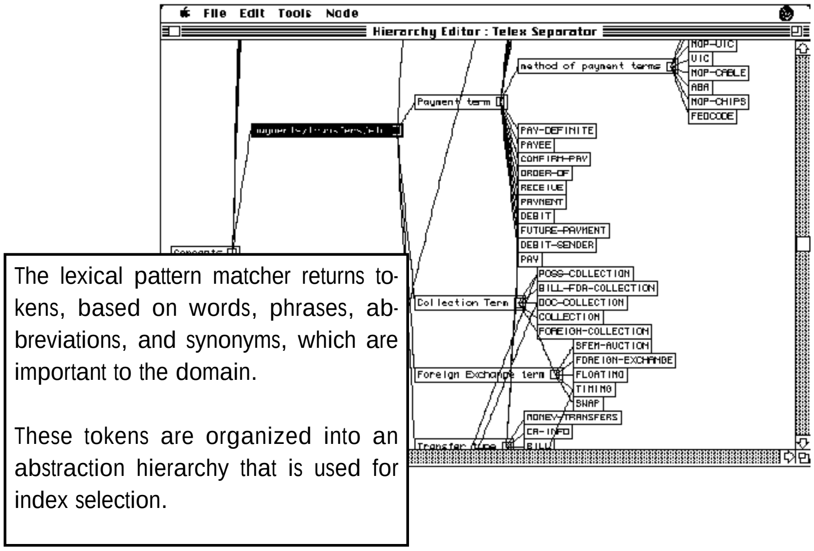
Figure 5. Classification of Messages.



Formulas may be created to derive abstract features that the tool may select as indices for case retrieval.

For example, the text of each telex is run through a lexical pattern matcher to derive abstract features (the tokenized message) from surface features (the input text).

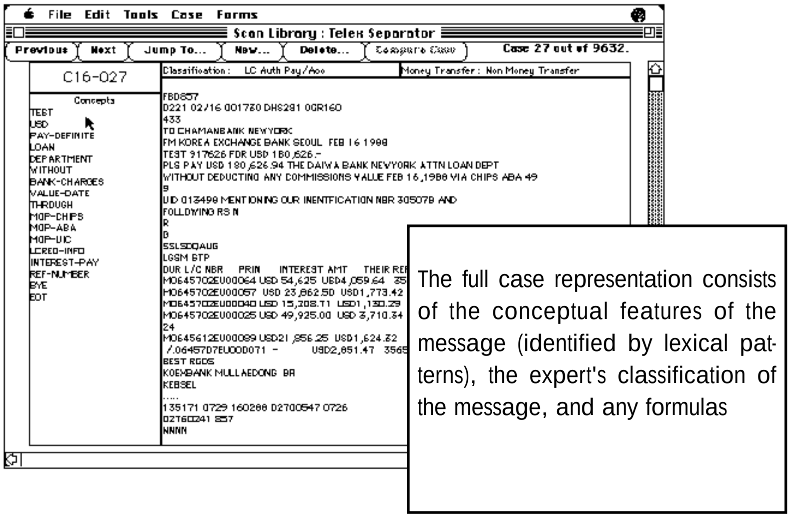
Figure 6. Creation of Formulas.



The lexical pattern matcher returns tokens, based on words, phrases, abbreviations, and synonyms, which are important to the domain.

These tokens are organized into an abstraction hierarchy that is used for index selection.

Figure 7. Symbolic Hierarchy of Tokens.



Case 27 out of 9632.

C16-027

Classification : LC Auth Pay/Acc

Money Transfer : Non Money Transfer

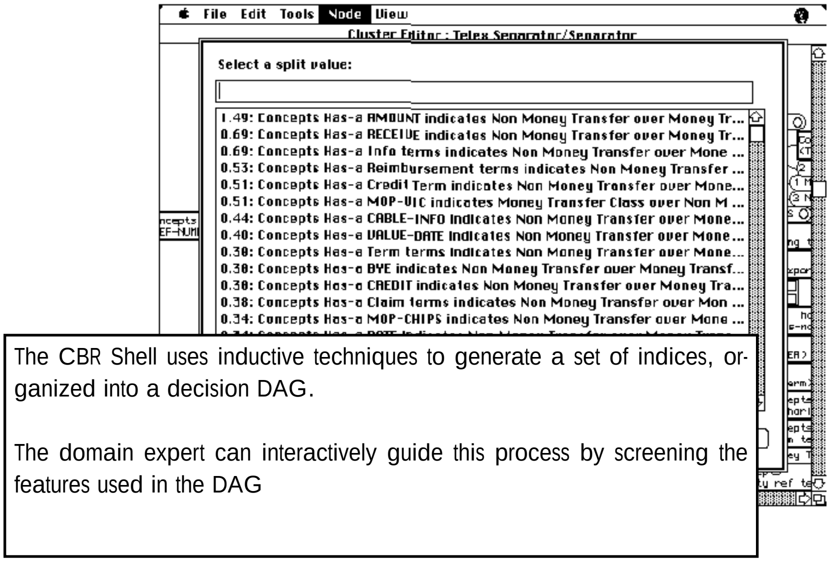
Concepts

TEST
USD
PAY-DEFINITE
LOAN
DEPARTMENT
WITHOUT
BANK-CHARGES
VALUE-DATE
THROUGH
MOP-CHIPS
MOP-ABA
MOP-UIC
LCRED-INFO
INTEREST-PAY
REF-NUMBER
BYE
EOT

```
FBD857
D221 02/16 001730 DHS291 0GR160
433
TO CHAMANBANK NEWYORK
FM KOREA EXCHANGE BANK SEOUL  FEB 16 1988
TEST 917626 FOR USD 180,626.-
PLS PAY USD 180,626.94 THE DAIWA BANK NEWYORK ATTN LOAN DEPT
WITHOUT DEDUCTING ANY COMMISSIONS VALUE FEB 16,1988 VIA CHIPS ABA 49
9
UID 013498 MENTIONING OUR INENTFICATION NBR 305078 AND
FOLLOWING RS N
R
B
SSLSDQAUG
LGSM BTP
DUR L/C NBR     PRIN      INTEREST AMT     THEIR REF
M0645702EU00064 USD 54,625  USD4,059.64   35
M0645702EU00057  USD 23,862.50  USD1,773.42
M0645702EU00040 USD 15,308.71  USD1,130.29
M0645702EU00025 USD 49,925.00  USD 3,710.34
24
M0645612EU00089 USD21,856.25  USD1,624.32
 /.064570 7EU00D071  -        USD2,851.47   3565
BEST RGDS
KOEXBANK MULLAEDONG  BR
KEBSEL
.....
135171 0729 160288 D2700547 0726
02760241 857
NNNN
```

The full case representation consists of the conceptual features of the message (identified by lexical patterns), the expert's classification of the message, and any formulas

Figure 8. Full Representation of Case.



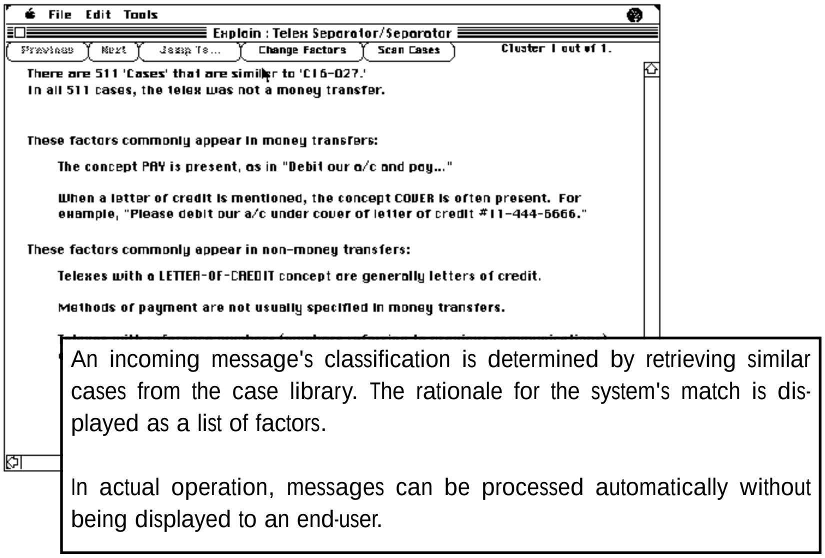
Select a split value:

```
1.49: Concepts Has-a AMOUNT indicates Non Money Transfer over Money Tr...
0.69: Concepts Has-a RECEIVE indicates Non Money Transfer over Money Tr...
0.69: Concepts Has-a Info terms indicates Non Money Transfer over Mone ...
0.53: Concepts Has-a Reimbursement terms indicates Non Money Transfer ...
0.51: Concepts Has-a Credit Term indicates Non Money Transfer over Mone...
0.51: Concepts Has-a MOP-VIC indicates Money Transfer Class over Non M ...
0.44: Concepts Has-a CABLE-INFO indicates Non Money Transfer over Mone...
0.40: Concepts Has-a VALUE-DATE Indicates Non Money Transfer over Mone...
0.38: Concepts Has-a Term terms indicates Non Money Transfer over Mone...
0.38: Concepts Has-a BYE indicates Non Money Transfer over Money Transf...
0.38: Concepts Has-a CREDIT indicates Non Money Transfer over Money Tra...
0.38: Concepts Has-a Claim terms indicates Non Money Transfer over Mon ...
0.34: Concepts Has-a MOP-CHIPS indicates Non Money Transfer over Mone ...
```

The CBR Shell uses inductive techniques to generate a set of indices, organized into a decision DAG.

The domain expert can interactively guide this process by screening the features used in the DAG

Figure 9. Clustering.



There are 511 'Cases' that are similar to 'C16-027.'
In all 511 cases, the telex was not a money transfer.

These factors commonly appear in money transfers:

The concept PAY is present, as in "Debit our a/c and pay..."

When a letter of credit is mentioned, the concept COVER is often present. For example, "Please debit our a/c under cover of letter of credit #11-444-5666."

These factors commonly appear in non-money transfers:

Telexes with a LETTER-OF-CREDIT concept are generally letters of credit.

Methods of payment are not usually specified in money transfers.

An incoming message's classification is determined by retrieving similar cases from the case library. The rationale for the system's match is displayed as a list of factors.

In actual operation, messages can be processed automatically without being displayed to an end-user.

Figure 10. Retrieving Cases.

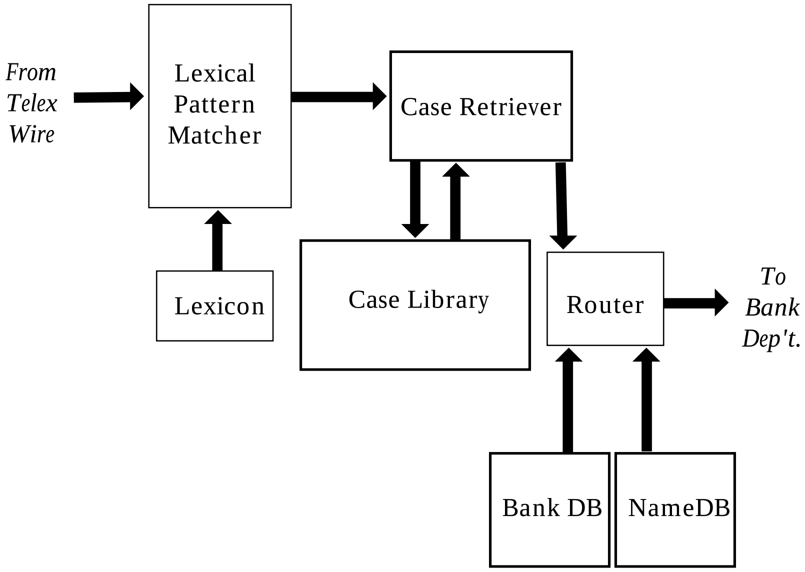


Figure 11. Case-Based Prism

for the MHT installation and about 30 for the AEBL installation), the knowledge engineering time to customize Prism for a new client is reduced. In fact, the total time required to customize both the AEBL and MHT systems was less than eight weeks. Further, improving the performance of the system is simply a matter of adding telexes that the system misses into its case library (which currently contains over 9600 telexes).

Adding the router (which selectively extracts information from the telex) increased Prism's total telex processing time to 30 seconds. Still, 30 seconds is a significant improvement over rule-based Prism's 44 seconds for each telex and a dramatic increase over each bank's previous telex processing procedure. Also, Prism (unlike human operators) is able to work effectively 24 hours a day. This ability to process telexes in a more timely and cost-effective fashion will allow MHT to reduce its current staff of telex operators from five people to between two and three people. Prism has also been able to guarantee higher consistency than previously possible.

Conclusions

Using case-based and inductive approaches, a system that can easily be maintained and customized was developed for classifying bank telexes.

The techniques used in developing Prism appear to have widespread implications for the development of message classification systems. Prism has enabled the processing of bank telexes in a more timely fashion and has allowed the staffing requirements to be reduced at three banks.

Acknowledgments

Case-based Prism would still be vaporware without the valuable support and assistance of the following people: Lt. Col. Bob Simpson who sponsored work on the CBR Tool at Darpa, Abe Waksman at the Air Force Office of Scientific Research who administered the case-based reasoning contract, Philomena Abbene at Chase L/C who pulled for Prism as hard as I pushed, Steve Mott and Pete Johnson who believed in the technology enough to try it, Anatole Gershman and Sudah Kailer who designed and built rule-based Prism, Joe Devlin and Andy Wilson who helped with early experiments in case-based message classification, Bob Strong and Jim Meehan who provided technical support and encouragement, Anurang Revri who worked on the router and router customization, BJ Patnode who acted as the project leader for message processing, and Ben Wyckoff who built most of the CBR Tool.

References

Brachman, R. 1978. A Structural Paradigm for Representing Knowledge, Technical Report 3605, Bolt Beranek and Newman.

Brownston, L.; Farrell, R.; Kant, E.; and Martin, N. 1985. *Programming Expert Systems in OPS5*. Reading, Mass.: Addison-Wesley.

Dyer, M. 1982. In-Depth Understanding; A Computer Model of Integrated Processing for Narrative Comprehension, Research Report, 219, Dept. of Computer Science, Yale Univ.

Goodman, M. 1989. CBR in Battle Planning. In *Second Proceedings of a Workshop on Case-Based Reasoning*, 264–269. San Mateo, Calif.: Morgan Kaufmann.

Goodman, M. 1988. Case-Based Retrieval for Knowledge-Based Systems Construction. Presented at the Sixth Intelligence Community Artificial Intelligence Symposium, Washington, D.C.

Riesbeck, C. 1989. CBR Process Model for Problem Solving, Defense Advanced Research Projects Agency case-based reasoning slides, Cognitive Systems, New Haven, Conn.

Riesbeck, C. 1988. An Interface for Case-Based Knowledge Acquisition. In *First Proceedings of a Workshop on Case-Based Reasoning*, 312–326. San Mateo, Calif.: Morgan Kaufmann.