ATRANS: Automatic Processing of Money Transfer Messages

Steven L. Lytinen and Anatole Gershman

Cognitive Systems, Inc.
234 Church Street
New Haven, CT. 06510

ABSTRACT

Unformatted natural-language money-transfer messages play an important role in the international banking system. Manually reading such messages and encoding them in the format understandable by a bank's automatic payment system is relatively slow and expensive. Due to the very restricted nature of the domain, the problem lends itself naturally to a Conceptual Dependency (CD), script-style solution. This paper illustrates the solutions to a number of problems that arise when an academic theory is applied to a real-world problem. In particular, we concentrate on the problem of context localization in the absence of reliable syntactic clues, such as sentence boundaries.

I. INTRODUCTION

This paper describes a real-world natural-language understanding system, ATRANS (Automatic Funds TRANSfer Telex-Reader), which extracts information from telex messages. The messages are requests for transfers of money which banks send to each other. ATRANS reads these messages, extracts the necessary information, and then outputs it in a form suitable for automatic execution of the transfer. This paper will present an overview of the problems presented by the domain, outline the general solution, and discuss in more detail the solution to one of the problems, namely context localization and the resolution of semantic lexical ambiguities.

ATRANS routinely processes a wide variety of money transfer messages sent by banks around the world. These telexes are often composed by people whose ideas of English spelling, sentence construction, standard abbreviations, amounts and date conventions are very different from Standard American English. In addition, since these messages were intended for human visual inspection, senders very often introduce various kinds of visual "embellishments" such as table formats, stars, dashes, frames, etc., which can easily confuse a purely linguistics-based analyzer. In spite of these difficulties, ATRANS correctly extracts approximately 80% of the desired information fields. About 15% of the information items are missed and 5% are identified incorrectly. (When ATRANS has any "doubts," an item is not filled, rather than filled incorrectly.) With about half of the messages, all information fields are processed completely and correctly. All messages are then verified and, if necessary, corrected by a human operator.

*The ATRANS System was developed by Steve Lytinen, Steve Miklos, Anatole Gershman, Michael Lipman, Richard Wyckoff, and Ignace D'Haenens.

In the next section we introduce the domain of international money transfer messages and outline some of the major difficulties it presents. Section 3 presents our general approach to the solution of the problem. Section 4 discusses the problem of context localization in greater detail.

II. THE DOMAIN OF MONEY-TRANSFER MESSAGES

We will begin by presenting two simple examples of international money-transfer telex messages.

```
FROM: GEBABEBB18A : GENERALE BANK ANTWERPEN
TO  : TLX CTIUS33XXX : BIG BANK NEW YORK NY
REF : 1977675454
MSG : NORMAL
TEST:      51375 : BRUSSELS  ON 1748 USD


TLXNO11/1909TB

VALUE 851118
DEBITING GENERALE BRUSSELS
CREEDIT USD 174.806,65

TO  : CREDIT LYONNAIS PARIS
REF : FX / CVDW / 96098 / 45492

COLL 174.806,65

X : 15/11/85 14 28  ISN : 00125
    15/11/85 14 40  OSN : 00005 BGBKUS33XXX

MEDIC REF ORG/NEW 10630/13769
```

This telex requests that $174,806.65 be transferred from the account of Societe General de Banque, Brussels (from "debiting Generale Brussels" in the text) to the account of Credit Lyonnais, Paris. Presumably both banks have accounts with Big Bank, New York. Thus, Big Bank should simply transfer this amount of money from one account to the other.

Most messages also include several other pieces of information. The value date of Nov. 18, 1985 (from "value 851118") means that any currency exchanges necessary for this transaction should be done using the exchange rates for this date. The test key, 51375, is used to verify the authenticity of the message. It is computed from the value date and the amount and currency of the transaction. Reference numbers such as "FX / CVDW / 96098 / 45492" are attached by the sender and the beneficiary to provide both a unique identification of the transfer and an audit trail.

All of this information is converted by ATRANS into a standard format, from which it then generates an output format appropriate for the client's payment-processing system. The following is a fragment of the standard format for the above message produced by ATRANS.

```
Test key:                51375
Amount:                  174806.65
Currency:                USD
Value Date:              Nov. 18, 1985
Sender name:             General Bank
Sender city:             Antwerpen
Sender ref:              TLXN011/2909TB
Beneficiary ref:         FX/CVDW/96098/45492
Credit party account:    12345678
Credit party name:       Credit Lyonnais
Credit party city:       Paris
Debit party account:     87654321
Debit party name:        General Bank
Debit party city:        Brussels
```

What is required to process a message such as the above example? First, most messages contain a great deal of irrelevant information. In this telex, there are strings of characters identifying telex lines, message numbers, etc. Some messages even contain greetings from telex operators or other irrelevant text. The program must, therefore, be quite robust, capable of accounting for, or ignoring, every word in the input.

Lexical access in the system must also be very robust. First, words are sometimes misspelt, such as "creedit" above. Second, the names of banks and customers are often given in the messages in non-standard ways. The above message mentions "Generale Brussels," which refers to a bank in Brussels whose full name is "Societe Generale de Banque." The same bank is also often referred to as SGB. The system must be able to identify which bank is referred to by these non-standard names.

The problem of bank and customer name recognition is very serious. There are many variations of what constitutes the "standard" name of a bank. The "standard" name of the New York branch of Barclays Bank is "Barclays Bank of New York" which is rarely used by telex senders. Instead, we often see something like "Barclays, New York." The Flemish branches of Societe Generale de Banque are called Generale Bankmaatshappij, the British Commonwealth branches of the same bank are called Belgian Bank, and the German branches, Belgische Bank. In most cases, people will use the name of the bank that is most common in their own country. Thus, a beneficiary of a transfer may be specified as "Societe Generale de Banque, Antwerpen," even though the telex receiver's database does not list a bank under that name in Antwerp.

This problem is compounded by the fact that there is no single complete database with "standard" bank names. Each bank uses its own, which in most cases was originally designed for mailing purposes and was typed in by several generations of secretaries. (In one such database, we found about 1200 entries beginning with "TO: ". A typical large bank's database of corresponding banks and commercial customers contains anywhere from 20,000 to 40,000 entries.)

Messages are often ungrammatical and are usually written in one very long sentence, which gives no clues as to where different sections of the message begin and end. In addition, the input often contains ambiguous lexical items. In this example, both the recipient of the telex message (Big Bank) and the beneficiary of the transaction (Credit Lyonnais) are marked by the word "to." Similarly, the word "credit" is used as a synonym for the word "pay," but it also appears as the first word in the name of a bank. Similar expressions are interpreted differently depending on where in the telex they are encountered.

The way in which numbers should be interpreted in this message also varies. After the word "value," the program must know to interpret the number "851118" as a date (Nov. 18, 1985). However, if the same string of numbers appeared after a currency type, such as "USD" (U.S. Dollars), then it would be interpreted as an amount, or $851,118.00. Similarly, after "ref," which indicates that a reference number follows, numbers must simply be treated as strings, copied verbatim into the reference field.

The above examples show that even in such a narrow domain as money-transfer telexes, a text-understanding system must show a great deal of flexibility, both in tolerating the appearance of lexical items in the text which are unknown to the program and in determining when known words or phrases are misspelled or referred to in non-standard ways. In addition, the extraction of standard fields for money transfers must proceed without explicit cues, such as separate sentences, that might indicate where the fields can be found, and must take place in the presence of lexical ambiguities that can complicate the process.

III. HOW ATRANS WORKS

To deal with the problems outlined in the last section effectively, ATRANS uses a knowledge-based approach to text analysis. Although the structure of telex messages can vary a great deal, their content is very predictable. We can use the predictability of the content to guide the parsing process and overcome the problems we discussed earlier.

Much of ATRANS' knowledge of the input domain can be organized in terms of a script [9] or a standard sequence of actions which we can expect to occur in a money transfer. The script is the following:

1. Customer OC (Originating Customer) in country A asks his local bank OB (Originating Bank) to send some money M to a beneficiary BC (Beneficiary Customer) in country B.
2. Bank OB asks a large international bank SB (Sender Bank) in country A to forward the money.
3. Bank SB sends a request (the message that we are reading) to its corresponding bank RB (Receipient Bank) in country B.
4. Bank RB pays the money to a local bank BB (Beneficiary Bank) with whom the beneficiary customer has an account.
5. Bank BB pays the beneficiary customer BC.
6. Bank RB wants to be reimbursed for the money it pays. According to the instructions contained in the message, it either debits SB's account with itself, or waits until the money is credited to one of its accounts with some other bank CB (Cover Bank).

There are a number of variations of the above script, including a number of intermediary banks, banks trading on their own accounts, different methods of payments, etc. A message can also request several payments to different beneficiaries.

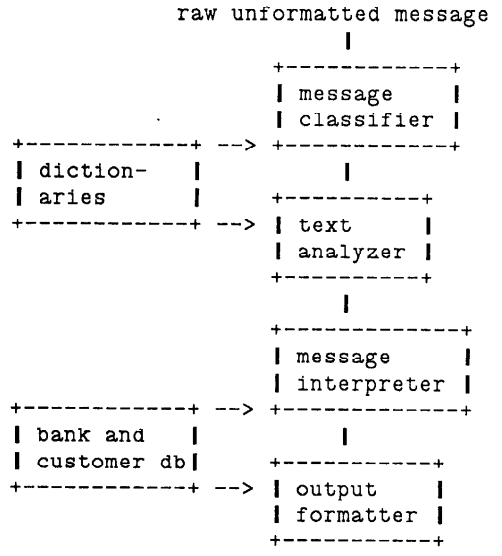

Figure 1: Structure of the ATRANS System

The ATRANS system consists of four parts, as illustrated in figure 1. The message-classification module determines the type of message being processed and chooses a variation of the transfer script to be applied. If the message contains multiple transfers, the module identifies the common portions of the transfer and composes several single transfer messages. "Visual" clues, such as table-like alignment of amounts and dates, play an important role in determining if a message contains a request for multiple payments.

The Text Analyzer is the heart of the system. It processes each telex from left to right in a deterministic manner, producing a Conceptual Dependency (CD) representation [8] of the telex content. The Analyzer follows the general line of semantically-based predictive conceptual analyzers (for details, see [7], [1], [6], and [5]). The basic script for international money transfers consists of a number of frames, some of which can occur only in a prescribed order and some of which can occur anywhere in the message text. Using the script, the dictionaries, and the context localization mechanism (described in the next section), the Analyzer identifies the frames being referred to by the text (e.g., payment, test, cover, etc.) and sets up expectations which interpret and extract information items completing those frames (e.g., amounts, dates, banks, etc.).

The same information items can be specified in different places within the same message. For example, the sender of the telex can be explicitly stated in the beginning of the message (e.g., "Here is ..." or "from ... "), at the end of the message (e.g., "Regards, ..."), or as a telegraphic answerback key (e.g., "918824 ESTNCO G"). Some of this information may not be 100% reliable, as when the sender uses somebody else's telex machine, producing a misleading answerback key. However, if different passages in the text confirm one another, we can conclude with a high degree of confidence that the telex was understood correctly.

The Analyzer does not verify the extracted information or check it for consistency. This is the job of the Message Interpreter. It verifies and consolidates the extracted information items, looks up in the data base the appropriate account numbers and customer addresses, and decides on the most appropriate method of payment. The result is represented internally in what we call a Universal Message Format. From this format the Output Generator produces the output in the form appropriate for the particular user of the system (e.g., SWIFT, CHIPS, Fedwire).

IV. CONTEXT LOCALIZATION IN ATRANS

Now that we have given an overview of the problems which must be solved in order to process messages in the domain of international money transfers, we will concentrate on the solution of one of these problems: context localization and, in particular, how it is used to resolve lexical ambiguities.

It is well-known that context can often eliminate semantic lexical ambiguities in texts. Words which in general have many different meanings often have only one possible meaning within a limited enough context. Riesbeck [7] presented the following example of this situation:

John and Mary were racing. John beat Mary.

In general, "beat" has several meanings, such as "to hit repeatedly," "to be victorious in a competition," or "to mix thoroughly" (e.g., to beat an egg). However, in the context of "racing," it is clear that "beat" means "to be victorious in a competition."

In script-based systems, particular contexts "prime" or give preference to particular senses of ambiguous words by using what is called "scriptal lexicons" [2] [3]. In the above example, the word "racing" would activate expectations associated with the concept of racing, including a specialized vocabulary of "racing terms" in which the word "beat" would have the single meaning of "to be victorious."

ATRANS uses an extension of the scriptal-lexicon idea to focus its expectations and resolve ambiguities. Instead of associating a scriptal lexicon with a relatively large script, ATRANS uses a hierarchy of local contexts, each of which uses a smaller "contextual lexicon." As is the case with every context-based system, the following issues must be addressed:

1. What is the mechanism by which a local context is activated?
2. How broad a range of word senses should a given context prime?
3. How long should a context be active (i.e., how do we know when the context has changed)?

To bring contextual information to bear on the resolution of ambiguities, ATRANS has a set of separate lexicons, each of which contains definitions for words or word senses which refer to a certain class of objects which the program must find. For example, one of the lexicons contains only names of banks. Another lexicon contains definitions of words which are likely to appear in addresses, such as "street," as well as names of cities and information about how to process numbers such as zip codes. Other lexicons contain only currency types, only words having to do with dates, or only non-bank customer names.

Within any single lexicon, lexical items are unambiguous. For example, in the Address lexicon, numbers are defined exclusively as zip codes or street numbers, not as dates or amounts. In the Bank-name lexicon, the word "Credit" is defined as the first word in the names of several banks, such as "Credit Lyonnais," but not as meaning the same thing as "pay."

During the processing of a telex message, ATRANS maintains a list of lexicons which are currently active. The

system has a set of rules which determine when this list should be altered, either by activating new lexicons or deactivating currently-active lexicons. Thus, potential ambiguities are resolved by virtue of which lexicons are active when the word is encountered. For example, if the Date lexicon is active, "851113" is interpreted as a date, because of the definition of a number in the Date lexicon. However, if the Currency lexicon were active, the definition of this same number would be interpreted as "$851,113.00." Similarly, if the Bank lexicon were active, the word "Credit" would cause the parser to try to match the input against bank names beginning with "Credit," rather than try to interpret the word as meaning "pay."

The types of lexicons we have described so far are appropriate when context predicts that a certain type of object will occur next in the input. For example, after the phrase "value date," it is very likely that a date will follow. Thus the Date lexicon is activated. At different times, however, the level of specificity of the expectations that context can provide varies a great deal. Because of this, ATRANS also has a range of lexicons which vary in their level of specificity.

Because ATRANS' job is to find the fillers of particular fields in a telex message which correspond to the most specific lexicons in the system, more general lexicons exist solely to determine when context can be refined enough to activate the specific lexicons. For example, the most general lexicon, called the Telex lexicon, contains definitions of words which mark general divisions of the telex message, such as the heading, the body, and the sign-off. This lexicon contains words such as "from" and "to," which mark the beginning of a message header; "pay" and "credit" (the sense meaning "pay"), which often mark the beginning of the body of the message; and words such as "regards," which mark the end of the body. Part of the definitions of these words is information that activates more specific contexts. For example, after the word "pay," it is likely that only certain information about the transaction will appear, such as information about the beneficiary and intermediate banks. Thus, one lexicon which "pay" activates is the Pay lexicon, which contains definitions of words such as "in favor of," "to" (meaning "beneficiary"), "account," etc. The definitions of these words contain information which in turn causes more specific lexicons to be activated. For instance, since the beneficiary is likely to follow immediately after "in favor of," this phrase activates the Bank lexicon and the Customer lexicon.

Because of the way in which lexicons in ATRANS activate each other, they can be viewed as being arranged into a hierarchy. Very general lexicons at the top of the hierarchy, such as the Telex lexicon, contain definitions of words which activate lexicons at the next level of the hierarchy, such as the Pay lexicon. These lexicons in turn contain definitions which activate lexicons at the next level down. This continues down to lexicons at the bottom of the hierarchy, such as the Date lexicon, the Bank lexicon, etc., which look for specific fields in the transaction.

We will now address the problem of what words should be included in a contextual lexicon. Clearly, the words which directly refer to the expected concepts should be included. In many cases, however, the meanings of words which only indirectly refer to expected concepts should also be favored over other meanings of these words. For example:

John went to a restaurant. He ordered a rare ...

At this point in the sentence, it is already possible to disambiguate "rare" to mean "not well-done" rather than "highly unusual." However, this word does not refer to one of the roles or events which are explicit in the restaurant script. It refers to a property of food, which is an explicit role-filler, but does not refer directly to the food.

In the ATRANS system, this problem is overcome in two ways. First, lexicons which contain word senses referring to particular objects also contain words referring to related concepts. For example, when mentioning the reimbursement account for a transaction, the telex message will often give the type of account or the branch of the sender bank to which the account belongs. Thus in the Reimbursement lexicon, although the type of object explicitly being looked for is an account, words and phrases such as "branch," "head office," and "foreign office" are also included in this lexicon. Secondly, lexicons are often paired together, so that one lexicon will always be activated whenever another lexicon is activated. For instance, whenever ATRANS looks for a customer, both the Customer lexicon and the Address lexicon are activated, because it is likely that an address will accompany the customer name in the telex message.

Finally, we have to address the issue of context deactivation. Once a set of word senses is primed, how long should they stay primed? For example:

John and Mary were racing. Mary won. John got mad and beat her.

At some point in this story, we must realize that the racing context no longer applies, and that "beat" therefore means "hit repeatedly."

The ATRANS system uses the hierarchical organization of its lexicons to determine when to switch contexts. At all times, the system maintains a stack of previously-active lexicons. This stack is maintained so that the system can return to previously-active, less specific contexts when the specific expectations of currently-active contexts are not met. Whenever the Analyzer encounters a word which is not defined in the current context but which does have a definition in one of the previous contexts on the stack, the Analyzer abandons the current context and restores the previous context. For example:

```
TO:     BIG BANK, NEW YORK

PAY USD 100,000 IN FAVOR OF BANK A
ACCOUNT WITH YOURSELVES

IN COVER OF CREDOC #133563

REGARDS,
BANK B
NEW YORK
```

The phrase "in cover of" activates a set of lexicons used to find information about reimbursement for the recipient bank. This set of lexicons includes the Bank lexicon, which contains bank names. However, in this particular message, no information about reimbursement is given. Therefore, the Analyzer needs to know when to stop looking for this information. When the word "regards" is reached, the Analyzer knows that the reimbursement context should be abandoned because "regards" is not defined in any of the currently-active lexicons but is defined in a previously-active lexicon, namely the Telex lexicon, which contains definitions of words which mark different sections of the telex message. Because of this, the context in which the Telex lexicon was

active is restored, thus de-activating the context set up by "in cover of." In this case, since the telex context which is re-activated was active several contexts ago, the popping of the context stack also eliminates the possibility that other, more recently-active, contexts might be re-activated, such as the "pay" context which looks for phrases such as "in favor of," "account," etc.

V. CONCLUSION

We have presented a knowledge-based text-understanding system which processes telex messages reliably and robustly in the domain of international money transfers. Although the input messages are noisy, including irrelevant text, misspellings, non-standard references to banks, and many ambiguities, the system's use of knowledge about the domain allows it to extract the important information in a robust manner.

We have presented in detail the solution to one of the issues that must be faced in such a system, namely the resolution of lexical ambiguities. ATRANS takes advantage of the fact that, in some contexts, words which in general are ambiguous can be treated as if they have only one meaning. Although the structure of telex messages gives us few contextual clues, ATRANS is able to use its knowledge of the domain to determne when particular contexts should be activated or de-activated.

To decide when a particular lexicon or set of lexicons should be activated, lexicons in ATRANS are arranged hierarchically. Thus, when expectations provided by context are very general, very general lexicons are used by the system. As context creates more specific expectations, more specific lexicons are activated. This approach also provides a natural solution to the problem of knowing when to de-activate a particular context. A stack of previous contexts is maintained by the system. Whenever a word or phrase which was defined in a previous context but not in the present context is encountered, this is taken as a signal that the present context should be abandoned and the previous context should be re-activated. In this way, the system is able to de-activate specific expectations at the appropriate times, and fall back on previously-active general expectations to determine what the next context in the message should be.

In addition to benefits in performance, the use of local lexicons in ATRANS proves to have organizational benefits as well. Because the system uses local contexts, different programmers were able to develop parsing rules for different contexts independently.

In contrast to other message-parsing systems such as FRUMP [4]or TESS [10], which concentrate primarily on message classification and summarization, ATRANS carefully analyzes every word in a message, producing a highly-detailed representation of its content. To the best of our knowledge, ATRANS is unique in its robust coverage of a domain at this level of detail.

Finally, we offer some implementational details. ATRANS is currently undergoing live testing at a major international bank. The system is implemented in the T dialect of LISP under the VAX/VMS operating system. The average processing time on a VAX-11/785 is under 20 seconds per telex.

REFERENCES

1. Birnbaum, L., and Selfridge, M. Problems in Conceptual Analysis of Natural Language. 168, Yale University Department of Computer Science, October, 1979.

2. Charniak, E. Ms. Malaprop: A Language Comprehension Program. Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Mass., August, 1977.

3. Cullingford, R. *Script Application: Computer Understanding of Newspaper Stories.* Ph.D. Th., Yale University, 1978. Research Report #116.

4. DeJong, G. *Skimming Stories in Real Time: An Experiment in Integrated Understanding.* Ph.D. Th., Yale University, May 1979. Research Report #158.

5. Dyer, M. *In depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension.* Ph.D. Th., Yale University, May 1982. Research Report #219.

6. Gershman, A. *Knowledge based Parsing.* Ph.D. Th., Yale University, 1979. Research Report #156.

7. Riesbeck, C. Conceptual Analysis. In *Conceptual Information Processing,* North-Holland, Amsterdam, 1975.

8. Schank, R.C. "Conceptual Dependency: A Theory of Natural Language Understanding". *Cognitive Psychology 3,* 4 (1972), 552-631.

9. Schank, R.C. and Abelson, R.. *Scripts, Plans, Goals and Understanding.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

10. Young, S. and Hayes, P. Automatic Classification and Summarization of Banking Telexes. Proceedings of the Second Conference on Artificial Intelligence Applications, IEEE Computer Society, December, 1985.