ROZMOWA :
a natural language retrieval system for lawyers

by

Danuta KAMINSKA-KEPA
Warsaw University - POLAND
Romuald KEPA
Warsaw technical University - POLAND
Ludmila OHLSON
Software system research center - SWEDEN

1. INTRODUCTION

We feel that time is ripe for computers to be equipped with natural language systems which can be used by persons who are not trained in any special computer language.

We are developing a system called ROZMOWA at the Institute of Informatics at Warsaw University. ROZMOWA includes Polish frontend with ability to understand and explicitely answer user's request basing on the large data base. It is designed to support work with specialized data base using a subset of real legal acts.Due to this the structure of ROZMOWA is not so complicated. Our system is a part of another system called DIALOG developed in WARSAW, too. DIALOG is a more general and versatile natural language system, but of course it is more research prototype of computer system than a system applicable in practice.

In spite of work being carried out using a set of the legal acts from environmental protection, the ideas of ROZMOWA can be directly applied to another domain.

This report describes the current state of two-years project aimed towards the development of natural language retrieval system for lawyers.

ROZMOWA is implemented in LISP language on IBM 370/145.

2.THE GOAL OF ROZMOWA

Our main goal is to allow a non-programmer to obtain information from data base with no prior training or experience. To realize this goal any ROZMOWA-type system must be able to understand a considerable set of user's natural language.

The system must of course, react to any user's sentence, but only questions or orders which describe legal act using its formal characteristics can cause system to retrieve the document from the data base. We must explain what the formal characteristics are. Each legal act, while published in the official gazette announcing current legislation, is described giving its title, text and some additional information such as : type of the document (act, resolution), promulgation, date of its pass, date of validity. This information plus publishers and the date of law abrogating constitute the formal characteristics of the document. Moreover, we give information of its membership to the subdomain of environmental protection. The user can retrieve any document or a set of documents specifying some of its formal characteristics. For example, if someone is interested in documents which were published in Dziennik Ustaw (Polish official gazette announcing current legislation) during a period 1.03.1955 – 15.07.1960 and which refer to animal protection, one can express this by :

> Give me all documents concerning animal protection which were published in Dziennik Ustaw between 1 March 1955 and 15 July 1960.

In this case we have three values of the formal characteristics, i.e. publisher (Dziennik Ustaw), the date of validity (1.03.1955 - 15.07.1960) and subdomain (animal protection). Using this information the system will retrieve a set of legal acts.

The questions or orders which can be addressed to ROZMOWA have to be sentences or phrases in grammatical Polish terminated by question mark, period or exclamation mark, as appropriate, and observing the comma punctuation.

Let us see how our system is organized and how it works.

3. THE STRUCTURE OF ROZMOWA

The processing of the user's request is divided into three main phases : natural language processing, evaluation and answer generation. Graphically the structure of our system is shown below in figure 1.

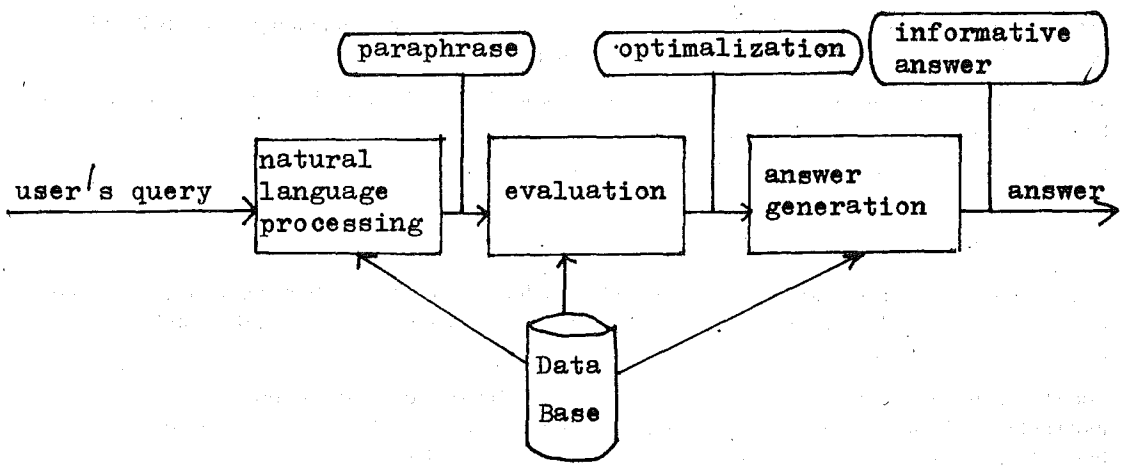

Fig. 1 – An overview of the ROZMOWA system

Below we will present individual parts of the system

4. NATURAL LANGUAGE PROCESSING

The natural language processing of the system is realized with the help of ATN (Augmented Transition Network) [1] [4].

Our installation of ATN interpreter and compiler has been realized on the basis of T. Finin's documentation [3] and was performed as the whole system in LISP. ROZMOWA system has used the ATN mechanism for natural language processing, i.e. to translate the natural Polish question into

SQUARE formula [2]. SQUARE (Specifying Queries as Relational Expressions) is a set-oriented data sublanguage for expressing queries to data base consisting of a collection of time-varying relations. This language does not require the sophisticated mathematical machinery of the predicate calculus in order to express simple references to the data base. Since we especially deal with this kind of demands in our system we mainly use a subset of SQUARE. The ATN grammar with the help of which translation to SQUARE is performed is a semantic grammar but it differs from R. Burton, W. Woods or T. Finin understanding [3] [6].

The semantic nets for processing a sentence pertaining to legal acts consist of main network for the sentence, called request net and additional net with the help of which the partial information such as date is described.

During the natural language processing the dictionary module is used. Every word in the dictionary is written as an s-expression in which the first element is the word and the rest- the syntactic and semantic features. The first element of the feature list is the category name and the second is the word representative. Semantic categories usually show to which formal characteristic of the document the word belongs. The categories may be specific representatives of the key words. We have not any morphological analysis because the morphological analysis of Polish language is very complicated and up to now there is no good method to do it. Our dictionary is small enough to store the inflexional forms of words. Words as dokument, dokumenty, dokumentami (word document with inflexional case endings) are represented in the dictionary as s-expression.

((dokument dokumenty dokumentami) substitute (dok))

Some words in natural language queries, from our point of view, do not contain any information, for instance words please, give or phrases I would like or I want to know, etc. Such words or phrases are substitute by the dictionary manager by the atom NIL in the LISP meaning. In the dictionary some of syntactic features are also stored, for instance the prepositions which played a really important role in the translation from natural language into SQUARE formula.

Let us see an example :

The user asks the system :

Gife me documents which were promulgated by Sejm until 1970.

In this case we have two formal characteristics which describe a set of documents demanded (Sejm, until 1970).

After natural language processing we get the SQUARE formula :

(MAPA DOK (TITLE PUBLISHER)(ORGAN-OF-POWER DATE-OF-PASS)((SEJM)(LT 1970)))

It has to be explained that when the user asks about documents and he did not specify which characteristics of the documents he wanted the system gives him title and the publisher of the documents demanded.

Directly from this structure the paraphrase of the user's question is generated. For our example it has

the form :

> You ask me about the title and publisher of documents which were promulgated by Sejm and which date of pass is no later than 1970.

The reader can immediately find the difference between the user's query and the paraphrase.

In our system the paraphraser has two main goals in its output :

- to identify the word sense selected for potentially ambiguous terms,
- to show how the system understood the user's query.

The paraphraser play really important role when anaphoric references are used.
Let us imagine the user question :

(1) What is the title of a document published in Dziennik Ustaw number 45, item 9 in 1970 ?

and the answer is

> Water Law

Then the user asks a question with the anaphoric reference :

(2) When was it pass ?

where *it* means the document which the system retrieved as the answer for the preceding order (1). For this kind of questions the paraphraser has to process the anaphoric references. So the appropriate paraphrase for question (2) will be :

> You ask me when the document published in Dziennik Ustaw 45/1970 item 9 was passed.

The paraphrase is presented to the user for approval. If it is correct (corresponds to the user's intent) than an appropriate SQUARE formula will be sent to the evaluation phase. Otherwise the user will be able to rephrase the input to circumvent the problem.

5. EVALUATION PHASE

The evaluation phase uses the SQUARE formula generated in the previous stage to search the data base and construct the answer. The data base uses a relational model of data [5]. We distinguish two basic parts of our data :

- real documents (texts of legal acts pertaining to environmental protection)
- formal characteristics of these documents.

These data are grouped into four relations.

The DOK relation has a row for every legal act, giving its : title, type of document, promulgation, date of pass, date of validity, publisher, date of its abrogating, information about amendments and serial number - the primary key for this relation.

The PART relation has a row for every item of legal act, giving its : serial number, description of the item, text of given part of the document. The first two attributes compose a primary key of this relation.

The DOMAIN relation has a row for every legal act, giving its : serial number, description of the document item, subdomain to which the document belongs. The primary key of this relation consists of all attributes because one document item can belong to many subdomains.

The last relation AMEND describes amendments of documents. This relation has a row for every document to which an amendment exists, giving its serial number and a serial number of the document which contains this amendment.

6. SEARCHING

Before executing the SQUARE query obtained from the natural language processing, the system will try to optimize this query. It means that the query will be reorganized in this way that a new query will take a shorter time to execute. Using some of the algebraic laws that apply to relational algebra operators [9] we optimize SQUARE formula. Our optimizer is based on the Ullman [9] algorithm.

After improvement, the searching module tries to choose the most effective algorithm which retrieves the documents relevant to the user's query. To achieve this purpose we examine at first conditions which deal with domain for which inverted lists exist. In this way we essentially restrict a set of documents, which will be verified taking into consideration the conditions connected with the remaining domains. This method in conjunction with direct access organization of the files essentially shortens the answer time of the system.

7. RESPONSE PHASE

During the process of retrieving the relevant documents from the data base we construct a special file, which contains selected information. The form of that answer, however, is not suitable to be given back to the user. It contains a lot of abbreviations and physical addresses. The response phase must make it readable for the user. In the natural language data base query system the organization of the response should be determined by the structure of the questions and not by the structure of the data base. Sometimes this response is negative. It is the situation where a user's request for data cannot be answered in the desired way because the data base does not contain the data requested. Let us see an example :

Which documents were published in 1790 ?

The direct answer for this query is none. But people typically do not ask questions to which they expect a negative of trivial response (such as zero, none). The system should be able to detect that the initial query in the dialogue incorrectly presumed something and response appropriately.

In our example there is a wrong assumption that in the data base are very old documents. In fact our data base contains only documents passed during 1900-1981. So the informative answer for that query should be :

There are only documents which were published between 1900-1981.

An informative answer to a failing query (when a direct answer is negative) will adjust the user's wrong assumptions about the content of the data base and often will free him from the need of asking additional queries [7] [8]. If we ask the question :

(1) Give me all documents published in XXX during 1950-1965.

let the direct answer be :

None (because there is no publisher called XXX).

The user can ask another question such as :

(2) Give me the documents which were published in XXX

still the answer is :

None

and the last question :

(3) Is there XXX in data base ?

No

As we see if we responded to the first question (1) in informative way, the user would not ask question (2) and (3). Answering informatively is essential to natural language system in a practical environment (such as ROZMOWA), because the fact that natural language is used in the interaction will imply to the users that the normal cooperative conventions followed in a human dialogue will be observed by the machine. Up till now ROZMOWA does not answer in the cooperative way, but we have been working on it. It will be the most important task of the response phase.

8. THE FINAL REMARKS

The natural language retrieval system ROZMOWA is used as testbed for verification of well-known mechanisms and methods. A number of factors contribute to make our problem much easier to be solved than the general problem of understanding unconstrained natural language. The most important is of course limited scope and domain of the system.

Danuta Kamińska-Kepa, *Institute of Informatics, Warsaw University*
Romuald Kepa, *Institute of Mathematics, Warsaw Technical University*
Ludmila Ohlsson, *Software System Research Center, Artificial Intelligence Laboratory.*

REFERENCES

[1] Bobrow B., Fraser J., *An augmented State Transition Network Analysis Procedure, Proc. IJCAI,* 1969.

[2] Boyce R.F. et al., *Specifying Queries as Relational Expressions, Proc. of IFIP Working Conference on Data Base Management,* 1974.

[3] Burton R.R., *Semantic Grammar : A Technique for Efficient Language Understanding in Limited Domain,* Doctoral Dissertation at Univ. of California, Irvine 1970.

[4] Burton R., Woods A., *A Compiling System for Augmented Transition Network, Proc. of COLING,* 1976

[5] Codd E.F., *A Relational Model for Large Shared Data Banks* CACM 13, p. 377-387, 1970

[6] Finin T.W., *An Interpreter and Compiler for Augmented Transition Networks,* Univ. of Illinois, Urbana Illinois 1977

[7] Janas J.M., *On the Feasibility of Informative Answer,* in Advances in Data Base Theory, vol. 1, Plenum Press, N.Y. 1981

[8] Kaplan S.J., *Cooperative Responses from Portable Natural Language Data Base Query System.* Doctoral Dissertation at Univ. of Pennsylvania, Philadelphia, Pennsylvania 1979

[9] Ullman J.D., *Principles of Database Theory, Pitman Publ. Ltd., 1980*

((UTWORZENIE WOLINSKIEGO PARKU NARODOWEGO)
(ROZPORZADZENIE)
($$$1960.03.03$)
($$$1960.03.03$)
((DZ.U)(NR 14/1960) (POZ 79))
( )
(PAR1 PAR2 PAR3 PAR4)
(PARAGRAF 1 TWORZY SIE WOLINSKI PARK NARODOWY O OBSZARZE OK 4691 HEKTARA ZWANY DALEJ PARKIEM POLOZONY W POWIECIE WOLINSKIM W WOJEWODZTWIE SZCZECINSKIM")
(PARAGRAF 2 W SKLAD PARKU WCHODZA 1/OBSZARY OBJETE OCHRONA REZERWATOWA O POWIERZCHNI OKOLO 4676 HEKTARA 2/OBSZARY WLACZONE DLA CELOW ADMINISTRACYJNYCH PARKU O POWIERZCHNI 15 HEKTAROW")
(PAR.3.USTEP.1 NA OBSZARZE PARKU OKRESLONYM W PARAGRAFIE 2 USTEP 1 PUNKT 1 WSZELKIE CZYNNOSCI GOSPODARCZE ICH CHARAKTER ZAKRES I SPOSOB WYKONYWANIA MUSZA BYC SCISLE DOSTOSOWANE DO POTRZEB I CELOW OCHRONY PRZYRODY"
USTEP 2 OGRANICZENIA WYNIKAJACE Z PRZEPISU USTEPU 1 W STOSUNKU DO TERENOW ZABUDOWANYCH I POZOSTAJACYCH POD UPRAWA ROLNA LAKOWA PASTWISKOWA LUB GOSPODARKA RYBACKA NIE DOTYCZA CZYNNOSCI GOSPODARCZYCH KTORYCH WYKONYWANIE KONIECZNE JEST ZE WZGLEDU NA RACJONALNE UZYTKOWANIE TYCH TERENOW% OD DECYZJI DYREKTORA PARKU W SPRAWACH UZNANIA CZYNNOSCI ZA KONIECZNE ZE WZGLEDU NA UZYTKOWANIE TERENU SLUZY ODWOLANIE DO WLASCIWEGO DO SPRAW OCHRONY PRZYRODY ORGANU PREZYDIUM WOJEWODZKIEJ RADY NARODOWEJ")
(PAR.4.UST.1 POSZCZEGOLNE CZESCI OBSZARU PARKU OBJETE OCHRONA REZERWATOWA PODLEGAJA OCHRONIE SCISLEJ REZERWATY SCISLE LUB OCHRONIE CZESCIOWEJ REZERWATY CZESCIOWE% W SKLAD REZERWATOW NIE WCHODZA TERENY ZABUDOWANE LUB POZOSTAJACE, POD UPRAWA ROLNA LAKOWA LUB PASTWISKOWA ORAZ TERENY WYKORZYSTYWANE PRZEZ WOJSKO"))

Fig. 7 The legal act after intial processing

(UWPN ROZPORZADZENIE RM 1960%.03%.03 1960%.03%.03 DU 14 1960
79 NIL (UPN) (PAR1 PAR2 PAR3 PAR4)(1265 1266 1268 1271 1281))

Fig. One row in DOK relation

```
34_(ROZMOWA)
                >> KONWERSACYJNY SYSTEM ROZMOWA <<

 >> CZY CHCIALBYS PRZESLEDZIC TRACE ANALIZY JEZYKOWEJ
...NIE
 >> PYTANIE UZYTKOWNIKA
 >> PAMIETAJ, JEZELI CHCESZ  ZAKONCZYC SESJE NAPISZ - KONIEC

...PODAJ DOKUMENTY UCHWALONE 8 MAJA 1977 ROKU .

 >> CZY CHCIALBYS ZOBACZYC PARAFRAZE PYTANIA ?
...TAK
 >> PARAFRAZA PYTANIA UZYTKOWNIKA :
    NALEZY PODAC TYTULY I ADRESY PUBLIKACJI DOKUMENTOW UCHWALONE
    W 1977.05.08

 >> CZY CHCESZ ZOBACZYC FORMULE
...TAK
 >> (MAPA DOK (TYTUL PISMO NUMER ROCZN POZYC)
       UCHWA
       ((EO 1977.05.08)))
 >> ODPOWIEDZ SKLADA SIE Z 7 ELEMENTOW. CZY MAM WYSWIETLIC ODPOWIEDZ (TAK NIE)?
...NIE
 >> CZY CHCIALBYŚ PRZESLEDZIC TRACE ANALIZY JEZYKOWEJ
...NIE
 >> PYTANIE UZYTKOWNIKA
 >> PAMIETAJ, JEZELI CHCESZ  ZAKONCZYC SESJE NAPISZ - KONIEC

...KTORE Z NICH ZOSTALY UCHWALONE PRZEZ SEJM .

 >> CZY CHCIALBYS ZOBACZYC PARAFRAZE PYTANIA ?
...TAK
 >> PARAFRAZA PYTANIA UZYTKOWNIKA :
    NALEZY PODAC TYTULY I ADRESY PUBLIKACJI DOKUMENTOW UCHWALONE
    W 1977.05.08 ORAZ  WYDANYCH PRZEZ  SEJM
 >> CZY CHCESZ ZOBACZYC FORMULE
...TAK
 >> (MAPA ANAPHOR (TYTUL PISMO NUMER ROCZN POZYC)
       (UCHWA ORGAN)
       (((EO 1977.05.08))
        (SEJM)))
 >> ODPOWIEDZ SKLADA SIE 1 ELEMENTU. CZY MAM WYSWIETLIC ODPOWIEDZ (TAK NIE)?
...TAK
 >> PRAWO WODNE, DZU 1977.05.08/1
 >> CZY CHCIALBYS PRZESLEDZIC TRACE ANALIZY JEZYKOWEJ
...NIE
 >> PYTANIE UZYTKOWNIKA
 >> PAMIETAJ, JEZELI CHCESZ  ZAKONCZYC SESJE NAPISZ - KONIEC

...KONIEC
 >> DZIEKUJE
```

Fig. 8 The part of ROZMOWA conversation