COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

TRANSFORMATION OF NATURAL LANGUAGE INTO LOGICAL FORMULAS

Leonard Bolc and Tomasz Strzalkowski

Institute of Informatics
Warsaw University
PKiN, pok. 850
00-901 Warszawa, POLAND

This paper presents an attempt of elaboration of a full parsing system for Polish natural language which is being worked out in the Institute of Informatics of Warsaw University. Our system was adapted to the parsing of the corpus of real medical texts which concern a subdomain of medicine. We made use of the experience of such famous authors as (6), (7), (8), (9), (10), (11), (12), (13), (14).

INTRODUCTION

The system described below could be used as an interface of natural language information systems, natural question-answering systems, expert systems or automatic understanding of texts. The authors paid close attention to the syntactical and semantical constraints of medical dialogue so that the system would be used by physicians without previous preparation. Although a subdomain of medicine is a current system application, the change or development of the conversation field may be facilitated. It requires only that a new dictionary will be established and some expert parts of semantical interpreter will be changed.

Our system contains two stages: syntactical analysis and semantical interpretation. Both stages cooperate with each other in such a way that the second stage checks up on the correctness of syntactical structures which have been built by the first one. Finally, the parser produces a formula of First Order Predicate Calculus which corresponds to the input sentence. Other outputs as MINSKY frames or FUZZY formulas are considered.

We used the CATN method (Cascaded ATN) (14) to implement the system. The CATN possesses all of the advantages which proved true in natural language processing. A high degree of universality is a very important feature of the system.

Here are some examples of sentences the parser can understand:

Alkohol podany doustnie powoduje wzmozone wydzielanie gastryny.
(Alcohol given per os cause greater secretion of gastrin.)
Alkohol zwieksza wydzielanie soku trzustkowego.
(Alcohol increases pancreatic juice secretion.)
Gastryna jest hormonem powodujacym wydzielanie kwasu solnego w zoladku.
(Gastrin is a hormon which cause gastric HCL secretion.)

Sekrytyna i pankreozymina stymuluja czynnosc wewnatrzwydzielnicza trzustki.
(Secretin and pancreocymin stimulate endicrin activities of pancreat)
wzrost napiecia miesniowki dwunastnicy moze byc przyczyna wzrostu cisnienia w przewodach trzustkowych.
(The increas of tonus of the tunica muscularis may cause higher pressure in the pancreatic ducts.)
Dlugotrwale dzialanie alkoholu powoduje prawdopodobnie bezposrednie uszkodzenie komorek wydzielniczych trzustki.
(Long action of alcohol propably cause direct injury in the pancreatic endocrin cells.)
Jakie sa kliniczne objawy ostrego zapalenia trzustki?
(How appears the clinical symptoms of acute pancreatifis?)
Co stymuluje czynnosc wewnatrzwydzielnicza trzustki?
(What stimulates endocrin activities of pancreat?)

CATN AS A TOOL

A Cascaded Augmented Transition Network looks like two or more "cascades" which succesively perform the same information. Each of them is an ATN grammar (1) which has, in addition, a new action called TRANSMIT. The TRANSMIT action may be set on every arc and causes a piece of information to be sent from the current "cascade" to the lower one. Whenever a TRANSMIT occurs each information about the current "cascade" is saved on the stack while the parser operates on the lower "cascade" until new information or data is required. Then the higher "cascade" is activated from the same point it has been stopped.

Two stages of our parsing system correspond to the CATN-cascades. In the present realisation the structure popped from the syntactical stage is TRANSMITed into semantic interpretation because a free word-order of Polish sentences prohibits another solution. Particulary, the places of the subject and the main verb in the sentence may be varying.

If the second stage is not able to find an appropriate interpretation for syntactical structure the first stage is activated to build an alternative parsing. When such a parsing cannot be rebuilt the parser fails.

In the other implementation of CATN we used the Earley's algorithm, a well-known context-free parsing method (10). In this case the syntactical analyser produces all possible parsings at once. The semantical interpreter has to verify them and reject each meaning-less parsing.

THE FIRST STAGE - SYNTACTICAL ANALYSIS

A surface structure of a sentence is received after the First Stage of the parser was applied to an utterance. It means that such elements as VERB/ACTION, SUBJECT, OBJECT (direct and indirect), PREPOSITION PHRASES etc. are found out.

Polish natural language is a typical example of a flexional language. One of its most characteristic features is a free word-order in a sentence. It is very important for the parser to know each lexical parameter of nouns, adjectives, adverbs, numbers, prepositions etc. These parameters are number, gender, case, person and degree. They

are carried over the whole phrase and decide about the role of the phrase in the sentence. A flexional form of the main verb also influences the construction of the sentence. Especially, however, the flexional properties of the main verb could help the parser to find out the subject and the direct object.

These problems and several others as post-modifiers problem, wh-movement, conjunction, etc. were solved succesfully.

The syntactical analysis comprises a wide subset of Polish language eg. simple affirmative sentences and questions, complements and relative clauses and certain types of complex sentences. We had to take into account a number of special properties of the medical dialect which rarely occur in a common conversation. The grammar is able to parse not only the common Polish but the "medical" Polish as well. It means, among others, a great deal of participles, gerunds, modal verbs (eg. moze - could, powinien - should) and vague adverbs (eg. prawdopodobnie - propably, czesto - frequently, rzadko - rarely, czasami - sometimes).

The syntactical analyser transforms an input sentence into an unflexional and ordered form. Some examples of the output of the First Stage are given below. The |-mark divides the whole sentence into phrases. An empty place between two |s points out a missing phrase. The S and END flags indicate the beginning and the ending of each simple clause in the sentence. If the DCL flag occurs just after S-mark in the top-level clause the sentence is dealt as an assertion. In a question there are one or more question words instead. The MODIFIERS flag divides a direct object (if any) into the main phrase and post modifiers. This last flag is an important one because the head word of a direct object phrase may be a predicative element of the clause. (eg. byc przyczyna - to be a cause). Notice, that a predicative element of the top-level clause becomes the main predicative element of the whole sentence.

alkohol podany doustnie powoduje wzmozone wydzielanie gastryny.
(alcohol given per os cause greater secretion of gastrin.)

(S DCL | | | POWODOWAC | | ALKOHOL | S | | | PODAC |DOUSTN* | |

ALKOHOL MODIFIERS | | | END | S | | | WYDZIELANIE | WZMOZON* | |

GASTRYNA MODIFIERS | | | END | | | END)

alkohol zwieksza wydzielanie soku trzustkowego.
(alcohol increases pancreatic juice secretion.)

(S DCL | | | ZWIEKSZAC | | ALKOHOL | S | | | WYDZIELANIE | | |

TRZUSTKOW* SOK MODIFIERS | | | END | | | END)

co stymuluje czynnosc wewnatrzwydzielnicza trzustki?
(what stimulates edocrin activities of pancreat?)

(S CO | | | STYMULOWAC | | | WEWNATRZWYDZIELNICZ* CZYNNOSC MODIFIERS

TRZUSTKA | | | END)

Nevertheless, because such information is not sufficient an interpretation in the Second Stage is needed.

The First Stage contains the main ATN net named SENTENCE which can perform Polish natural sentences. There are four special subnets: NOUN_PHR, ADJ_PHRA, ADV_PHRA, Q_EXPR which can recognize different types of phrases eg. nominal phrases, adjectival phrases, adverbial phrases and question expressions respectively.

The First Stage uses a syntactical dictionary which contains the flexional forms of the words,

THE SECOND STAGE - SEMANTICAL INTERPRETATION

When the syntactical analysis has been completed the Second Stage of the parser tries to find out a semantical interpretation ofthe syntactical structure. The main predicative element of this structure (eg. VERB/ACTION or OBJECT) creates one or more instances of framework describing an event. That framework looks like a pattern-concept pair (8), (12), nevertheless there are more frameindicating verbs (7). For example the following verbs and verb expressions: powodowac (cause), stymulowac (stimulate), prowadzic do (conclude), byc przyczyna (to be a cause), byc skutkiem (to be a result), etc. refer to the conceptualization #IMPLY and podac (to give), stosowac (to apply), etc. to the conceptualization #APPLY.

The pattern determines which phrases may be expected round the predicate and which of them must occur. The interpretation process is driven by such a pattern so it is called expectation-driven. It may be called structure-driven too because there are structural conditions in the pattern which must hold true during the parsing time.

A concept is a notation that represents the meaning of a clause. Together this pair associates different forms of an utterance with its meaning.

The #APPLY conceptualization looks like:

```
(APPLY TYPE TREATMENT
       AGT  ( () HUM OPT )
       OBJ  ( () MEDIC OBL )
       MANNER  ( () MOA OPT )
       CONCEPT  (BUILDQ
                ((#APPLY 3) + + +) AGT OBJ MANNER)
)
```

where TYPE is an indicator which points out that the described event is a treatment. AGT, OBJ, MANNER determine that there may be three phrases round the predicate, but only one of them must occur in an utterance. (OBL means obligatory parameter, OPT - optional one). None of these phrases could have a preposition before it - (). The AGT-phrase (agent that applies something) must be a human; the OBJ-phrase (object which is applied) must be a medicament; the MANNER slot may be filled when the manner of application is specified (eg. doustnie - per os). The CONCEPT indicator describes the way an atomic formula has to be built. As it is seen above, we shall receive a 3-nary pre-

dicate called #APPLY which arguments will be constructed during the interpretation process. The BUILDQ function is a special ATN form which provides BUILDing of Quoted expressions (see (1) for details).

A filling of frame slots is done after the syntactical and semantical requirements were satisfied. When the whole pattern were completed an atomic formula would be generated. Therefore, the interpretation process is an attempt to squeeze the syntactical structure of a sentence into one or more instances of framework of an event. Beside the main predicate(s), a great deal of additional information would be joined the output formula. These facts are stored in part in pattern-concept pairs and in expert subnets of interpreter. They create a system knowledge. It is necessary for the system to have such a knowledge because none of the real text corps is able to describe completely a domain of the real world.

A great deal of context information may also be used from the special context stack. It helps to solve the problems of pronoun references and ellipsis.

If the "squeezing" could not be made the First Stage is actvated again.

In addition, the semantical dictionary is appended to the Second Stage. It keeps all patterns of frameworks mentioned above. It contains some special entities too for indicating the reference between verbs and patterns.

The Second Stage also contains the main ATN net named FORMULA. It guides the interpretation process and controls the semantical correctness of utterances. There are aslo some expert nets which can recognize special medical expressions (eg. names of sicknesses and symptoms organs, treatments, etc.). These subnets are a changeable part of the system and they decide about the system knowledge. The expert subnets may communicate with the main net through the middle level of interpreter - the CASES net. This net handles nomimal phrase structures eg. prepositions, conjunctions and post-modifiers.

The Second Stage produces a formula of the First Order Predicate Calculus corresponding to the input sentence. The formula has an implicative form where the main predicate of the utterance is a conclusion and other generated facts are presumptions.

Two generated formulas are given below. First of them is an assertion, the remaining one denotes a question. They are in LISP notation so a clarification is needed. IMPLSYM and KONJSYM marks are the logical operators IMPLY (=>) and AND (&). An integer just after the KONJSYM mark indicates how many factors were joined. Each predicate name is preceded by a hash-mark (#) and followed by an integer to indicate a number of arguments. Arguments look like a pair or triple which determines the type of argument, the name of a variable and a constant (if any) respectively.

Alkohol zwieksza wydzielanie soku trzustkowego.
(Alcohol increases pancreatic juice secretion.)

```
(IMPLSYM (KONJSYM 3
           ((#BADMEDIC 1)(MEDIC X0002585))
           ((#MEDICAMENT 2)(MEDIC X0002585)(mname X0002586 ALKOHOL))
```

```
        (IMPLSYM (KONJSYM 4
                    .((#ORGAN 2)(ORGAN X0002589)(ONAME X0002590
                                                    TRZUSTKA))
                     ((#WYDZIELNICZ*-NARZAD 1)(ORGAN X0002589))
                     ((#JUICE 1)(LIQUID X0002588))
                     ((#LIQUID 3)(LIQUID X0002588)
                                 (LNAME X0002591)(ORGAN X0002589)))
                  ((#SICKNESS 4)(SICK X0002587)(STYPE X0002592 FI)
                                (SNAME X0002593 wydzielanie)
                                (BODY X0002588))))
      ((#RAISE 2)(etio X0002585)(SYMPTOM X0002587)) ))
```

Co powoduje alkohol podany doustnie?
(What damages cause alcohol drinking?)

```
((X39) (IMPLSYM (KONJSYM 3
                  ((#BADMEDIC 1)(MEDIC X30))
                  ((#MEDICAMENT 2)(MEDIC X30)(MNAME X31 ALKOHOL))
                  (IMPLSYM (KONJSYM 2
                              ((#BADMEDIC 1)(MEDIC X36))
                              ((#MEDICAMENT 2)(MEDIC X36)
                                              (MNAME X37 ALKOHOL)))
                           ((#APPLY 3)(anim X38)(MEDIC X36)
                                      (MANNER X33 DOUSTN*))))
               ((#IMPLY 2)(ETIO X30)(SICKNESS X39)) )))
```

The parser can also produce other kinds of formal representation of natural lamguage.

CONCLUSION

The parsing system described above is an attempt to build an universal parser for natural language analysis, The authors incline to the fashionable thesis that the syntactical and semantical componemts should act in the same time, nevertheless with a domination of the syntax over the semantics. This remark is an important one for the Polish language. This approach, however, provides no less efficiency of the parsing process than in the semantic-dominate systems (7), (8) (11), (12) and certainly greater universality of the system. This provides among others most of the advantages of regularity of natural language.

BIBLIOGRAPHY

(1) Bates, M, The Theory and Practice of Augmented Transition Network Grammars, in (2).
(2) Bolc, L., (ed), Natural Language Communication with Computers, in Lecture notes in Comp. Sci., vol 63,(Springer-Verlag,Berlin, Heidelberg, New York 1978).
(3) Bolc,L. (ed), Natural Language Based Computer Systems, (Hanser Verlag and Macmillàn Press, London 1980).
(4) Bolc, L.(ed), Natural Language Question Answerig Systems,(as (3)).
(5) Bolc, L.(ed), Representation and Processing of Natural Lnaguage, (as (3)).
(6) Burton, R., Brown, J.S.,Semantic Grammars: A Technique of Constructing Natural Language Interfaces to Industral Systems, (BBN Rep. No. 3587, Bolt Beranek and Newman Inc. Cambridge MA 1977).
(7) Carbonell, J.G., Multy-Strategy Parsing,(DEpt. of Comp. Sci., Carnegie-Mellon Univ., Pittsburgh PA, 1981).

(8) Gershman, A.V., Knowledge-Based Parsing,(Reaearch Rep. 156, Yale University, Dept. of Comp. Sci, 1979)
(9) Landsbergen, J., Adaptation of Montague Grammar to the Requirements of Parsing, (reprint from MC Tract 136, Formal Methods in the Study of Language, J.A.G. Groenendijk, T.M.V. Jassen, M.B.J. Stokhof (eds.) 1981).
(10) Martin, W.A., Church, K.W., Patil, R.S., Preliminary Analysis of a Breadth-First Parsimg Algorithm, (MIT Laboratory for Comp. Sci., 1981).
(11) Schank, R.C., Lebowitz, M.,Birnbaum, L.A., Integrated Partial Parsing, (Research Rep. 143, Yale Univ., Dept. of Comp. Sci.1978)
(12) Wilensky, R., Arens, Y., PHRAN - A Knowledge-Based Approach to Natural Language Analysis,(Dept. of Comp. Sci., Univ. of California, Berkeley 1980).
(13) Woods, W.A., An Experimental Parsing Syntem for Transition Network Grammars, (BBN Rep. No. 2362, Bolt Beranek and Newman Inc. Cambridge MA, =1972).
(14) Woods, W.A., Cascaded ATN Grammars, (in American Jrnl. of Comp. Linguistics, vol. 6, no. 1, 1980).