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

NATURAL LANGUAGE UNDERSTANDING AND THE PERSPECTIVES OF QUESTION ANSWERING

Petr Sgall

Department of Applied Mathematics
Faculty of Mathematics and Physics
Charles University
Prague
Czechoslovakia

A method of automatic answering of questions in natural language, based only on input texts and a set of rules of inference, is described. A first experimental system including a grammatico-semantic analysis of the input texts and questions, a procedure of inferencing, a search for appropriate answers to individual questions and a synthesis of the answers are being implemented, mainly in the language Q and PL/1. The output of the analysis, the underlying representations of the utterances of the input text, serves as a base of the knowledge representation scheme, on which the inference rules (mapping dependency trees into dependency trees) operate.

The important, though partial possibilities of automatic understanding of natural language gave rise to different kinds of experimental systems, ranging from sophisticated systems of machine translation through various kinds of modelling of dialogue (with robots, data bases, etc.) to question answering.[1] From a linguistic viewpoint the main challenge consists in attempting to transfer the burden of the communication between humans and computers to the latter, who should be able to react in an appropriate way to the user's input texts formulated in her or his native language, without serious restrictions. The necessity of thousands of human beings preparing data "for computers" (not only encoding messages, but also compiling data bases) should be removed.

This challenge constitutes one of the central tasks of modern linguistics; an explicit description of the main features of the language system, which is necessary for these purposes, must be based on a sound theoretical framework suitable for the description of grammar as well as of the linguistically patterned aspects of semantics and pragmatics. A close cooperation of linguistics with logic, computer science and cognitive science has become urgent. This task presents also an effective way of checking the results of theoretical linguistics in various important fields.

These considerations have led the group of algebraic linguistics in Prague (now belonging to the department of applied mathematics, faculty of mathematics and physics, Charles University) to start working on an experimental system based on the approach called TIBAQ (Text-and-Inference Based Answering of Questions).[2] Its four main procedures are (1) grammatico-semantic analysis, (2) rules of inference, (3) identification of a full (direct) or partial answer, and (4) synthesis; see the overall scheme in Fig. 1.

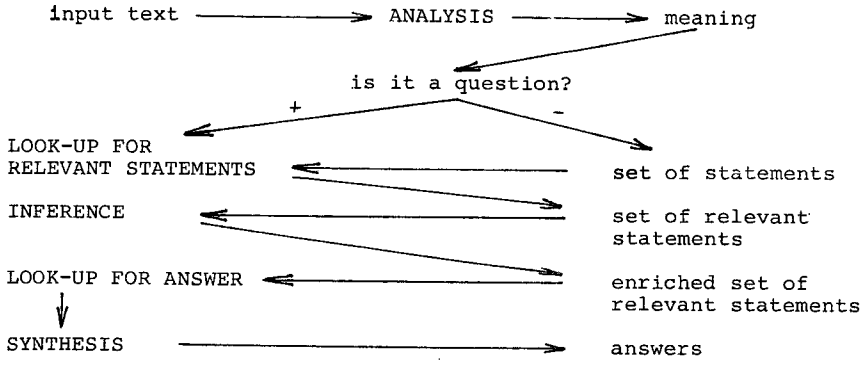

Fig. 1
An overall scheme of a system based on the method TIBAQ

(1) The automatic grammatico-semantic a n a l y s i s[3] is being prepared in such a form that it can handle Czech and English polytechnical texts (papers, reports, monographs) in their usual shape, and also questions formulated in Czech. Thus there will be no need for the user to "cope with the needs of the computer system". The procedure of analysis has the following characteristic properties distinguishing it from a mere parsing procedure:

(i) The analysis procedure is based on a systematic theoretical account of the structure of natural language, the functional generative description; this linguistic approach, elaborated in the Prague group of algebraic linguistics,[4] makes use of the results of the empirical research carried out in the frame of European structural linguistics, and also of the methodological requirements formulated by Chomsky and the different wings that developed from his school. The resulting linguistic approach is used as a general base ensuring that the particular practical solutions (in ambiguity removal, etc.) chosen for a restricted area can be replaced by more generally valid sets of rules, whenever it appears as necessary to cross the boundaries of this narrow area (e.g. when applying the method to a new kind of texts, to a new polytechnical domain, etc.). This is ensured thanks to the universal character of natural language and to the fact that the linguistic framework (if appropriately chosen) provides means for an adequate description of all its subdomains (cf. Hajičová and Sgall, 1980a).

(ii) In connection with this requirement the analysis procedure is designed to transfer the input sentences from their outer form to a disambiguated notation of their meanings (which can be identified with their underlying structures, in the framework of functional

generative description). The level of meaning of sentences includes such syntactic units as Actor, Objective, Addressee and other participants or cases, Manner, Instrument, Place, Direction and other free adverbial modifications, as well as lexical and morphological meanings (the latter including e.g. number, tense, modalities). This level is formulated as a linguistic counterpart of intensional structure, which makes it possible to define the concept of strict synonymy of expressions and to ensure an algorithmic transition to a postulated universal formal language of intensional logic[5] (among the trends that started with Montague, our account of meaning stands close to that by David Lewis, though the form of formal language we prefer has much in common with Tichý's framework). The representations of the meanings of sentences serve as the main components of knowledge representation in the semantic networks of the systems based on the method TIBAQ. They can be illustrated by the representation in Fig. 2.

(iii) As can be seen from this representation, our approach works with dependency trees as the form of meanings of sentences. This allows us to work with relatively simple underlying structures in which such notions as "head" and "modifier", or "noun" phrase vs. "verb" phrase, as well as the relations described by Fillmore as cases find an economical treatment.

(iv) Not only the roles of the elements of syntactic relations, but also the topic-focus articulation of sentences finds its proper place in the representations yielded by this procedure of analysis. Also the whole pragmatically based interplay of topic, focus, contextual boundness and communicative dynamism, as combined with the recursive properties of sentence structure can in principle be rendered in the chosen form of representations of the meanings of sentences.[6] Analysis of written texts does not allow for a complete identification of all the items relevant for the topic-focus articulation, and the present form of our algorithms gives results which are not fully reliable, but the errors appear to be neither too numerous nor too grave for the given purpose. The main rules consiat in understanding the parts of a sentence standing to the left of the finite verb as belonging to the topic, while the verb itself (if it is not semantically void, as the copula, or become, carry out, etc.) and the elements following it are classed as belonging to the focus in the Czech polytechnical texts.[7] Such a treatment appears as sufficient for ensuring that those cases in which the topic-focus articulation is semantically relevant will be handled appropriately. This concerns the relative scopes of quantifiers in such sentences as Every car has several wheels and the 'holistic' understanding of the topic e.g. in Smoking is dangerous, as well as Kuno's "exhaustive listing" and the difference between thetic and categorical judgements; even more important is the relevance of the boundary between topic and focus for the determination of the scope of negation, and thus also for the identification of presuppositions in some cases: Many arrows didn't hit the target does not imply that the target wasn't hit by many arrows, and The king of France didn't come to COLING 82 does not presuppose the existence of a king of France. The relevance of topic and focus for natural language understanding is most clearly recognized in connection with the assignment of reference to definite noun phrases (and other expressions).

(v) The procedure of analysis provides also for a treatment of the interconnections between the individual assertions (which are stored in the shape of the meanings of sentences). This is done by means of two main devices: first, in the representation of each lexical meaning in the lexicon there is an indication of the relations

of synonymy and hyponymy (subordination, superordination) of the given item to others, and also semantic features are used (for a partial modelling of the object domain pertinent to the treated area of polytechnical texts);[8] second, the relation between an object and the occurrences of expressions referring to it in the texts is handled by means of a register or concordance, supplying addresses of all the occurrences of a given unit in the whole set of knowledge representation.

After having examined different means of implementation of the analysis procedure, esp. Kay's parser, Wood's ATN, the Grenoble system and others, we decided that among the systems actually available to us the framework elaborated in the T.A.U.M. group, based on Colmerauer's Q-systems, can serve best our aims. Thanks to the Canadian colleagues we got the possibility to implement Q-systems (through Fortran) on such computers as IBM 360, EC 1040 (Robotron)and others (by means of a procedure given at our disposal by B. Thouin who together with R.Kittredge introduced us to the intricacies of their systems). It appeared that Q-systems are a means flexible enough to be used for our purposes, in spite of the fact that several major differences can be found between the original goals Q-systems were designed for and between our goals: after a couple of years of experience our programmers (first of all Z.Kirschner and K.Oliva) are able to use Q-systems for a dependency-based analysis attempting to penetrate into the underlying structures of sentences (which is necessary also for translation between typologically different languages). The trees Q-systems were designed to operate on can be readily interpreted as standing close to our dependency trees (though instead of each of the nodes exemplified in Fig.2 it is necessary to have a whole subtree composed of several nodes, since Q-language works only with elementary node labels). Moreover, it became also clear that Q-systems are a suitable means to handle inflectional languages exhibiting complicated systems of morphemic ambiguity and synonymy,[9] as well as the so-called free word order (which is not free at all, but determined by the topic-focus articulation, esp. by communicative dynamism, in a much more straightforward way than is the case in English). It is not necessary to work with individual rules for the different permutations of the elements of a sentence, since an approach working - roughly speaking - with an elementary dependency tree for every tentative clause (a finite verb and its neighbours on both sides) is possible, including the use of list variables for the irrelevant parts of the tree.[10]

The strong combinatoric power of Q-systems, as well as its relative transparency, made it possible to formulate a procedure of analysis, which is by far not yet complete, but which accounts already for hundreds of kinds of phenomena from the syntax of Czech. These include a relatively complete analysis of the structure of noun phrases, achieved by means of checking the agreement of an adjective with its governing noun, and preferring a noun in the genitive case to be understood as an adjunct of an immediately preceding noun, whenever this is possible, while with the other oblique cases (simple and prepositional) there is a complex scale, elaborated by J.Panevová, deciding whether the given noun functions as an adjunct of this or that preceding noun or as a modifier of the verb (the indices of the given nouns, verbs and morphemic means are used to determine the specific dependency relation). The participants modifying the verb are identified with the help of lexical data concerning valency (obligatory and optional modifications and their usual morphemic forms). We mentioned already the identification of topic and focus, achieved precisely on the base of the "free" word order.

Thus it seems that a syntactico-semantic analysis of the texts of a limited polytechnical domain (we started with texts on operational amplifiers) is feasible. In other words, it is possible to obtain in an automatic way an image of the input text having the shape of s set of disambiguated underlying representations of sentences (called statements in the sequel),interconnected by means of pointers based on the lexicon and on the paradigmatic relations registered there (hyponymy, etc.).

Whenever a user´s input question is analyzed (by the same analysis procedure as the statements are), the system goes over to other procedures, which operate on the set of statements gained by the grammatico-semantic analysis.

(2) First of all, the whole set of statements is searched through (by means of the concordance we mentioned in (v) above), to identify the subset of statements possibly relevant to the given question (in the first experiments, a non-empty intersection of the two sets of autosemantic lexical units being treated as a sufficient condition for these statements). The rules of i n f e r e n c e , which are then applied to this restricted set of statements, are described (together with the procedure of identification of appropriate answers) in the short communication presented by P.Jirků and J.Hajič, who are the main authors of the respective programmes; we can limit ourselves here to a few illustrations of these two procedures. In the rules of inference such modifications of the statements are included as the deletion of an adverbial under certain conditions (e.g. from "It is possible to maintain X without employing Y" it fellows that it is possible to maintain X), or several shifts of verbal modalities, a shift of Actor and Instrument in some cases, and also a conjuction or a similar connection of two statements; e.g. "X is a device with the property Y" and "X can be applied to handle Z" are combined to yield "X is a device that has the property Y and can be applied to handle Z".

In the first experiment the inference rules are applied only during the handling of a given question. In case a procedure checking all newly analyzed statements for compatibility with the already given pieces of information is formulated at a later stage of the research, then it will also be necessary to decide which inference rules should be applied already during that procedure (i.e. independently on questions asked by the user), and which types of consequences should be included permanently in the stock of data. It will also be necessary, in further experiments, to use heuristic strategies for the choice of the inference rules to be applied at a given time point. The growth of the enriched set of statements must be controlled and limited.

(3) The enriched set of relevant statements is then searched through by means of a procedure of the choice of an a n s w e r . The representation of the question is compared with the statements belonging to the enriched set, with three kinds of possible results:

(a) the statement is found to give a full answer to the question, if the two representations differ only in that the answer includes specific lexical units (perhaps a whole subtree) in the position occupied by the question word in the question (this position being shifted to the end of our representation of the question);

(b) the statement contains information which probably can be of interest to the user, though either some of the parts of the two representations are not identical, e.g. these representations differ in what concerns hyponymy, or in semantically relevant aspects of their word order (communicative dynamism);

(c) the statement is not relevant for the given question, if either the sequence of edges of the tree going from the root to the question word does not have a corresponding counterpart there, or if the two representations are radically different in their other parts.

In case (b) the representation of the answer is assigned the prefix "I (only) know that ..." to point out that the answer is not complete.

(4) An answer undergoes then the procedure of s y n t h e s i s, transducing the underlying representation to the graphemic shape of a Czech sentence. This procedure has been implemented in PL/1 on the computer EC 1040 and is being checked within a rather broad system of random generation of Czech sentences, which encompasses several hundreds of rules covering most different grammatical phenomena of all levels (cf. Panevová's paper presented at this conference).

The system prepared for the first experiments with the method TIBAQ is limited in several respects. An enrichment concerning the linguistic aspects (broadening of the lexicon, inclusion of yes/no questions) does not seem to be too difficult, since the grammatical patterning has already been included in the algorithms to a rather large extent. Thus the two main problems that have to be solved in adapting the system to handle open texts from a chosen branch of polytechnics or science in an appropriate way consist in

(i) the relation of instantiation (or of different objects bearing tha same lexical denomination), i.e. of the assignment of reference to definite noun phrases and other expressions has to be solved (in the texts processed in the first experiments only general concepts are present, so that up to now this step was not necessary); at least three kinds of means should be used here, namely the degrees of salience of the images of individual objects in the stock of knowledge shared by the speaker and the hearer (see Hajičová and Vrbová, this volume, about preparatory studies in this direction), further an evaluation of the known tendency to keep the topic of an utterance in its function also in the next utterance of a connected text, and, thirdly, rules concerning the role of factual konwledge in the determination of reference; this last point, which goes beyond the linguistic structuring, is probably restricted to a rather narrow domain in well-formulated technical texts;

(ii) an enlargement of the rules of inference (from about thirty that were already formulated to hundreds of them); it may be necessary to add rules of new shapes and to have a procedure for checking what effect a specific rule of inference will have in connection with the individual lexical and grammatical phenomena; in this respect only the first steps have been done in the empirical research, so that when enriching the lexicon we may face new problems of checking all the already formulated rules of inference. Only when more experience in these new fields in gained will it be possible to formulate regular patterns and general procedures which could be adequate for these new areas of artificial intelligence. Such an inquiry certainly belongs to most promising directions leading to a deeper insight into the relationships between communication and cognition.

FOOTNOTES:

1 We do not have in mind here the systems including only an elementary or marginal linguistic equipment, though some of them can well serve the purposes of text information retrieval (cf. e.g. the method MOSAIC, intended for automatic indexing and for extracting,

prepared by Z. Kirschner in the Prague group), or of natural language front-end contact with data bases. Systems belonging to the domain of artificial intelligence and serving for man-machine communication in natural language need a much more complete linguistic elaboration. This concerns the systems intended for open sets of instructions for a robot and for a dialogue with it (Winograd's SHRDLU), with which the robot's reactions can serve as a criterion for checking whether the input was "understood" by the system. The investigations of KRL by Bobrow, Winograd, Kay and others, the task oriented dialogue system prepared at SRI (Robinson, Hendrix, Hobbs, Grosz and others), as well as e.g. the models of dialogues constructed by the group of W. von Hahn in Hamburg may be classed with the systems of natural language understanding. However, these systems (and also those designed to analyze or generate narrative and other texts on the base of scripts, scenarios and similarly) differ in the level and completeness of the linguistic approaches underlying them. As for machine translation, it is interesting that most of the linguistically well equipped systems (those of Vauquois and his group, of Kulagina, of Apresjan and of T.A.U.M.) concern French.

2 The first characterization of a question-answering system of this kind was presented at the 6th International conference on computational linguistics, Ottawa; see Hajičová (1976).

3 A preliminary characteristics of this procedure can be found in Panevová and Sgall(1979); as for an account including illustrations of its technical aspects, see Panevová and Oliva (in press).

4 See Sgall et al. (1969); Hajičová and Sgall (1980a); Sgall, Hajičová and Panevová (in prep.).

5 Sgall, Hajičová and Procházka (1977); Sgall (1980).

6 For a short empirical and formal characterization of this interplay see Hajičová and Sgall (1980b); more details are given in Sgall, Hajičová and Panevová (in prep.).

7 In English the situation is more difficult, since even in printed texts it is usual here that the intonation pattern of a sentence is marked, esp. with adverbials of time and place following the intonation centre (which cannot be readily recognized by an automatic analysis of the written sentence); these adverbials in such a position belong to the topic: We came to PRAGUE yesterday differs from We came to Prague YESTERDAY; see Hajičová and Sgall (1975;1980b), where some "rules of thumb" for the identification of topic and focus in such sentences were given.

8 For our example in Fig. 2 with the lexical unit device there are pointers to such subordinated units as operational amplifier, filter, bandpass filter, stopband filter, etc.; all these units are assigned the semantic feature "device"; apply has a pointer to its synonym use; design has a semantic feature of an action noun and a pointer to its synonym project, while system is assigned a semantic feature of "intellectual category".

9 The morphemic analysis of Czech was implemented in PL/1 in the seventies, see Králíková, Weisheitelová and Sgall (1982).

10 Cf. Panevová and Oliva (1982); a German translation of Colmerauer's definition of Q-systems will appear in Prague Bull. of Mathematical Linguistics 38, 1982.

REFERENCES:

[1] Hajičová, E., Question and answer in linguistics and in man-machine communication, Statistical Methods in Linguistics (SMIL) (1976), No. 1, 36-46.

[2] Hajičová, E. and Sgall, P., Topic and focus in transformational grammar, Papers in Linguistics 8 (1975) 3-58.

[3] Hajičová, E. and Sgall, P., Linguistic meaning and knowledge representation in automatic understanding of natural language, in: COLING 80 - Proceedings of the 8th Int. Conference on Computational Linguistics (Tokio) 67-75; reprinted in Prague Bull. of Mathematical Linguistics 34 (1980a) 5-21.

[4] Hajičová, E. and Sgall, P., Dependency-based specification of topic and focus , Statistical Methods in Linguistics (SMIL) (1980b), No.1-2, 93-140.

[5] Králíková, K., Weisheitelová, J. and Sgall, P., Automatic morphemic analysis of Czech, Explizite Beschreibung der Sprache und automatische Textbearbeitung VII (Prague, 1982).

[6] Panevová, J. and Sgall, P., Towards an automatic parser for Czech, Int. Review of Slavic Linguistics 4 (1979) 433-445.

[7] Panevová, J. and Oliva, K., On the use of Q-language for syntactic analysis of Czech, in: Explizite Beschreibung der Sprache und automatische Textbearbeitung VIII (Prague, 1982).

[8] Sgall,P., Towards a pragmatically based theory of meaning, in: Searle, J.R., Kiefer, F. and Bierwisch, M. (eds.), Speech act theory and pragmatics (D. Reidel, Dordrecht, 1980, 233-246).

[9] Sgall,P., Hajičová,E. and Panevová, J., The meaning of the sentence in its semantic and pragmatic aspects (Academia, Prague, in prep.).

[10] Sgall, P., Hajičová, E. and Procházka, O., On the role of linguistic semantics, Theoretical linguistics 4 (1977) 31-59.

[11] Sgall, P., Nebeský, L., Goralčíková, A. and Hajičová, E., A functional approach to syntax in the generative description of language (American Elsevier, New York, 1969).

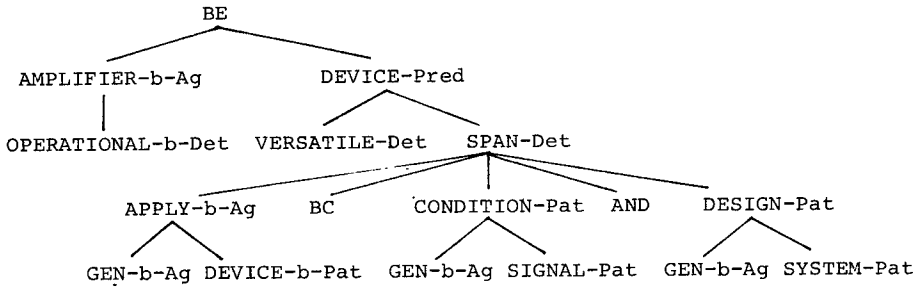

Fig. 2: A TR of "Operational amplifier is a versatile device with applications spanning signal conditioning and systems design"