Linguistic Information in the Databases as a Basis for Linguistic Parsing Algorithms.

Apollonskaya Tatiana A.

Beliaeva Larissa N.

Piotrowski Raimund G.

Applied Linguistics Department

Herzen Pedagogical Institute

Moika emb. 48

191186 Leningrad USSR

The focus of this paper is investigation of linguistic data base design in conjugation with parsing algorithms. The structure of linguistic data base in natural language processing systems, the structure of lexicon items and the structure and the volume of linguistic information in automatic dictionary is the base for linguistic parsing organization.

The avalanche-like flow of documents in natural Languages (NL) calls for a reliable cybernetic means to conduct its intellectual processing and formalized catalogization and classification.The most effective instrument helping to achieve these tasks is Linguistic Automaton (LA). LA is an all-round complex of hard-, soft-, lingua-, and partly tutorware.

During recent years, the linguistic research activity at Leningrad Speech Statistics Group (SpStGr) on natural language processing was concentrated on the pursuit of two objectives:

first, the lexico-semantical,morphological and pragmatical problems of automatical dictionary (AD)

and second, the construction of parsing programs.

At the same time, it had long been asserted that semantic and pragmatic information contained in AD and in LDB must be used to resolve many of the lexical and grammatical ambiguities that occur in the text. The adequate resolution of ambiguities is often critical to the MT process,since often ambiguities which occur in source language cannot be maintained in target language.

The creation of such a complex needs, on one hand, extensive theoretical investigations in the field of systemic linguistics and consideration of possible practical contributions in such diverse natural language processing (NLP) areas as machine translation information retrieval,indexing,automatic abstracting etc.On the other hand, all these systems need special parsing algorithms and special structure of automatic dictionary (AD).

The conjugation of AD structure and parsing hierarchy is the focus of this paper. This conjugation is hindered by a series of antinomies, the principal of which are two paradoxes:

1.The linearization paradox consists of non-additivity of text undestanding while human text processing.The process of text undestanding is sumultaneous with text reception. When modelling this process on computer, mental sumultaneous-associative processes are successively linearized during parsing.

2. The static and dynamic paradox consists of the necessity to model the dynamically and constantly enriching process of text generation and reception during the human intellectual activity with the help of previously fixed procedures on the basis of a static model of averaged professional competence, stated in LDB.

As a matter of fact, the creation of NLP system is a process of gradual overcoming these paradoxes. The success of such a process is determined by:

- the correctness of the elaborated models of professional competence;
- the database organization model and the professional competence model level;
- the level of the model of language competence, and correspondingly,
- the level of linguistic algorithms and program elaboration;
- the optimum of parsing realization;
- the level of computer development.

Thus, when designing NLP system it is necessary to conjugate the three previously established models in a united technological structure which allows to minimize the influence of the described paradoxes on the NLP rezult.

The basis of this conjugation is both the organisation of data processing (parsing) and the organisation in LDB.

The LDB organization must answer to the next requirements:

1) the data, which are inserted in the LDB, and the data descriptions must be structurized in accordance with the procedures, which are realized in a specific NLP system;
2) the LDB must be organized optimally concerning the problems, which the specific NLP system is tuned on.

The optimum LDB organization requires a modular design which consists in realization of LDB as a set of nonrigidly-linked modules. This modularity allows to arrange a LDB as modules are ready and eliminates the data duplication. Besides, it allows the step-by-step solving of NLP problems.

Besides, we must orient the structure of the system as it is and the structure of the linguware on system pragmatics which demands to investigate

- the specialist needs (express-information, signal translation, high-quality post-edited translation);
- the details of information flow (document types, volume of document and document flow, source language types, possibility of pre-, inter- and post-editing)
- the peculiarities of terminology and syntax of a special domain.

The organisation of LNP system implies the systemic principle, that determines conditions of

- the description of lexicon and morphology of source and target languages,
- the description of source and target languages syntaxis;
- the interface between LDB and software.

In accordance with this we can establish the main principles of MT system design. They are as follows:

1. The principle of modular and hierarchical organisation.
2. The principle of separation of basic and problem-oriented modules of lingua- and software.
3. The principle of the transfer as the translation process basis.

The main feature of our LA design approach is a tend to separate the groups of interconnected processes in a complicated ATP process as a whole. This separation is to be done so that their interaction both give certain system stability for different input data and allow to preserve open modular structure.

At the same time these principal points in NLP system development inevitably lead to dimention crisis. That's why in the eleborated system the hierarchy of translation levels is clearly defined. The development of the hierarchy structure of the system is realized in a descending line, "from top to bottom". This point of view implies the following:

- the exact analysis levels definition and the levels hierarchy ascertainment;
- the volume and goals definition, that means the definition of the goal of each analysis level from above, the definition of information volume of a word entry and of information distribution in word areas;
- the availability of an open modular system.

In accordance with this the procedure of translation is devided into subprocesses (levels) each having its own functional value. The results of development of each level form the basis for processing on a higher level. Thus a phrase level, a sentence level, a functional component level, a functional unit level, a lexical unit level are separated. Each level is connected with the translation process. Translation is regarded here as a multi-level process, each of its procedures translates a component of the special level.

It means that the source structures of each level are transformed into output structures which may be modified on a higher level in accordance whith the structural features of this higher level.

Thus the translation process is simulated in the system in question as a composition of lexical and semantic-syntactic translation process.During the lexical translation process the identification of text and dictionary units and the extraction of dictionary information from the lexicon blocks are carried out.During semantic-syntactic process the interlanguage structure transfer which uses the whole information received on the lexical translation phase and joins up grammar and semantic LDB blocks is carried out.This transfer process is simulated as an aggregate of verticaly conjugated subsystems, the hierarchy of the components which are extracted from the text.

The Linguist's aim in this conception of translation process is to define all the levels of translation and analysis, to formulate the set of characteristics which are necessary for the source structure modification into the target structure of the definite level and to definite the specification of the next higher levels.

Proceeding from the stated idea of the NLP system design let's analyse the structure of AD and the reciprocal correlation of grammar and dictionary on each of the determined levels in the analysis and translation of the predicate of the sentence.

During the verb entry elaboration it is necessary to choose the most important, key structural elements (which determines the dictionary volume), and to state a set of rules for the singled out linguistic elements functioning (which determines the grammar volume and the principles of parsing).

For a multilanguage ATP system the choice of AD item is determined both by word- and formbuilding principles different in specific languages as well as by the representation features of semantic text items. Besides that the choice of a basic dictionary item is determined by the tasks of NLP system and the LDB universality level.

In the Soviet NLP systems the Russian language is used as a metalanguage for source text definition as well as the target language. The unity of the target language enables to unify its definition for all NLP systems from foreign languages into Russian and to unify the procedures of morphological synthesis of a Russian wordforms.

When we design MT system for translation from the Russian the procedures of the morphological analysis are unified as well. In any case machine morphology definition of the Russian language constitutes a separate module and is used in all versions of the system.

SILOD-MULTIS AD includes source word dictionaries, which are organized as dictionaries of word usages and dictionaries of stems, source phrase dictionaries,target stems definitions and machine morphology for different languages.

Any AD that characterizes a specific language includes a universal structure set of dictionary items and machine morphology. All the source language ADs have the same function and a united scheme organisation.

This scheme allows to unify such procedures of the source language text processing as a selection of minimum text units, the morphological analysis, the identification of the text with AD items, the organization of the dictionary information file.
Any lexical unit (LU) in AD acquires a description on the morphological, syntactic, semantic and functional levels as an appropriate characteristic set.

The basic version of the system includes dictionary items (DI), which consist of the following characteristics:

-the head LU as it is: a stem, a word or a phrase;

-the lexical and syntactic code (LSC), which depends on the typological features of the source language, its grammar and parsing algorithms which are realized in the system in question;
-the translation, which is stored as references to the corresponding target language items (stems and lexical and grammatical characteristics).

For analytical languages the most expedient is the introduction of separate word forms, as it allows to increase the speed of the system while the growth of the dictionary volume is negligible. For synthetical languages machine stems are the head LU in the DI and the input AD is filled up with machine morphology.

In order to reduce the memory volume for AD location we resort to the artificial morphology transformation, i.e. to the insertion of the agglutinative morphology. The essence of the latter consists in the process of the selection in any word usage a machine stem and an affix "sticking"to it.

The concept of the inserting of machine affix allows to elaborate the Russian machine grammar, formed as a set of paradigms - machine affix chains. Each typical paradigm correlates with the grammatical characteristics of stems and the word formation mode. The link between a machine stem and a paradigm is realized with the help of a special code, which characterizes all the word forms which can be generated from the stem in question.

The use of this machine morphology allows to realize the wordform generation proceduress in accordance with the lexical and grammatical characteristics which are formed in the course of MT, and to make this procedure a universal one for any language pair.

Accordingly, the elaborated Russian stem dictionary permits to identify automatically the text words with dictionary items and to ascribe their morphological characteristics accurately to case homonyms. The result of morphological analysis, which is received with the help of LDB and special lexical and morphological analysis algorithms, is a source for parsing and transferring algorithms for Russian-English MT.

A two-layer system of lexical and semantic coding is realized in the LDB of SILOD-MULTIS system. The upper level of this coding is constituted by 30-element LSC which is formed in DI immediately. LSC formation is created in accordance with the coding tables elaborated for every source system languages. This information can be formed on-line.

The levels discussed above specify the lexical and grammatical description of LU in LDB. The syntactic definition covers the functional LU characteristics which determine their potential capacities to accomplish a specific role in syntactical sentence structure. The semantic definition which constitutes in a distinct, internal level is concerned with the transfer from the linguistic phenomena proper to the extralinguistic ones. The formation of this definition is based on the structural investigation of the domain, that is to be manifested.

Let's consider the structure of information on the example of verb entry of French-Russian MT system, which is the base for parsing system.

On the lexical level of the analysis the predicate equal to the morphological verb-form is development. In the French-Russian MT system the verb is presented in two ways: as word-forms for the irregular and suppletive verbs (avoir, etre, aller, vouloir) and as machine-stems with their standard paradigm.

Each source standard paradigm includes information sufficient to establish a link with a definite stem and a corresponding word entry (item).The analysis procedure is performed according to the

morphological tree.

On the functional unit level a verb and nominal segments are identified. The structures of this level include verb segments equal to the complex verb, tense of the pronominal verbs and of the verbs in active and passive forms. The procedure is performed on the information contained in various positions of theverb entry:

- the information of the verbs belonging to the auxiliary class are contained in the LSC. This information is necessary for the discrimination of the complex verb tenses. Position Six of the LSC of the verbs "aller", "venir" contains the information necessary for "Immediate" tenses identification;
- the passive form identification footholds on Position Eight (transitivity notes), but for its translation the corresponding rules of Position Eleven are to be used. This position contains the information of the possibility of the shortened passive participle form usage ("est ouvert" - opened), the pronounal form usage ("est prepare" -- is prepaired), the active form usage ("est suivi" - follows).

The pronounal form is translated according to the information of Position Twelve of the verb entry. The compound nominal predicate identification and translation is performed on the basis of Position Fourteen.

As to the designing of the grammar rules which direct the analysis and translation of impersonal construction it is prescribed by the information of Position Fifteen.

The inner verb class relations are of fixed character. This makes it possible to présent a verb segment as a frame including all verb-connected elements (the objective pronouns, the negative and limiting partycles) and verb elements (the auxiliary verbs and the participles of a conjugated verb) . During the analysis on the functional segment level the procedure of homonymy elimination is realized.

The result of the procedure on this level is a chain of source and target functional segment. Together with this the target functional segment (a verb group) gets a certain set of indications necessary for the next level - the sentence level analysis.

The peculiarity of verb elements analysis is their immediate functioning on the sentence level, as to the nominal groups, they have an additional stage - the stage of functioning components formation. This is explained by the diversity in the interrelations of the nominal group elements.

Thus up to the begining of the sentence level analysis the structure of the verb functional segment is known, the ways of the given verb structure presentation are defined; the verb elements homonymy is eliminated. The designed output structure gets the total set of indications necessary for its analysis on the sentence level.

This set is compiled of the active form verb entry information:

- the indication of the obligatory direct object according to Position Eight;
- the indication of the possible information distribution according to Position Six;
- the indication of the possible object or adverbial modifier according to Position Nine.

This set is also compiled of the information ascribed to the pronominal verbs (the type of government), according to Position Thirteen, and to the passive form verbs according to Position Ten; and the indications formed in the translation process on the preceding levels of the analysis (tense, number, person and others) of the compound verb constructions in all mentioned forms.

By the sentence level analysis stage a number of "refusals",

got on the previous levels, are piled because of various causes (ambiguity of the structure in a bilingual situation, uneliminated homonymy, impossibility of the analysis on the preceding stages of a number of constructions (infinitive, passive, impersonal, pronominal) requiring the subject-object transfoormations for a correct translation). Thus it is possible to pass over to the choice of the translational structure of the whole sentence only after the functional of the nominal and verb groups as sentence members is defined.

While choosing the translational equivalent on the sentence level some difficulties arise in the case of the input and output structures inadequacy.

Then it is possible to resort to the subject-object transformations. The subject-object transformations may be realized either with the help of the sentence members rearrangement or by the case forms of the target structure change or by the conversives search.

The conversives search practically leads to the increase of the number of the verb translational equivalents. More productive is the way of subject-object transformations, connected not with the sentence members rearrangement but with the case relations change in the output structure. The results of the sentence level elaboration is the obtaining of the output sentence structure.

On the phrase level the translation of the whole complex sentence is performed. Here the subordinate clause translation is corrected. In particular the testing of the correct choice of the conjunctions and relative pronouns, introducing the subordinate clauses. Thus for a correct choice of the translational equivalent of an homonymous form "que" (what, so that, which) it is necessary to resort to Position Eight of the word entry information. The information contained in it gives an opportunity to choose the corect form (indicative or subjunctive) for the subordinate clause verb translation. The same process takes place when translating the subordinate clause with "clout". The correct choice of the translational equivalentfor the whole subordinate clause is realized only with the orientation to te indication of Position Nine of the main clause verb.

Thus the chosen point of view on the MT system elaboration makes it possible to realize the whole volume of the research goals. In this circumstance that is an indipensable facility for the designing of the interaction of grammar and dictionary on each of the system levels.

Hence this conception creates the necessary faciolities for the development of the systems forecasting the analysis of newly arising situations on the basis of the once elaborated situations.

W.J.Hutchins. Machine Translation: Past,Present,Future. Chichester: Ellis Horwood , 1986;382p.

Ju.Kondratieva, R.Piotrowski, S.Sokolova. Organization of the Russian-English MT-algorithm. SCCAC Newsletter, No 4, Bowling Green, 1988, p.21.

N.Maruyama, M.Morohashi, S.Umeda, E.Sumita. A Japanese sentence analyzer. -In: IBM Journal of Research and Development, Vol.32, No 2, March 1988, pp.238-250.

S.Nirenburg(ed). Machine Traslation: Theoretical and Methodological Issues. Cambridge: Cambridge University Press, 1987. -350p.

SILOD. A Russian-English Translation Support. New-Delhi: Elorg-Computronics India, 1988. -16p.