METHODS FOR TRANSLATION MODELING IN THE AUTOMATIC TEXT PROCESSING SYSTEMS

Zarifa Guliyeva

Institute of Information Technology of ANAS, Baku, Azerbaijan
depart15@iit.ab.az

In the making of the "information society" the raising of the management efficiency in the field of the scientific research and development can be achieved by automation of the information processing and implementation of Information Search Systems (ISS) and Systems for Automatic Text Processing (ATP), which currently are used in various automated systems for natural language text processing, in economics.

The ATP technology appeared in late 70s of the 20th century when the main research issues were such problems as solving ambiguity, generative models of language, issues of formal grammars and their features, etc. The machine translation systems in the form of expert systems and the information search systems belong to the intellectual systems that process text in natural languages. They are the components of the ATP technology that are designed for the natural language interaction with user. The researchers in the automatic text procession (ATP) orderly move from the simplest analytical methods to more advanced ones and gradually approach to such text presentation that corresponds to human perception. On one hand the research in the field of logic-mathematical and high level software product creation has commenced that would allow presenting the machine procedures and the information used by them in the format that is more transparent, compact and suitable for human perception so as the developer would be able to concentrate on the content of the problem that is solved without being diverted to the technical problems of his solution record. On the other the formal models of natural language and language activity have commenced to be developed in which facts and events are comprehended in terms of there prospective machine processing and presented in the form suitable for the purpose of such a processing.

But complete imitation of the human language activity is not feasible even on modern computers. Any language algorithm can only describe some big or small language processes. The models can be full or partial ones. The partial models include single morphological events of a particular language. They do not take into account the errors in the entrance presentation therefore assembling the partial models in one full system which imitates all the language mechanisms (from morphology to semantics) will want special effort. The full models to which belong all big systems for machine translation and full text analysis are generally created by a team of linguists which after a long period of collective effort begin to represent a new scientific approach in the applied linguistics where currently growing attention is paid to problems of realizing natural language (NL) and ATP.

1. Classification of the translation models. Translation model is a conventional description of a number of operations by executing which the program or machine can complete translation of the whole matter or its portion. The task of the translation model is just to describe the actions consequence by which the translation task can be fulfilled at the specified terms of translation.

The translation model is of conventional character as it does not necessarily reflect the real activity of translator when creating the text of translation. Most of such models have limited explaining power and do not pretend of being the basis for translating any text with required degree of adequacy. The translation models disclose some aspects of the translation linguistic mechanism functioning. In his practical work translator can reach the required result by a method which does not coincide with any of known translation models but knowledge of such models can help him with solving difficult translation problems. At the current stage of the translation modeling evolution the following model types can be noted:

1) Situation model;

2) Transformation semantic model;
3) Psycho linguistic model;
4) Translation adequacy model.

The Situation model considers the translation process as the process of description by the translation language of the situation that was described by the language of the original text. When perceiving the original text the translator compares the component units of the text with language units of the source language which are known to him and by interpreting their meaning in the context determines what a situation of the reality the original text describes. The situation model works most correctly when dealing with lexical units having no equivalents. Even if the created adequacies (by transcribing or calquing) refer directly to the translated word of phrase they can only be chosen on the basis of correct comprehending the situation and when the described situation determines the choice of the translation variant. That is there is only one way of describing the situation in the translation language regardless of how it is described in the original text.

Transformation - semantic model differs from the situative one and based on the suggestion that in the process of translation transfer of the meaning of the original text unit is executed. It considers the translation process as a series of transformations by which the text is converted from the units of the original language into the units of the translation units and sets relations of equivalency between them. The method of the translation presentation that is used in the transformation-semantic model very much corresponds to the translator's intuition who is very often eager to find how to transfer a sema in the meaning of the original text. The transformation-semantic model does not simulate the translation process when figurative and other text related associations should be transferred.

The first two translation models present its conventional depiction and do not fully correspond to the real translator's activity. To more fully reflect the activity of the translator the model should include description of psychic processes underlying such an activity.

The psycho-linguistic model of translation is developed on the basis of the statements of the speech activity theory. According to those first an inner program of the further communication is formed which then is transformed in speech statement. When executing the translation process according the psycho-linguistic model the translator first transforms his conception of the original text content and then transforms the program in the translation text. This process can be identified with translation from the source language to the inner code and then translation from the code to the translation language.

However the model does not fully correspond to the ideal translation model as we do not know how such coding is executed and how one of the possible ways of such a program realization is chosen in the translation text.

The model of the translation correspondences has some advantages as compared with the three ones presented above. Translation, being a special process of inter-language transformation deals with various language levels: morphology, vocabulary, syntax and semantics. In the translation process complex interaction of these levels takes place resulting in appearance of new units: translation units (TU), translation correspondences (TC), etc. These units belong to different language levels. This translation model reflects the hierarchy of these language levels. The novelty in the modeling that uses translation correspondences is that the latter are placed in the center of the model and the whole process of modeling [3]. To obtain integral model of MT the sequence of actions with the formalized original information should be united in the consistent system not only for translation of phrases and sentences but also of the complex of the sentences – integral text.

2. Realization of the translation model. The adequacy and optimality of the system work depends on the proper comprehending and formalized description of the human translation activity as an organized sequence of operations and procedures that could find its analog in the TM systems. To realize the linguistic model of the information processing in a bilingual situation that consists of a few subsequent stages it is necessary:

1. Realization of the following techniques:

a) lexical TM – includes creation of acting bilingual dictionaries of the icon type;

b) grammatical TM – includes creation and machine realization of acting algorithms of the morphological and syntactical analysis and synthesis;

c) semantic TM – includes creation and machine realization of acting algorithms for eliminating the lexical ambiguity.

2. Association of the created techniques in integral system and its complex realization.

3. Further permanent modification of the used techniques and acting engineering-linguistic models and development of more perfect systems on the basis of modern achievements of science and technology.

Lexical TM is based on the methods of creation of automatic dictionary for the system of the machine translation which is of practical and theoretical significance for language researchers in the fields of general linguistics, comparativistics, applied linguistics and artificial intellect. In the translation system the AD is the main means for keeping language information. It is directly connected with all levels of language hierarchy and participates in all stages of the translation process. The data contained in the AD are used for morphological, syntactical and semantic analysis. However, being a component of the system the AD interacts with the knowledge base without which the data of the dictionary will not be of special significance in the SMT.

Grammatical analysis. The first step in the processing the natural language text is the block recognizing (analyzing) and creating (synthesizing) various forms of words – the block of the morphological analysis and synthesis. Morphological processing of the text includes solving of all or some of the following problems. The morphological analysis is the identification of the original word – lexema by the wordform and also morphological characteristics of the wordform such as gender, case, number, plurality or singularity, etc. The morphological synthesis is the process of the converse comparison – finding the form of a required word by the defined case, plurality or singularity, etc.

In the process of the morphological analysis two widespread approaches can be observed which can be conventionally called "right to left" and "left to right" analysis.

In the first case the word is preliminary isolated from the context. Then, without using dictionary finite set of affixes is cut from it and thus the word base is accentuated. Then the attempt to find the translation correspondence for the base in the base dictionary is done. This method is perhaps the most widespread one. More widely to this type any methods can be attributed that use reduction of the word without dictionary to the form allowing search of the chain in the dictionary.

In the second case the implied base of the analyzed word is searched in the base dictionary and then based on the dictionary information obtained an attempt is done to interpret the residual right part of the word as the set of affixes.

The third possible approach is description of the language in the form of rules and algorithms. In this case the "left to right" analysis becomes impossible as to find hypothetic bases you should cut endings and apply alteration before referring to the dictionary. The advantages of this method are natural and linguistic description of the considered events. The weakness of the method is the branched hierarchy of rules but the current tools of the information technologies bring wide opportunities for implementation of even very algorythms. The two methods mentioned above can be combined. However the language description in the form of rules is more natural for systems rendering synthesis rather than analysis of wordforms. Such a system was called the two-level model [1].The system is a language independent one for a wide class of languages of various structures. The morphological analyzer and synthesizer working by the finite automate principle.

The two-level model is a kind of formal grammar. The grammar and dictionary determine the linguistic model by forming the main portion of linguistic data. Separation of the grammars and algorythms is important for practice as it allows changing grammar rules without changing algorythms (and the programs thevselves) dealing with grammars. But such separation is often impossible. The most clear separation of grammars and algorythms can be observed in systems dealing with context-free (CF) grammars where the language model is a grammar with finite number of states and the algorythm should provide for any sentence the tree of its output by the grammare rules and if there are a few such outputs it should enumerate those. Such an algorythm is called the analyzer. Analyzers are created for grammar classes though taking into account of the grammar specific features can increase the analyzer's efficiency. There is wide choice of analyzer creation techniques for the context-free grammars (CFG). However, just CFG themselves are not sufficient for description of a natural language. Therefore they were added with CFG sets of *transformation rules* that deal with trees of components. There are many of recording methods for context conditions and all of them are expansions of CF rules. Putting a number of limitations on the kind of the used transformation rules allows program realizing of the rule interpretor in the form of a finite automate that results in increazing the calculation efficiency of the analyzer. The two-level presentation only that corresponds to the original text and the analyzis result is used there unlike multi-level models. The sets of rules are considered as unordered sets and do not used alternately but simultaneously. The latter point makes the text transformation rule set simmetrical relatively the both levels that is relatively the wordform analysis and synthesis operations. Thus both the morphological analyzer and the program for synthesis are created on the basis of the same set of linguistic rules and the same program interpreting these rules.

An alternative approach that is currently developed abroad is using the models of so called paradigmatic morphology. In terms of this approach attention is paid mainly to determination of the hierarchic system of morphological classes and description of the mechanisms of inheritance by subclasses. Such an approach is suitable for description of highly flective languages e.g., Latin and some Slavonic languages.

3. Morpho-syntactic analyzer in MTS. Currently, we are developing a linguistic analyzer on the basis of the system for machine translation from English to Azerbaijani. The analyzer renders full morphological, superficial syntactic and partial semantic analysis and the text synthesis deriving from the specified program realization. This analyzer consists of the data base that is presented by the automatic dictionary and lists of exemptions attached to it and also of knowledge base presented by the set of rules for recognition and generation of transformation grammatical forms of the input language (English) with subsequent getting to formation of equivalent forms of the output language (Azerbaijani). Saying figuratively, these rules consist of right and left sides. The left side contains formal description of the analysis of a grammatical form of the input language and the right one describes the means for formation of the identical form of translation correspondence in the output language.

In this development of the machine translation system the two-level model can be applied but we should take into account that the text morpho-syntactic analysis is not an independent task. Rather it is a subtask of a particular applied task, in this case a subtask of the ATP system and is rendered with specific program means with their characteristic algorithms, etc. Therefore, this model should be applied combined with the "left to right" and "right to left" approaches in accordance with the terms of formal description of the working language pair. That is the combined approach for the formal language description should be applied.

At the current stage of the development of the English-Azerbaijani analyzer the analysis of an English sentence is executed first on the basis of recognizing grammatical wordforms within phrases (nominal, verbal, prepositional, phraseological ones, etc.) that is by morpho-syntactic analysis. When such an analysis is impossible the word-for-word translation of the sentence is rendered.

The most practical sequence of procedures for the morpho-syntactic analysis for translation of an English sentence to Azerbaijani is the following one:

1. Analysis of the sentence left to right according to structure of the English sentence.

2. Dividing the text into phrases. Dividing is executed a priori without using lexical information. If the grammar form of the phrase components coincide with algorithmic description of this language category then the synthesis of the translation correspondences of these forms in the output language is executed based on the generative grammar rules.
3. If there is no coincidence with any algorithm description the text is re-divided in single wordforms and their analysis is executed.
4. Matching with the dictionary. If the word is found in the dictionary the analysis is completed.
5. If the word is not found in the dictionary "right to left" analysis is conducted and by recognition and generating rules proximate segregation of the wordform bases is executed. One word can be corresponded to a few hypothetical bases. During the subsequent analysis each of such hypotheses is checked. As a rule one hypothetical base only corresponds to a word.
6. If the base of the word is found in the dictionary then by the lexical and grammatical information attibuted to it and according to the generation rules it is checked if a particular ending can be attached to that base. Then grammatical characterics of the wordform are determined and by this the analysis of the grammatical form is completed and there is the turn of the synthesis algorythms.

References

1. Gelbuh A.F. Efficiently realized model for the morphology of the flective natural languauage. Dissertation. UDK 801.73:681.3 (043.3) Vserossiyskiy Institut Nauchnoy I Tehnicheskoy Informatsii (In Russian).
2. Marchuk Yu.N., Tihomirova B.D., Scherbinin V.I. System for machine translation from English to Russian// Machine translation and automation of information process. M., 1975. (In Russian)
3. Melchuk I.A. Experience of the *Meaning* <=> *text* linguistic models. M.: Nauka, 1974. (In Russian).