Development of a Linguistic Automaton on the Basis of Statistics of Speech

R. G. PIOTROWSKI
St-Peterburg Pedagogical Gertsen University, St-Peterburg, Russia

YURI TAMBOVTSEV
Lviv Lesotechnical University, Ukraine

Abstract

The methodology and developmental technology were created for a multifunctional modular linguistic automaton which functions as a computer analogue of human verbal and mental process. A linguistic automaton (LA) is a complex system described through a multi-aspectal representation in the form of various models and diagrams. It includes hardware, software, the so-called 'lingware' and other components. Two description strategies are presented: structural/functional and management/decision.

Structural/Functional Description of a Linguistic Automaton (LA)

This type of description disregards the LA physical substrate. It represents the automaton as a hierarchical system comprising the following three planes (layers) (Zubov, 1985; Chizhakovskij, 1988; SILOD, 1988; Piotrowski, 1990).

(1) The upper layer, implemented for an LA in the form of man–machine interaction, is viewed as an analogue of motivation and partly as a communicative-pragmatic operator in the representation of human verbal/mental activity.

(2) The middle layer, represented as a set of automatic processing (ATP) subsystems:

$$F = (U,Y,F,I,A,E,P,P_Y),$$

where U is the subsystem ordering text units (letters, letter combinations, word forms, and word combinations) alphabetically, according to frequency, alphabetically by last letters, etc.; Y is the subsystem which determines the language of the text; F is the fragmentation subsystem; I is the subsystem for indexing texts and text fragments; A is the subsystem that constructs the document search pattern and annotation/abstracts; E is the expert subsystem operating in a man–machine dialogue mode; P is the machine translation subsystem; P_Y is the subsystem for thematic machine translation of article and book titles. The number of functional subsystems could be increased. Experimental text generation modules (including verse generation) are constructed which produce a text proceeding from a specified semantic 'motive'. This idea is also used in generation of output texts as an element of machine translation (Mel'chuk and Ravich, 1967; Piotrovski, 1975, 1990; Zubov, 1985; SILOD, 1988; Piotrovskaja, 1991; Popeskul, 1991). Attempts have been made to use a linguistic apparatus for language learning. Providing a system with linguodidactic programming (CALLware), we convert it to a linguistic teaching automaton (LTA) (Piotrowski, 1975, 1990; Prekup, 1989).

(3) The lower layer, described by set of functional modules M, consists of two subsets, M_a and M_s. The former combines analysis modules

$$M = \{d,c,l_k,l,m,L_k,L,g,s_1,s_2\},$$

where d is the text decoding module; c is the text correction module; l_k is a module for lexical analysis of text keywords; l is the module for word/phrase-based (lexical) analysis of all lexical units (LU) of the text; m is the module for autonomous morphological analysis of word occurrences in the text; L_k is the module for lexical/morphological analysis of key LUs; L is the module for lexical/morphological analysis of all LUs; g is the module for analysis of superficial text structure; s_1 is the module for analysis of deep (theme/rheme) text structures; s_2 is the module for semantic/pragmatic text analysis.

The second system includes synthesizing module

$$M_s = \{k,c,l',L',g',s'_1,s'_2\},$$

where k is the module for graphic or phonetic representation (quantification) of text; c is the correction module; l' is the lexical synthesis module (selection from automatic dictionary of lexical equivalence of input word occurrences and word combinations); L' is the module for lexical/morphological synthesis of word occurrences and combinations; g' is the module for synthesis of superficial output text structure; s'_1 is the module for synthesis of the semantic/pragmatic text image.

These modules, assigned to various levels and sublevels of message generation and analysis, function as computer analogues of these levels. For comparison of various ATP systems, one can operate with the concept of the working space of an LA (Mel'chuk, 1967; Zubov, 1985; Prekup, 1989; Piotrovskaja, 1991; Popeskul, 1991). For an ideal LA this is organized as Cartesian products $= F \times M$, which includes all doublets f_i m_j (f_i is a subsystem, m_j is module). For the working space $S = S$ of each existing LA it includes only those binary relations $f\,m$ which hold for the respective LA. For example, the working space of an elementary LA which performs word-for-word machine translation (Piotrovskaja, 1991) is organized as follows:

$$S^* =, (\Pi d, \Pi l, \Pi l', \Pi k).$$

Correspondence: Professor Dr Y. A. Tambovtsev, Chairman Department of Linguistics and Foreign Languages, Lviv Lesotechnical University, 290044 Lviv-44, Pushkin Street 84, PO Box 8834, Ukraine.

The working space of a more complex LA, which performs text fragmentation prepares a search pattern and generates a rough translation of a French or British patent specification fragmented according to subject headings (Chizhakovskij, 1988; Kondrat'eva and Sokolova, 1988) and is represented by the following set (see Fig. 1):

$$S = \{\Phi d, \Phi L_k, \Phi L', \Phi k, Ad, Al_k, Al', Ak, \Pi d, \Pi L_k, \Pi L', \Pi k\}.$$

Decision Arrangement of a Linguistic Automaton

An ATP system involves recognition operations which are performed under uncertainty presented in algorithmic blocks of a linguistic database by a set of versions, from among which the LA selects the correct decision. LA architecture is described in decision-making terms as well as from the structural/functional point of view. An LA is represented as a mapping of input lexical units (input texts) T into a set of output lexical units (output texts) T performed under control of G, which is an analogue of an operator. In other words, we have

$$LA^* : T_i^{in} \times G_j^i \rightarrow T^{out}, \; i = \overline{1, e} \; . \; j = \overline{1, n}. \quad (1)$$

where l is the number of subsystems of the given LA, and n is the number of modules used in the ith subsystem (Mel'chuk and Ravich, 1967; Zubov, 1985).

By analogy with control systems, the decision process described by formula (1) can be represented as a hierarchy of the following planes (layers): (1) self-organization, (2) LA adaptation to texts processed, and (3) selection of solution method for the problem being processed.

At the first level—usually in dialogue with the computer—the strategy for solution of general problem $\overline{P}$ is elaborated. Subsystems and modules are created and assembled according to this strategy to enable the automaton to execute its task.

Proceeding to solution of a problem, the LA is normally in conditions of uncertainty due to multiple meanings of vocabulary LUs, ambiguity of morphological forms, syntactic patterns in the text, and to lack of linguistic and encyclopedic knowledge in linguistic DBs (IDBs). The decision architecture must have tools for adaptation to reduce uncertainty, which includes filter algorithms (Piotrovski, 1975) (see below) and techniques for LA adaptation to texts processed. These techniques include updating automated systems to expand the list of geographical and proper names, terminological word forms and combinations of the relevant language subset, and creation of new and restructuring of existing algorithms. This additional LA learning is accomplished in the course of man–machine dialogue or in an autonomous regime, with the machine selecting the most frequent alternatives (Muzalevskaja, 1988). All these techniques constitute the second adaptive level of LA decision organization.

Input Text

/ 19 / république francaise
/ 11 / 2. 071. 055
/ 21 / 69. 43575
/ 15 / brevet d'invention
/ 22 / 16 decembre 1969,16 h. 45 mn.

revendication 2. 071. 055. 1. mechanisme de transmission qui comporte un moteur dispose longitudinalement par rapport a l'axe du vehicule un embrayage une boete de vitesse qui comprend au moins un arbre d'entree et un arbre de sortie paralleles, un mechanisme de differentiel dispose entre le moteur et la boete de vitesse.

patent search pattern (word/phrase-based annotation):

patent claims,mechanism,transmission,engine,axle (pin), transportation vehicle,clutch,gearbox,input shaft, differential mechanism,differential

differentiation into conceptual fields (frame-questionnaire) and translation quasiabstracting:

/ 19 / country of patent:
France
/ 11 / patent number:
2. 071. 055
/ 21 / application registration number:
69. 43575
/ 15 / publication type:
patent
/ 22 / application date:
16 December 1969, 16 hr 45 min

patent claim 2. 071. 055

title of intention:
1. transmission mechanism
combination of restrictive characteristics of invention: which is comprised of an engine situated lengthwise relative to the axle (pin) of a transportation vehicle, clutch, and gearbox.

combination of restrictive characteristics of the object of invention:
which contains at least one input shaft and one output shaft parallel:
the differential mechanism is situated between (is included in) the engine gearbox

Fig. 1 Conceptual processing of a French patent text and its French patent text and its machine translation

Arrangement and operation of mechanisms of the third layer are essential to the LA concept. We discuss this aspect in some detail.

Choice of Efficient Method for Problem Solution

The Speech Statistics Group has developed several techniques to find optimal decisions which take into account engineering and linguistic limitations of LA. Some of these techniques have already been programmed for computer usage. The two methods which follow deserve special comment.

The first method is based on hierarchial organization of operation of subsystems and modules, which can also function autonomously. It is implemented through the following rules:

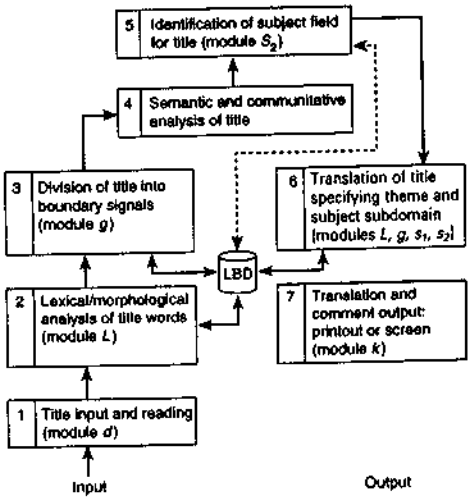

Fig. 2 Structural/functional flowchart of LA which executes theme/rheme machine translation of German titles

—the top control level is the man–machine decision-making block (see above).

—subsystems and modules of top levels control purposive operation of respective blocks of lower levels.

—if lower level subsystems and modules operating in autonomous mode are incapable of making decisions or adopt several alternatives, the results of text processing are sent to the upper hierarchical level, where the final decision is made (Botnaru, 1985; Prekup, 1989).

We illustrate this procedure by LA (Fig. 2) which performs theme/rheme translation of German article titles in the subject field 'Waste water and effluent' (Muzalevskaja, 1988). Lexical–grammatical processing of a title begins in block 2, which represents the lower level of LA. It makes use of a German–Russian dictionary and computer morphology included in LDB. It represents not only lexical/morphological blank forms for translation but also information on morphological boundary signals and semantic data utilized later for decision-making at higher levels.

In the third block, titles are divided into semantic/syntactic segments on the basis of morphological indicators produced by the second block and syntactic boundary signals extracted from LDB. A string of filters serves to determine the communicative (thematic and rhematic) character of these segments in blocks 3–5. Filters help with the selection of alternative decisions.

The first filter included in the third block is probabilistic/syntactic. Preliminary statistical studies of German titles indicate that in almost 90% of cases the first segment coincides with the theme or is included in it. The final segments of a title belong to the thematic component in almost 70% of cases. The second and third segments are assigned to the theme or the rheme with less certainty. Positional segmentation of a title does not always yield an unequivocal decision. Results produced by blocks 2 and 3 must be transmitted to a higher level (block 4), which performs further communicative analysis of the text and tests the results of segmentation performed by block 3. Block 4 includes a filter which tests all words and stems from the title against LUs kept in lists of rhematic and thematic indicators of the LDB. The results of this comparison are matched against information transmitted from the lower level (block 3). This operation leads to one of the following outcomes.

(1) Lexical units identified as rhematic indicators are found in initial segments, while thematic indicators are found in final segments. Results of analysis at both levels thus coincide, and LA makes the decision: the segment end is the theme, the initial segment is the rheme. Attributive segments adjacent and rhematic segments may be qualifiers of the theme or the rheme.

(2) Information extracted by the third block is at variance with information generated by the fourth block. In that case, theme/rheme segmentation of the title produced by the fourth (top) block is assigned priority.

(3) The fourth block does not yield an unequivocal decision concerning theme/rheme segmentation of the title. In that case, title parameters obtained in blocks 3 and 4 are transferred to the top (fifth) block. It compares the title with subfields of the subject area concerned. To this end, the system employs indices which determine that terminological word forms belong not only to the 'waste/effluent' field but also to other related subject areas discussed in texts studied (the pertinent indices are given in vocabulary entries of input terms).

Statistical analytic data indicate that terms of the 'waste water/effluent' field are more likely to occur in the thematic component of the title, while terms of the other subject in the rhematic segment. The presence in a segment of a word form belonging to a subject field can serve as an additional indicator of the rhematic or thematic function of the segment. This is illustrated by the theme/rheme analysis of a German language article from the journal *Wasserwirtschaft-Wassertechnik* (no. 1, p. 4, 1985). A fragment of the printout resulting from analysis and translation is presented in Fig. 3. Title processing by blocks 2–4 does not yield a final decision as to the identification of the theme or the rheme. All parameters are transferred to the top fifth block. After processing of term indices, the title is assigned to one of the mental space (subdomains) of the subject field studied. Terms with indices of mental spaces are treated as weak indicators of the theme. Word forms and combinations with an index of a different subject field (other than 'waste water/effluent') indicate the rheme. According to this rule, segment 'eines Auswerterrechners' is included in the rhematic part of the title. Segment 'Trinkwasseraufbereitung' has a subject index, confirming that the word form belongs to the thematic segment.

Title: Einsatz ! eines Auswerterrechners ! bei Verfahrensuntersuchungen ! in der Trinkwasseraufbereitung.

Einsatz—S,N/D/Ac,Sg,r / 941 /
Auswerterrechners—Cp, S,Cc,Sg,M / 105 /
Trinkwasseraufbereitung—Cp,S,Cc,Sg,TW / 4001 /
Einsatz eines Auswerterrechners—rheme
bei Verfahrensuntersuchungen—no solution
Auswerterrechners—'machine-equipment' enterprise
Trinkwassersaufbereitung—'water management' enterprise
Translation: application of computer in studies of drinking water treatment methods.

Symbolism
! = segment boundaries; Ac = accusative; Art = article; Cc = common case; Cp = compound word; D = dative; G = genitive; Gs = boundary signal; M = machine and equipment enterprise; N = nominative; R = rheme; S = noun; Sg = singular; TW = water management enterprise (Russian equivalent addresses given in parentheses).

Fig. 3 Theme/rheme machine translation of a German title

The communicative character of segment 'bei Verfahrensuntersuchungen' remains uncertain and must be decided in a dialogue between LA and the user.

The second method of finding an optimal decision rests on the capacity of LA for decomposition or simplifying modification of general problem P if it cannot be solved or requires a prohibitive amount of time and computer memory.

After decomposition, the general problem is presented as a set of partial problems:

$$\overline{P} = \{P_1, P_2, \ldots, P_2, \ldots, P_k\}.$$

As an illustration, we describe a problem that arose in development of an experimental Turkish–Russian machine translation system. Nominal and verbal paradigms of the Turkish and Russian languages are utterly nonisomorphic. Lexical/morphological modules L and L' have to interact with analysis and synthesis modules of superficial and deep sentence structure (g/g', s/s') to generate Russian word forms and combinations that correspond morphologically to Turkish input word occurrences. Since modules g/g' and s/s' for Turkish–Russian machine translation have not yet been completed, the lexical/morphological task of L/L' is subdivided into three independent subtasks:

P_1—analysis of Turkish word occurrence. As a result, the word is subdivided into its stem (initial form) and component affixes (compare module m);

P_2—determination of grammatical characteristics of each affix (module m);

P_3—translation of stem(modules l/l').

Proceeding from information obtained after execution of these subtasks, the user prepares the final translation of the input Turkish sentence. A typical example where general problem P is replaced by simplified modification P is switching LA to a word-by-word or phrase-by-phrase translation when the system lacks morphological and semantic/syntactic resources to build superficial and deep structures of the input sentence. Decomposition and simplification of P enhances the vitability of LA by allowing the system to escape from deadlocks that occur when the automaton fails to follow a prespecified text processing format

References

Beliaeva, L. N., Kondratjeva, Ju., Piotrowski, R. and Sokolova, S. (1989/90). In K. M. Schmidt, R. A. Boggs *et al* (eds), *Abstracts from the Leningrad MT Project: Society for Conceptual and Content Analysis by Computer*. Bowling Green State University Newsletter, No. 5, pp. 26–35.

Botnaru, R. V. (1985). *Automatic Recognition of the Meaning of a Special Text on the Basis of Relator Frames*. In Russian. Dissertation for the Degree of Candidate of Philosophy, LGU, Leningrad, pp. 29–30.

Chizhakovskij, V. A. (1988). *Semiotic and Communicative Aspects of Automatic Processing of Scientific Titles*. In Russian. Dissertation for the Degree of Doctor of Philosophy, LGU, Leningrad.

Kondrat'eva, Yu. N. and Sokolova, S. V. (1988). Organisation principles of machine translation algorithms. In: *Speech Statistics and Automatic Text Processing*. In Russian. RGPU im. A. I. Gertsena, Leningrad.

Mel'chuk, I. A. and Ravich, R. D. (1967). *Automatic Translation. 1949–1963: A Critical Bibliography*. In Russian. VINITI, IYa Akad. Nauk SSSR, Moscow.

Muzalevskaja, V. M. (1988). *Reproducing an Engineering Linguistic Model of Logical and Semantic Analysis and Translation (with Special Reference to British and US Patents in Printed Circuit Board Production Technology)*. In Russian. Dissertation for the Degree of Candidate of Philosophy, Voennyj Krasnoznam Institut, Moscow, pp. 296–298.

Piotrowskaja, K. R. (1991). Modern computerised linguodidactics. *Nauchno-Technicheskaja Informatsija*, Ser. 2, no. 4, p. 26.

Piotrowski, R. G. (1975). *Text/Machine/Man*. In Russian. Nauka, Leningrad, pp. 91–145.

—— (1990). Linguistic automation and computer-assisted language learning via microcomputer. In: *Proceedings of CALL '89: International Conference of Computer-Assisted Language Learning at the Institute of Applied Linguistics*, Wilhelm Piek University, November 15–17, 1989. University of Rostock Press, Rostock, N6, pp. 106–110.

Popeskul, A. N. (1991). *Production-Network Approach to Modeling the Meaning of a Scientific Text*. In Russian. Dissertation for the Degree of Doctor of Technical Sciences, Vinnitsa Polytechnical Institute.

Prekup, A. V. (1989). *Modeling and Implementation of a Stratificational Linguistic Machine*. In Russian. Dissertation for the Degree of Candidate of Technical Sciences, Kishinev Polytechnical Institute, pp. 43–45.

SILOD (1988). A Russian–English Translation Support. Computronics India, Delhi.

Zubov, A. V. (1985). *Probability-Algorithm Model of Text Generation (The Semantic/Syntactic Aspect)*. In Russian. Dissertation for the Degree of Doctor of Philosophy, Voennyj Krasnoznam Institut, Moscow.