Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/6010
DOI URL: http://dx.doi.org/10.21474/IJAR01/6010

RESEARCH ARTICLE

PRINCIPLES OF ALGORITHMIZATION OF MACHINE TRANSLATION IN THE THEORIES OF LANGUAGE EVOLUTION (BASED ON THE UKRAINIAN-ENGLISH TRANSLATION).

Rita Kalko.
by-st. Vokzalny, 68, Sloviansk Donetsk region, Ukraine 84109.

Manuscript Info

Manuscript History

Received: 09 October 2017
Final Accepted: 11 November 2017
Published: December 2017

Key words:-
Algorithms, a Machine Translation, a Language Evolution, Linguistic Theories.

Abstract

This paper deals with the issue of the main principles of algorithmization of machine translation in the theories of language evolution. The aim of this article is to offer the original algorithmic principles of machine translation of natural languages of the Indo-European group with the optimal mathematical dimension of the text and the accuracy acceptable for practical translation tasks based on the classic linguistic theories of A.Schleicher, I.Schmidt, etc., using the phenomenon of the weak self-similarity of Mandelbrot. One of the most disputed questions in the linguistics is the problem of machine translation in the theories of language evolution. Reference has been made to the reasons that determine the necessity of the research into the linguistic theories. Attention has been paid to the main theories of language evolution: a content-forming "family-tree" theory by A.Schleicher and J.Schmidt's theory of "waves". It is stated that the involvement of linguistic theories allows us to create algorithms that reject a large number of non-grammatical constructions and create automatic criteria for choosing the best translation variant.

Introduction:-

Since the 19th century, there has been a rapid rise in the linguistic studies. Scientists have mostly focused their research on the problem of studying the evolution of different language groups. In general, such studies had tended to focus on the two main linguistic theories: a content-forming "family-tree" theory by A.Schleicher and J.Schmidt's theory of "waves". The theory of "family-tree" by A.Schleicher has explained the origin of the languages of the Indo-European family by means of divergence, that is, the gradual division of the dialect. The theory was presented in the form of a scheme that resembles a branchy tree. It is important to note, that Schleicher's "family tree" image based on the mathematical theory of bifurcation, i. e. "fractal tree". It can theoretically simulate any natural phenomena, including language. It should be emphasized, that in the image of Schleicher's "genealogy tree", it is useless to look for the strongest type of self-similarity, when the fractal looks the same at different magnifications as in abstract mathematical objects of the type of Cantor's set. A mathematically similar process will be described by models that distinguish between "almost self-similarity" and "statistical self-similarity". The isolated phenomenon of the self-similarity of a fractal tree, even in the weakest statistical version, theoretically enables researchers to continue the process of fragmentation of the branchy structure of the imaginary term indefinitely with the help of the idea of "recursive self-similarity" by G.Leibniz. We have to admit that for our study the phenomenon of self-occurrence within the "genealogy tree" of A.Schleicher is more important.

Corresponding Author:- Rita Kalko.
Address:- by-st. Vokzalny, 68, Sloviansk Donetsk region, Ukraine 84109.

Theoretical Reviews:-

In general, the "family tree" theory by Schleicher was imitating the structure of biological theories of evolution. Therefore, this theory has not escaped criticism. The main criticism of opponents reminded Darwinism criticism. It was the lack of the known intermediate forms between selected biological or linguistic phenomena. The model of "waves" by I.Schmidt was an alternative to the Schleicher's theory. According to I.Schmidt, all modern languages of the Indo-European family have relationships that make it impossible to identify the language in a separate group. He stated, that these languages are the links of one chain. Therefore, along with the theory of "family tree" there is a theory that modeled the origin of languages in the form of gradual transitions from one language group to another on the background of close contact (given on the convergence), founded on the principle of linguistic continuity.

The indicated language models (Schleicher, Schmidt) correspond to the main characteristics of the sets theory. The theory of sets of Zermelo-Frenkel (ZF) is considered the most known formal (syntactic) theory of the language **(Ebbinghaus, Heinz-Dieter, 2007)**. ZF is a formal theory of the language with one predicate symbol and double symbol of equality and theoretically can explain all the phenomena of the natural and artificial speech. At this point, it is sufficient to use the general theory of sets to simplify the conceptual apparatus for explaining the mentioned phenomena. Consequently, symbols, alphabets and languages will be further studied by classical mathematical sets. In our study, the notion of the theory of sets will appear in the general and mathematical language. All the conclusions can theoretically be extended to the formal language of the theory of ZF. Using the notion of the self-similarity of the mathematical model of "genealogy tree" as to the Indo-European languages (A and B), we use the basic notions of set theory in the context of translation practice. Therefore, for the translation from the language (A) into the language (B), the following basic relations are possible:

1. The "inclusion" relationship (denoted by A □ B): the translation of the word will be identical.
2. The intersection (or product) (denoted as A ∩ B) is a set that includes all the objects that are elements of the set (A) and the set (B) at the same time: the translation of the word will be partially similar. It can be explained by the transformation ascending material.
3. The empty set, which does not contain any element (denoted as Ø): the translation of the word will be completely non-original. Thus, it cannot be explained by the transformation of the ascending material **(Kalko R. 2017)**.

The studied language (A or B) will be a plurality of individual words as a tuple of characters encoded by symbols of a certain alphabet with the main types of relations between them: A □ B; A ∩ B and Ø. In what follows, we will focus on "Kleene's recursion theorem" as the theoretical basis of these relations **(Kleene St., 1938)**. Taking advantage of "Kleene's recursion theorem", we pay special attention to the topological correspondences of this theorem, in particular the Borsuk-Ulam theorem **(Matoušek J. 2003)**. Brauer's Theorem and Banach's Theorem prove that the transformation of the initial set of language A into the set of language B can be used to determine the correspondence in the form of "immutable point" or "recursion point".

Today, the content-forming for understanding the theory of language is considered a phonemic principle. According to this principle, each language of the world can contain a limited number of characteristic sounds, as usually from 25 to 35. The common intellectual roots and the global process of transition from the hieroglyphic signs to the phonemes determine the self-similarity of European alphabets. Thus, the phonemic alphabet as a system of symbols owes its appearance to hieroglyphic signs of the Eastern (Egyptian) type. At this point, the evolution from the symbol to the phoneme in the alphabet can be traced by the following example:

1. the Egyptian hieroglyph "Water" was marked by a wavy icon that resembles (M + N);
2. The Phoenicians had a sign "Mem" ( );
3. The Greeks used the letter μ;
4. Latin and other European languages had the letter M.

Part of the Phoenician sign "Mem" is the "Nun" Snake (fish). Its image () gives the Greek letter η and Latin N. In general, the Phoenician alphabet as the basis of the European writing consisted of 22 letters: (). Having originated from the pictograms, they were used to denote consonant sounds **(Encyclopedia and dictionaries. The New Book of knowledge. 2002)**. Thus, in the east, the Phoenician alphabet was the basis of the Aramaic writing. The Aramaic writing has given rise to others modern writing systems such as Hindi, Persian and Arabic. The Phoenician alphabet was the basis of ancient Greek literature. The Greeks called their alphabet "Phoenician Signs". It included 22

consonant signs and 5 vowels. Only 15 signs of 25 characters of the Greek alphabet: (α; β; γ; δ; ε; ζ; η; θ; ι; κ; λ; μ; ν; ξ; ο; π; ρ; ς; σ; τ; υ; φ; χ; ψ; ω) coincide with the corresponding signs of the Phoenician alphabet. As to the rest of the signs, the Greeks simply did not have the corresponding phonemes. The Greek alphabet from the Etruscans moved into a new form of the Latin alphabet: (a; b; c; d; e; f; g; h; i; j; k; l; m; n; o; p; q; r; s; t; u; v; w; x; y; z) and became the basis of the creation of the Slavic Cyrillic (Cyrillic). So, almost all alphabets of the world originated from the Phoenician and Greek alphabets, including the Ukrainian alphabet. On the account of the facts, it can be asserted, that the evolution of ethnic abbreviations repeats the model of "family tree" with the branchy structure of probabilistic branching and weak self-similarity.

Procedure and Methodology:-

The proposed model can be mathematically implemented in the form of a number (system of permissions) and an algorithm (sequential commands). The calculus has the form of a mathematical system that includes:

1. the output (primary or not defined) notions whose names form the "alphabet of symbols". In the research, it is the alphabet of natural languages: from (A) to (B) where the machine translation is carried out;
2. the primary statements about the connections between these notions or based on the linguistic theories of the evolution of the axiom on the constructive recursive relationship between the sets of alphabets (A') and (B') arising from the common root through divergence;
3. the rules for the withdrawal of new allegations from existing ones: use of Kleene's theorems on recursion and its topological correspondences (Borsuk's theorem - Ulam, Brauer and Banach). The numerology allows us to specify, with the help of the end-device, all the objects of a certain set, including the infinite (for example, all sentences of a given language). In linguistics, this property of calculations is used, dealing with a very large or potentially infinite number of units **(Nikitina F., 1987)**.

The main problem of machine translation, based on the "travelling salesman problem", will be quadratic slowdown while searching for the optimal variant of translation. For example, you need to translate a simple sentence of five simple words of the A language to the language B. "Simple" word of language A has five variants of the language B. In this case, the algorithm of calculation has to choose the variant of translation from the following options: (5 × 5 × 5 × 5 × 5 = 3125).

Taking into account the phenomenon of self-similarity of the branchy structures, the machine translation program of the set of coding alphabet A' of language A into the language B, composed of the set of alphabet B', will be based on the transformation of the word into the sum of the tuple of digits, which is calculated on the principle:

$$\Sigma = (1 \times N) + (2 \times N') + (3 \times N'') + ...$$

It should be pointed out, that the numbers: 1,2,3 … are the ordinal numerals of letters in the word. Accordingly, N, N', N".... ordinal numerals of letters in the national alphabet. At above-mentioned formula, Σ is equal to the sum of the constituent letters of the alphabet X' at the root of the word X. The resulting number is compared with the results obtained from the same procedure with the variants of the translation. Taking into account the dimension of the fractal tree, an option is chosen among the variants the number of which will be closer to the original text.

Having translated the word "red" from Ukrainian into English we got some possible translation options: 1) Red (рудий, рум'яний, *червоний*); 2) Ruddy (рум'яний, яскраво-червоний, *червоний*); 3) Pink (ліберальний, нестерпний, *червоний*, рожевий); 4) Blushful (рум'яний, сором'язливий, соромливий, *червоний*); 5) Florid (крикливий, *червоний*, яскравий). The Ukrainian language has an alphabet of 33 symbols а, б, в, г, ґ, д, е, є, ж, з, и, і, ї, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ь, ю, я).

Having expanded the root of the word "ЧЕРВ" according to the formula 1: Ч + Е + Р + В = 1× Ч + 2 × Е + 3 × Р + 4 × В = 1 × 28 + 2 × 7 + 3 × 21 + 4 × 3 = 117. The English alphabet consists of 26 characters (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z). Using the same methodology, we analyze the roots of the English words "RED" = R + E + D = 1 × 18 + 2 × 5 + 3 × 4 = 40; "RUD" = R + U + D = 1 × 18 + 2 × 21 + 3 × 4 = 72; "PINK" = 1 × 16 + 2 × 9 + 3 × 14 + 4 × 11 = 120; "BLUSH" = 1 × 2 + 2 × 12 + 3 × 21 + 4 × 19 + 5 × 8 = 205; "FLOR" = 1 × 6 + 2 × 12 + 3 × 15 + 4 × 18 = 147.

That is, we have the word (Red) of the language A is codified by the alphabet A' of 33 symbols. Its correspondences in the language B with the alphabet B' of 26 characters will have 5 variants (Red, Ruddy, Pink, Blushful, Florid). After processing the original algorithm, created on the basis of evolutionary theories of language development and based on the idea of self-similarity of branchy structures, we get tuples that can be objectively compared with each

other. The ascending word A of the alphabet A' corresponds to the number 117. As follows, the words B of the alphabet B' to the numbers 40, 72, 120, 205, 147. Since the set of the alphabet A' is greater than B' (33> 26), the dimension of the ascending word will also be greater than its corresponding ones, that is, variants (Pink, Blushful, Florid) (120, 205, 147) are rejected as non-grammatical. The most successful version of the translation will be the second (72 and 117) or "Ruddy", in tune with the Ukrainian word "Rudy".

In order to prove this, the image of the Dark Lady of the Shakespearean Sonnetary have been used for interpreting. The direct translation of the word Dark (безрадісний, порочний, смаглявий, темний, чорний, неосвітлений); (joyous, vicious, dark, dark, black, unlit) allows non-grammatical forms up to the "Devil Woman". Considering the theories of the evolution of the Indo-European languages, we tend to variants of the translation **"**Dark Lady**",** which emphasizes the mysterious nature of the terrible ambivalent image of Frey's type:

І сяючі карі очі мого кохання / Так схожі на краї лелечих крил / Що носять темне вбрання / На цвинтарі жалобних брил.

Ні, не схожі з сонцем її очі / І блиск коралу червоніший за її вуста / Якщо ж сніг білий, чому ж душа її пітьма / Волосся грубе наче дріт з плечей її звиса **(Kalko R. 2010)**.

Conclusions:-

Traditional programs of machine translation are based on solving the mathematical "travelling salesman problem". It involves a mechanical override of theoretically possible options. Taking into account, the quadratic slowdown, such a task becomes almost impossible when there comes a situation of "information explosion" and the program is lost among a large number of equal choices. In our opinion, the involvement of linguistic theories allows us to create algorithms that reject a large number of non-grammatical constructions and create automatic criteria for choosing the best translation options. The theoretical basis for the creation of such programs will be the idea of translation as interaction set of alphabets of natural languages, taking into account the weak self-similarity of branchy structures.

References:-

1. Ebbinghaus, Heinz-Dieter. (2007): Ernst Zermelo: An Approach to His Life and Work. Springer.
2. Encyclopedia and dictionaries. The New Book of knowledge. (2002) Grolier Incorporated, Danbury, Connecticut. Longmans Green and Compony.
3. Kalko R. M. Interpretatsii perekladiv vybranoho sonetariiu V. Shekspira: monohrafiia / Kalko Rita Mykolaivna. – Luhansk: «Svitlytsia», 2010. – 89 s.
4. Kalko R. Universalii movnoi svidomosti (2017) LAP LAMBERT Academic Publishing RU.
5. Kleene St. (1938): On notation for ordinal numbers.The Journal of Symbolic Logic: 150 – 155.
6. Matoušek J. (2003) Using the Borsuk-Ulam theorem. Springer Verlag, Berlin,.
7. Nikitina F. O. Strukturni ta matematychni metody vyvchennia movy / F. O. Nikitina. – K.: Vyd-vo pry Kyiv. derzh. un-ti, 1987. – 36 s.