AUTOMATION OF INFORMATION BASE DEVELOPMENT FOR MULTILINGUAL ADAPTIVE TRAINING TECHNOLOGIES

Prof. Dr. Alena Stupina[1]
Assoc. Prof. Dr. Irina Bagdasaryan[1]
Assoc. Prof. Dr. Margarita Karaseva[2]
Assoc. Prof. Dr. Olga Almabekova[1]
[1]Siberian Federal University**, Russia**
[2] Siberian State Aerospace University**, Russia**

ABSTRACT

The issue of effective and efficient foreign language teaching is relevant in today's global world. More and more frequently specialized computer programs for foreign vocabulary training are used. They are not expensive and simple to create and use in comparison with published analogues and not less efficient.

The paper considers systemic aspects of information base development for multilingual adaptive training technology, specifically the process of information collection of latent lexical relations and its employment in information terminological basis. The authors offer the system of primary text processing, its algorithms and structure of output data.

Creations of computer system aimed at training foreign vocabulary is presented in the consequence of steps, based on the Markov processes theory.

Employment of text processing subsystems, created by integration of subsystem of frequency dictionary generation and subsystem of latent vocabulary relations search is a successful solution in building of information technological basis.

Using latent lexical relations increases the efficiency of training system of foreign vocabulary in general. The developed structure and algorithm are nor resource consuming, thus making them economical. .

The developed structure of output data for preliminary text processing subsystem provides information flexibility and integrity.

Keywords: multilingual technology, adaptive training, terminological basis, automation.

INTRODUCTION

The problem of the effective foreign language training has always been important for people. Various methods of teaching, a lot of textbooks, dictionaries (including electronic ones) have been created. There were attempts to build a unified language of rather simple and logical structure. The results of these efforts are quite debatable but one thing is certain: all efforts were caused by the necessity to understand foreign speech, ability to speak within the special field. In addition, there was an urgent need for the qualified interpreters. But, whatever qualifications a translator has, he is unable to comprehend the diversity of human activity fields with specialized terms. Besides, the meanings of ones and the same terms in different fields are different. And that is why,

as we know, an interpreter helps us to communicate [1] with representatives of other language groups in all spheres of human activity.

Having summarized all the above mentioned, we can prove that nowadays it is preferable that the specialist personally could communicate in foreign languages with foreign colleagues or partners, at least within its special field. In Russian practice, unfortunately, this trend is developing slowly. And that is why there exists the necessity to develop effective methods of the specialized vocabulary training.

Today more and more often they use special computer programs training foreign vocabulary. They are relatively cheap and simple in development in comparison with the paper ones and of the same effectiveness. Such programs usually have a rather flexible structure that allows us to update the databases and replace them for training the vocabulary of other special fields [2].

It is noteworthy that the training process using such programs is individual, and a student is able to interrupt or resume the training process at any convenient time. This is also important that such training programs usually have some additional functions in training including media. The effectiveness of such programs also increases greatly.

The only disadvantage of such tools of training compared with paper ones is that some people fundamentally do not accept computer training at all [3]. The reasons for this are obvious. The fact is that training with a help of the computer requires some perseverance and patience from a student that not everyone can have. Naturally, a user must have the ability to work with a computer, at least at the fundamental level.

But, nevertheless, the majority of professionals such training programs are focused on should not experience any difficulties in their application. The reason for this is the higher level demands of today's professionals. It means that such training programs will be in demand in the market of foreign language training.

Recently modern specialists are required to know the terminology of several foreign languages.

The solution to this problem by means of the software products described above has some disadvantages. There often exists a mismatch of linguistic analogues. Often some analogues are forgotten faster than others. And it intensifies the training process.

Of course, this is primarily due to the fact that the process of training has some in stages. Multilingual adaptive – training technology (ML - technology) [4] offers an alternative approach to training several foreign languages sequentially, that is, training any foreign language taking into account the knowledge of the other previously studied foreign language. ML-the technology is based on the mechanisms of perception and memory of a person and provides the adaptation of the system to a specific user.

Nowadays in the framework of ML technology a number of studies are conducted and ML-technology as a core of the training system acquires new methods and systems of their implementation.

The development of the computer system training foreign terminology (in particular, on the basis of ML - technology) can be represented in the form of several successive stages.

The formation of the shell and the mechanism for the system movement includes various mechanisms of adaptation and additional functions.

The formation of terminological basis [5]:

- Development of the lexical database appropriately, and adequately reflecting the specifics of the certain field of any foreign language (some foreign languages, in the context of the ML - technology).

a. Search for the texts belonging to the particular special field.
b. Word processing and garbage collection.
c. Building a lexical database as a frequency dictionary.

- Formation of the multilingual terminological information basis using different methods including methods of the structural basis optimization.
- Quite often to describe technological processes and control the theory of Markov processes is used.

The Markov process is a process where for each moment of time the probability of any state of the object depends in the future only on the state of the object at this very moment and does not depend on how the object came into this state [6, 7]. Such a process is said to have the property of the Markov process.

The Markov chain is called the Markov process with the discrete time given in the measurable space.

To describe the Markov processes they use the Markov models that include some states, a set of transitions among these states and the probabilistic characteristics of these transitions (the transition probability or the probability of the transition). For the appropriate use of the transition probabilities, they are often written as a matrix (the matrix of the transition probabilities).

The example of the Markov chain is given in Fig. 1, which illustrates the pronunciation of the word “корова” in two different ways. According to this figure, a word "корова" with a probability of 0.7 to be pronounced as [karova], with a probability of 0.3 as [корова].

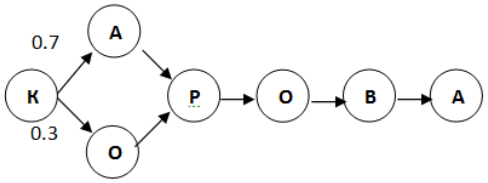

Fig.1The example of the Markov chain

All the given states except the initial one are functions of the previous state. The set of states given in this example, it is logical to consider as the combination of the bigrams (sequences of two words) because each state (except the initial one) is a function of only one previous state.

When a certain state depends on several previous states, then the combination of these states is logically to be considered as a set of trigrams, tetragrams and so on (the latter one is not often used in solving technological problems, as the construction of the system based on them is rather expensive). Here it is appropriate to talk about the wideness of the notion “the Markov process”.

If the Markov chain includes at least one state that can be achieved by single transitions from several previous states (Fig. 1, state No. 3), such a chain is called a hidden Markov chain (has a hidden Markov property).

We will refer to arrows (reflecting the possibility of the transition of the hidden Markov chain) as hidden Markov dependencies (relations), and in the context of this paper, taking into account the subject – hidden lexical relations.

Now it should be noted that the structure of the computer system of foreign language vocabulary training does not take into account hidden Markov dependencies among terms (hereinafter refer to hidden lexical relations).

However, it is obvious that such dependencies (primarily links between terms but not lexemes) can serve as a sufficiently strong associative mechanism when memorizing a set of lexemes of a language (languages) [8]. Hence, the study of this mechanism and possibilities to manipulate it fully corresponds to the objectives of the training systems.

Let's suppose that there exists a mechanism for data using of the hidden lexical relations in building a terminological information basis.

Then it is necessary to build a system that would allow us to find and present information about the hidden lexical relations so that they could be used by the previously mentioned mechanism. Two basic requirements are presented to such a system:

- the given system should fit the overall structure of the computer training system (the lowest resource consumption, synchronization with separate subsystems);
- the universality and simplicity of data presentation (mechanisms of use may be varied the same as the requirements for the representation of their input data).

It is naturally to refer the integrated system to the second stage in developing the system of foreign language training (pre-processing of texts, building lexical databases).

Also it would be natural to place it in the given structure either before the application of the subsystem generating a frequency dictionary or after that.

But if you examine the structure of both these systems in details, it becomes obvious that they are based on the same algorithm of word processing – the algorithm "Pattern search in a line". And as the resource consumption of the problem with this algorithm application is directly proportional to the volume of the processed text. Then a problem arises about these two systems combination.

The successful solution of this problem will greatly reduce the resource consumption of the stages of the lexical base in comparison to the sequential application of these two systems [8, 9].

In order to understand how efficient and easier to organize the presentation of output data of the system, it is necessary to know what these data are.

All the hidden lexical relations in the text can be represented as a directed weighted graph. Those vertices correspond to individual lexemes, oriented arc corresponds to the hidden lexical context, and the weight corresponds to the probabilities of the transition among lexemes (Fig.2).

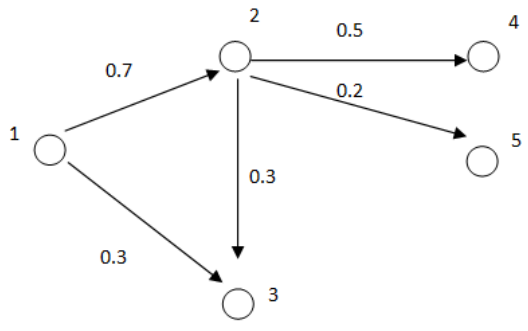

Fig. 2 Example of the digraph part reflecting the hidden lexical text dependencies of some subject field

Table 1 Frequency characteristics of lexemes

№	Lexemes	Data from frequency dictionary
1	Computer	0.007
2	System	0.002
3	Design	0.0006
4	Architecture	0.0001
5	Method	0.003

The interpretation of this graph with the respect to the given example, should be done in the following way:

- Vertex №1:
- Lexeme: Computer

After the lexeme «Computer» the following lexemes are given:

- «System» with the probability 0.7,
- «Design» with the probability 0.3.

It's the same about the vertices: 2,3,4,5.

Digraphs and any other graphs more often have the machine representation as a set of adjacency matrix and incidence. Such matrixes for the digraph shown in Fig. 1 will be as follows:

Table 2. Adjacency matrix

№	1	2	3	4	5
1	0	1	1	0	0
2	1	0	1	1	1
3	1	1	0	0	0
4	0	1	0	0	0
5	0	1	0	0	0

Table 3. Incidence matrix with edge weights

№	12	13	23	24	25
1	-0.7	-0.3	0	0	0
2	0.7	0	-0.3	-0.5	-0.2
3	0	0.3	0.3	0	0
4	0	0	0	0.5	0
5	0	0	0	0	0.2

For easy storage and minimal resource consumption data about the hidden lexical relations, we consider the following modification of the adjacency matrices and incidence (which represents the only one table) and it is equivalent to the matrix of transition probabilities, usually used in the context of the Markov chains.

Table 4. Matrix of transition probabilities

№	1	2	3	4	5
1	0	0.7	0.3	0	0
2	0	0	0.3	0.5	0.2
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0

This matrix should be read regarding the lines where, for example, the entry in the cell M[1,2] is"0.7" means that the relation weight 0.7 is included into the point 2 from the point 1. In other words "0.7" is the value of the probability transition from lexeme No. 1 to the lexeme No 2.

The ability to combine the adjacency matrix and incidence matrix were due to the fact that two digraph vertices of the hidden lexical relations connects one and only one edge that allows not to index the them; and the presence or absence of edges gives clearly a non-zero value of the corresponding cell of the matrix of the transition probabilities [8].

The output of the building system of the frequency dictionary can be represented, for example, in the form of a small database:

Table 5.A frequency dictionary

ID	Frequency	English	German	Russian
1	0.007	Computer	Computer	Компьютер
2	0.002	System	System	Система
3	0.0006	Design	Design	Дизайн
4	0.0001	Architecture	Architektur	Архитектура
5	0.003	Method	Methode	Метод
..	…	…	…	…

This is useful first of all due to its flexibility and reliability of data storage structures where flexibility depends on the chosen methods of forming terminological basis and of the selected training methods. It is possible to modify the database by adding, for example, an attribute: transcription for each language version. The reliability is ensured by the structure of the database.

The matrix of transition probabilities and frequency characteristics of lexemes must be synchronized; it is offered to represent a matrix of transition probabilities in the form of a database and to combine with the database of the frequency dictionary by ID (a unique number):

Table 6.The hidden lexical relations

ID	1	2	3	4	5
1	0	0.007	0.023	0	0
2	0	0	0.03	0.07	0.08
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0

Thus, the presentation of the output data of the search system for the hidden lexical relations in a matrix of transition probabilities will be synchronized with the output of the frequency dictionary and it is to be the least resource-intensive.

The algorithm is presented at a certain level of abstraction; therefore, it does not include the detailed descriptions of some items. The patterns search in the line here can be produced by any of the existing special methods. The authors of the paper, based on the results of their own research in the field of such problems solving, recommend the method of "Quick search", with minimum resource consumption.

Input: a text that has undergone the preliminary processing (garbage gathering, etc.).

The lexeme is selected starting with the first position of the text. Let's call it the main one because it represents the state of the process in real time and it takes into account the frequency for frequency dictionary. The lexeme is written into the database: in to the frequency dictionary and into the matrix of transition probabilities as an element (if they have not already met as an attribute) as a new attribute with the name identical to its own ID of the primary lexeme, the attribute meaning is temporarily equal to zero.

The current frequency for the primary lexeme is changed. The corresponding note is made in the frequency dictionary as a database.

The lexeme following the basis one is selected. Let's call it as a related lexeme. It presents the state of the process at the future moment of time and reflects a hidden lexical relation. If the associated lexeme has not met in the current pair, then it is put into the database: in the frequency dictionary as a new element with its ID and into the matrix of transition probabilities as a new attribute.

The frequency value for the current pair of lexemes is changed. The corresponding notes are put into the database (a matrix of transition probabilities).

The lexeme identical to the basis lexeme is found in the text (by any methods of the patterns finding in the string).

If the search was successful and the desired lexeme is found, the algorithm goes to point No. 2 and continues working.

If the observation of the text is complete, and the required lexeme is not found the algorithm goes to point No. 1 giving to the lexeme that follows it the first entry into the text of the main lexeme the status of the basis one. The algorithm starts working relatively it. And it works till the whole text will not be passed this way up to the end.

Output: of the database.

The use of the subsystem that processes texts obtained as a result of the unity of the subsystem generating a frequency dictionary and the subsystem of the hidden lexical relations search is very important in building information and terminological basis. On one side there is an opportunity to engage the student’s strong previously unavailable associative mechanisms of memory perception while training and on the other hand, the capacity of the basis building will be much lower than if we used two previously mentioned subsystems.

Conclusion: Thus using the hidden lexical relations, in general, the effectiveness of the foreign language vocabulary system training increases. Thus the developed structure and algorithm of subsystem texts pre-processing when the use of this subsystem in the information and terminological basis formation will be the least resource-intensive [1,7].

The structure of the output data of the subsystem of the texts pre-processing provides the flexibility and integrity of the information.

And in the effective use of data about the hidden lexical relations directly in the training process opens the opportunity for the new research in this field.

REFERENCES

[1] Karaseva M.V., Leskov V.O., The system of the information and terminological basis formation for multilingual adaptive-training technology, Vestnik of SibSAU, Vol. 4, Krasnoyarsk, Russia, 2007, pp 31-35.

[2] Agapova O.I., Krivosheev A.O., Ushakov A.S., About three generations of computer technology training, Informatika I obrasovanie, №2, Russia, 1994, pp 34-40.

[3] Azimov E.G., To the typology of educational computer programs for Russian as a foreign language, Sovremennie tekhnicheskie sredstva v obuchenii russkomu yaziku kak inostrannomu, Russia, 1990, pp 136-143.

[4] Alexandrov G.N., Programming training and new information technologies in education, Informatika I obrazovanie, №5, Russia, 1993, pp7-19.

[5] Brusilovsky P.L. Intelligent educational systems, Informatika, Informatsionnie tehnologii, Sredstva I sistemi, №2, Russia, 1990, pp 3-22.

[6] Karaseva M.V., Kustov D.V., Information and terminological basis in multilingual adaptive-training technology, Krasnoyarsk,Vestnik of SibSAU, Vol. 3(36), Russia, 2011, pp 29-31.

[7] Kovalev I.V., Suzdaleva E.A. The information base of the multilingual information and training technology formation, Nauchno-innovatsionnoe sotrudnichestvo: Works of the scientific conference. Scientific session of MIFI, Vol. 3, Moscow, Russia, 2002, p 137.

[8] Klarlashuk V.I., Training programs, Solon-R, Russia, 2001.

[9] Johns M.T. Programming of the artificial intelligence, USA, 2004.