Institute of Information and
Computational Technologies

ISSN : 2788-7677 (Online)
ISSN : 2788-7987 (Print)

ADVANCED TECHNOLOGIES AND COMPUTER SCIENCE

2020
№1

www.atcs.iict.kz

Institute of Information and Computational Technologies,

Advanced Technologies and computer science

№1

Almaty 2020

Institute of Information and Computational Technologies,

Advanced Technologies and computer science

The publisher, the authors, and the editors are safe to assume that the advice and information in this journal are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published works and institutional affiliations.

28 Shevchenko str., Almaty,Republic of Kazakhstan
7 (727) 272-37-11
atcs@iict.kz

About the Journal

Advance technologies and computer science is a bilingual scientific peer-reviewed, interdisciplinary, electronic journal of open access, including thematic areas:

- Section **"Applied mathematics, computer science and control theory"** includes papers describing modern problems in these areas.
- Section **"Information and telecommunication technologies"** also includes the following topics:
  - Data transmission systems and networks.
  - Internet technologies.
  - Cloud technologies.
  - Parallel computing.
  - Distributed computing.
  - Supercomputer and cluster systems.
  - Big data processing (Big-data).
  - Geographic Information Systems and Technologies.
- In the section **"Artificial intelligence technologies"** in addition to technology, there are works on topics:
  - Intelligent Management Systems.
  - Speech technology and computer linguistics.
  - Pattern Recognition and Image Processing.
  - Bioinformatics and biometric systems.
  - Human-machine interaction.
  - Machine learning.
  - Intelligent Robotic Systems.
- The section **"Information Security and Data Protection"** also covers topics:
  - Software and hardware information protection.
  - Mathematical methods for ensuring information security of complex systems.
- The section **"Modeling and optimization of complex systems and business processes"** may include:
  - Computational mathematics, numerical analysis and programming, mathematical logic.
  - Theory of Statistics.
  - Statistical Methods.

Editorial Team

Chief Editor

Editorial team

Contents

IRSTI 76.13.23; 15.81.00

UDC 004.94-57.090

CONSTRUCTION OF A MATHEMATICAL MODEL OF THE INFLUENCE OF VARIOUS ENVIRONMENTAL FACTORS ON THE PLANT BIOMASS ON THE CONTAMINATED SOIL BY TOXIC ELEMENTS

T. Mazakov[1], Sh. Dzhomartova[2], P. Kisala[3], Ch. Nurzhanov[4], A. Mazakova[5], B. Zhakyp[6]
[1]Institute of Information and Computational Technologies,Almaty, Kazakhstan
[3]Lublin Technical University, Poland
[1,2,4,5,6]al-Farabi Kazakh National University, Almaty, Kazakhstan
[1]tmazakov@mail.ru, [2]jomartova@mail.ru, , [3]p.kisala@pollub.pl ,
[4]DarkEremite@yandex.kz, [5]aigerym97@mail.ru, [6]zhakyp.botagoz@mail.ru
[1]ORCID ID: https://orcid.org/0000-0001-9345-5167
[2]ORCID ID: https://orcid.org/0000-0002-5882-5588
[4]ORCID ID: https://orcid.org/0000-0002-9760-3101

Abstract. The article has been devoted the development of a program for constructing a mathematical model based on the processing of experimental data for modeling plant biomass on soil contaminated with toxic elements.

In general, the problem under consideration is relevant in areas where forecasting based on analysis of data for previous points in time is required in order to take into account the relationship between the values of a certain set of factors and the behavior of an object or in the process which represented by a time series

The proposed software product for solving the problem of analyzing multidimensional time series is universal and can be applied in various fields: in ecology, medicine, chemistry, pharmacology, economics, and others. Models constructed using this method are superior to regression models in their predictive properties, due to the fact that, thanks to the application of heuristic principles of self-organization, automatic selection of informative input variables and selection of the structure of the optimal regression model. The construction of mathematical models from experimental data is carried out automatically. In the process of modeling, the main task is the identity of the form dependence and the choice of factors that have a significant impact on the dependent variable. Moreover, the structure of the model, in contrast to the regression analysis, is not fixed in advance, but is selected from a variety of options according to the absolute error criterion.

Using the self-organization program, we estimated the dependence of plant biomass (*Miscanthus*), which has a high biological absorption of heavy metals from contaminated soil from the precipitation, air temperature, potential evaporation of moisture in the soil and photosynthetic active radiation during the year, taking into account the approximation confidence value. Was obtained the mathematical models of biomass ‘plant depending from the environmental factors. The analysis showed that the predictors that have the greatest effect on plant biomass growing on soil contaminated with heavy metals are evaporation of soil moisture, photosynthetic active radiation and precipitation. The data obtained are relevant for predicting the processes of cleaning contaminated soil using plants.

Keywords: self-organization, selection, regression equation, time series, identification, heuristics, pollutants, environment, toxic elements.

Introduction

Acceleration of scientific and technological progress is largely determined by the level of application of computer technology in scientific research, in design and construction work, as well as for managing various production processes. Conducting applied research related to product quality control, process optimization, certification of complex technical products, solving sociological and economic problems, biomedical and agricultural research, research in demography, etc. often leads to the need for data analysis.

The task of analyzing time series is a very urgent problem in various fields of science and technology. Mathematical models that describe the statistical relationships between quantities, the change in time due to factors hidden from the observer, are an instrument for studying complex systems of processes that occur in the world around us. In a large number of cases, the initial information used to build these models and evaluate their adequacy is a time series (a sequence of results of measurements of the current values of one or more parameters carried out at ordered points in time). To explain the reasons for this or that behavior of a complex system or process that gave rise to a given time series, to identify and explain the laws of their dynamics, it is necessary to solve the analysis problem. To predict the dynamics of the development of a complex system or process, as a rule, based on the results of the analysis, the task of synthesizing a time series model is solved.

In general, the problem under consideration is relevant in areas where a forecasting problem is required based on analysis of data from previous time instants in order to take into account the relationship between the values of a certain set of factors and the behavior of the object or process which represented by a time series.

The current situation necessitates the study and formalization of the processes of building, analyzing and optimizing decision support systems based on time series analysis from the point of view of the technical, algorithmic and structural aspects and confirms the relevance and scientific and practical significance of this development.

Among the tasks successfully simulated on computers, a special place is occupied by biological sciences, in particular, molecular biology aimed at genomic research [1], an agricultural complex for planning, forecasting, analysis and modeling of agricultural processes [2, 3] and ecology aimed at preventing the harmful effects of economic and other activities on natural ecological systems [4-6]. Improving the environmental information system with increasing levels of soil/water pollution by various toxic elements as a result of human activities is one of the priority areas in the field of information technology. The concentration of toxic substances in the environment is an acute environmental problem, especially in agricultural and industrial regions [7-11].

The practical value of the obtained results lies in the creation of a set of tools applicable for the construction of operational decision support systems based on the trend analysis of time series.

Software implementation

The use of the concepts and ideas of cybernetics in matters of physics, chemistry, biology, sociology, psychology and other sciences gave excellent shoots, allowed to deeply advance into the essence of the processes taking place in inanimate and living nature. There is no doubt that the coming 21st century and the progress of natural science and science throughout will proceed along the lines of studying the laws of control processes in complexly organized systems. A self-organizing system is a cognitive model of 21st century science.

One of the methods of self-organization is the method of group accounting of arguments (MGUA) [12-13]. The MGUA is based on the principle of multi-row selection of self-organization models, and the MGUA algorithms reproduce the scheme of mass selection. In the MSUA algorithms, the members of the generalized Kolmogorov – Gabor polynomial are synthesized and selected in a special way [14-15]. This synthesis and selection is carried out with increasing complication, and it is impossible to predict in advance what final form the generalized polynomial will have. First, simple pairwise combinations of the initial features are usually considered, from which the equations of the decisive functions are composed, usually not higher than the second order. Each equation is analyzed as an independent decisive function, and according to the training development in one way or another, the values of the parameters of the composed equations are found. Then, in a certain sense, the best ones are selected from the obtained set of decision functions. The selected partial decision functions are considered below as intermediate variables that serve as initial arguments for a similar synthesis of new decision functions, etc. The process of such hierarchical synthesis continues until an extreme criterion of the quality of the decision function is reached, which in practice is manifested in deterioration of this quality when trying to further increase the order of members of the polynomial relative to the original features.

Self-organization algorithms are used to solve problems of pattern recognition, predicting random processes, identifying multi-extreme static and dynamic characteristics and optimal control of complex objects.

Suppose that there is a set of input data in the form of a matrix *X* of *n* observations in the space of variable variables of dimension *m* > 1, which is characteristic of the standard problem of multiple regression. Let a training sequence of examples be formed in which each row of the matrix *X* is associated with a known value of the response *Y*, measured in a quantitative scale. It is necessary, using self-organization methods, to obtain a model expressing the law of change in the response ***Y*** depending on the specific values of the independent variables ***X*** [16].

The essence of the proposed algorithm of the multi-row heuristic method of self-organization is that it reproduces the selection scheme. Here is the full description of the object

$$y = f(x_1, x_2, ..., x_n) \tag{1}$$

replaced by several rows of private descriptions:
First row;

$$\begin{aligned} &z_1 = a_{11} * x_1 + b_{11} * x_2 + c_{11} * x_1 * x_2,\ z_2 = a_{12} * x_2 + b_{12} * x_3 + c_{12} * x_2 * x_3, \\ &...,\ z_k = a_{1k} * x_{n-1} + b_{1k} * x_n + c_{1k} * x_{n-1} * x_n \end{aligned} \tag{2}$$

where $k = n*(n+1)/2$.
Second row:

$$\begin{aligned} &\varphi_1 = a_{21} * z_1 + b_{21} * z_2 + c_{21} * z_1 * z_2,\ \varphi_2 = a_{22} * z_2 + b_{22} * z_3 + c_{22} * z_2 * z_3, \\ &...,\ \varphi_k = a_{2k} * z_{n-1} + b_{2k} * z_n + c_{2k} * z_{n-1} * z_n \end{aligned}$$

etc.

For a linear model, the coefficients c_{ij} are taken equal to zero.

Each particular description is a function of only two variables. Therefore, the coefficients of such a regression equation can be easily determined even by a small number of observations using the least squares method.

Using nonlinear private descriptions (2), models of almost any complexity can be obtained, since the degree of the polynomial doubles on each selection row.

The choice of polynomials is due to the property that, according by Weierstrass theorem [17-18], any function continuous on a finite interval can be represented with arbitrarily high accuracy as a polynomial of a certain degree.

The degree of a complete description of an object increases with each row of selection, and as a result, it is possible to determine the numerical values of the coefficients of an arbitrarily complex full description from a small number of field (experimental data).

During the selection process, variables are selected in accordance with the criteria for minimizing the absolute error functional:

$$F = \sum_{t=1}^{M} (y_t - \overset{\wedge}{y_t})^2 , \tag{3}$$

where y_t is the value of the indicator (1) at the time t; $\overset{\wedge}{y_t}$ is the predicted value of the indicator at time t.

At each step of the selection, the coefficients of the regression equation are determined using the least squares method for the three arguments.

Let the function $Y = Y(U,V)$ be given by the table, that is, the numbers U_i, V_i, $Y_i (i = 1,...,n)$ are known from experience. We will look for the relationship between these data in the form:

$$Y(U,V) = a*U + b*V + c*U*V , \tag{4}$$

where a, b, c are unknown parameters.

We choose the values of these parameters so that the smallest sum of the squared deviations of the experimental data Y$_i$ and theoretical

$$Y_i^{\wedge} = a * U_i + b * V_i + c * U_i * V_i ,$$

i.e. the amount:

$$\sigma = \sum_{t=1}^{M} (Y_i^{\wedge} - a * U_i - b * V_i - c * U_i * V_i)^2 -> \min \tag{5}$$

The value σ is a function of the three variables a, b, c. A necessary and sufficient condition for the existence of a minimum of this function is the equality to zero of the partial derivatives of the function σ with respect to all variables, i.e.

$$\frac{\partial \sigma}{\partial \mathrm{a}} = 0, \frac{\partial \sigma}{\partial \mathrm{b}} = 0, \frac{\partial \sigma}{\partial \mathrm{c}} = 0 \tag{6}$$

Since

$$\begin{aligned} \frac{\partial \sigma}{\partial \mathrm{a}} &= -2\sum_{i=1}^{n} (Y_i - aU_i - bV_i - cU_iV_i)U_i \\ \frac{\partial \sigma}{\partial \mathrm{b}} &= -2\sum_{i=1}^{n} (Y_i - aU_i - bV_i - cU_iV_i)V_i \\ \frac{\partial \sigma}{\partial \mathrm{c}} &= -2\sum_{i=1}^{n} (Y_i - aU_i - bV_i - cU_iV_i)U_iV_i \end{aligned} \tag{7}$$

then the system for finding a, b, c will have the form:

$$\begin{aligned} a\sum_{i=1}^{n} U_i^2 + b\sum_{i=1}^{n} U_iV_i + c\sum_{i=1}^{n} U_i^2V_i &= \sum_{i=1}^{n} Y_iU_i \\ a\sum_{i=1}^{n} U_iV_i + b\sum_{i=1}^{n} V_i^2 + c\sum_{i=1}^{n} V_i^2U_i &= \sum_{i=1}^{n} Y_iV_i \\ a\sum_{i=1}^{n} U_i^2V_i + b\sum_{i=1}^{n} U_iV_i^2 + c\sum_{i=1}^{n} U_i^2V_i^2 &= \sum_{i=1}^{n} Y_iU_iV_i \end{aligned} \tag{8}$$

This system is solved by the Cramer method [19-20].

In accordance with the self-organization algorithm, after each series of selection, K regression equations of the form are selected:

first row $z = f(x_i, x_j)$,

second row $\varphi = f(z_i, z_j)$,

third row $v = f(\varphi_i, \varphi_j)$,

fourth row $\omega = f(v_i, v_j)$, etc.

After completing the first selection row, the first N equations with the smallest error are selected. The resulting regression equations are denoted by $z_k = f(x_i, x_j), \mathrm{k} = \overline{1, \mathrm{N}}$.

The second and subsequent series of breeding are constructed similarly to the first.

Stopping the generation of models on subsequent rows occurs when, with an increase in the layer number, i.e., with a complication of the models, the external criterion of the best model does not decrease (3). The complication of the model ceases when further improvement of the selection criterion does not exceed a certain number ε (algorithm parameter).

The developed program is universal and can be applied in various fields to solve the problem of analyzing multidimensional time series - in space physics, seismology, medicine, finance, and others.

The program is implemented in the Delphi programming language [21-22].

Application of a software product

Next, we consider the application of the developed program to build a mathematical model of the influence of various environmental factors on the biomass of a plant, which has a high ability to absorb toxic elements from contaminated soil.

Let us consider the use of a multi-row self-organization algorithm for analyzing the relationship between the dynamics of the biomass plant (Miscanthus) and the following parameters: 1 – evaporation of soil moisture (PE), 2 – photosynthetic radiation activity (PAR), 3 – precipitation (Rainfall), 4 – temperature air (Temperature). The following information is located in the FisxP.txt file:

```
4 100 2
Biomass
PE
PAR
Rainfall
Temperature.
```

In the first line, the first number indicates the number of parameters, the second number indicates the amount of data, the third number indicates the type of model (linear – 1, nonlinear -– 2).

The FisxD.txt file contains line-by-line experimental data.

In the line, the first number indicates the value of the effective parameter, then the values of the auxiliary parameters are listed.

The result of the program is displayed in the Rezult.txt file, the contents of which are presented below.

```
          SELF-ORGANIZATION program parameters
Parametric source data
Number of arguments – 4
Amount of points –100
Initial data


    Biomass
  1 - PE
  2 - PAR
  3 - RAINFALL
  4 - Temperature
  1)     1.050 =      2.408     8.280     0.250    15.500
  2)     2.532 =      3.515     7.314     7.110    17.000
  3)     3.535 =      2.152     3.542     5.080    14.000
  . . . . . . . .
100)   1264.565 =      2.014     6.624     0.000    22.500
Nonlinear model
=========================================
SELF-ORGANIZATION RESULT
          Max      Min     Avg
Y     =   1264.565     1.050    544.753
```

```
X( 1) =     4.807     0.000     2.291
X( 2) =    14.030     1.460     7.782
X( 3) =    38.100     0.000     5.684
X( 4) =   111.000     6.500    21.242

       1 row results
              Error      Coef-t-1     Coef-t-2     Coef-t-3
 3( 1- 4) =     16.442745     0.687433     0.639918    -1.078038
 2( 1- 3) =     16.641309     0.686506     0.656516    -1.019773
 1( 1- 2) =     16.738916     0.790674     0.763545    -1.305119
 4( 2- 3) =     17.036969     0.715034     0.855330    -1.359676
 5( 2- 4) =     17.412309     0.649833     0.670954    -1.035383
 6( 3- 4) =     17.424851     0.742313     0.770697    -1.275642

z1(1) = 0,687*X(1) + 0,640*X(4) - 1,078*X(1)*X(4)
z1(2) = 0,687*X(1) + 0,657*X(3) - 1,020*X(1)*X(3)
z1(3) = 0,791*X(1) + 0,764*X(2) - 1,305*X(1)*X(2)
z1(4) = 0,715*X(2) + 0,855*X(3) - 1,360*X(2)*X(3)

       2 row results
              Error      Coef-t-1     Coef-t-2     Coef-t-3
 3( 1- 3) =     14.452810     1.618656     1.360954    -9.061547
 4( 1- 4) =     14.871161     1.020455     1.879339    -8.893387
 7( 3- 4) =     15.046972     1.463690     0.953528    -3.064651
 6( 2- 4) =     15.327753     1.094837     1.005669    -2.758885
 5( 2- 3) =     16.314729     1.077392     1.105448    -2.506673
 1( 1- 2) =     16.442745     1.000000     0.000000     0.000000
 2( 1- 2) =     16.571151     1.459143     1.204295    -4.596856

z2(1) = 1,619*Z1(1) + 1,361*Z1(3) - 9,062*Z1(1)*Z1(3)
z2(2) = 1,020*Z1(1) + 1,879*Z1(4) - 8,893*Z1(1)*Z1(4)
z2(3) = 1,464*Z1(3) + 0,954*Z1(4) - 3,065*Z1(3)*Z1(4)
z2(4) = 1,095*Z1(2) + 1,006*Z1(4) - 2,759*Z1(2)*Z1(4)

       3 row results
              Error      Coef-t-1     Coef-t-2     Coef-t-3
 7( 3- 4) =     11.940320     0.909004     0.998533    -1.489588
 6( 2- 4) =     12.143334     0.783708     0.993105    -1.172293
 4( 1- 4) =     12.239301     0.735509     1.109891    -1.440375
 5( 2- 3) =     12.419342     1.026507     1.106805    -2.174982
 2( 1- 2) =     12.480062     1.197777     1.237831    -3.162864
 3( 1- 3) =     12.616615     0.925663     1.116772    -2.057170
 1( 1- 2) =     14.452810     1.000000     0.000000     0.000000

z3(1) = 0,909*Z2(3) + 0,999*Z2(4) - 1,490*Z2(3)*Z2(4)
z3(2) = 0,784*Z2(2) + 0,993*Z2(4) - 1,172*Z2(2)*Z2(4)
z3(3) = 0,736*Z2(1) + 1,110*Z2(4) - 1,440*Z2(1)*Z2(4)
z3(4) = 1,027*Z2(2) + 1,107*Z2(3) - 2,175*Z2(2)*Z2(3)

       4 row results
              Error      Coef-t-1     Coef-t-2     Coef-t-3
 1( 1- 2) =     11.940320     1.000000     0.000000     0.000000
 4( 1- 4) =     12.802728     1.009403     1.719933    -4.853063
 3( 1- 3) =     12.880619     1.218947     1.385676    -4.595693
 6( 2- 4) =     13.302587     1.158373     1.204906    -3.611096
 7( 3- 4) =     14.603084     0.890151     0.937029    -1.608685
 5( 2- 3) =     15.269376     0.872656     1.045008    -2.020681
 2( 1- 2) =     15.358824     1.041783     1.226586    -3.371989

z4(1) = 0,909*Z3(3) + 0,999*Z3(4) - 1,490*Z3(3)*Z3(4)
z4(2) = 1,000*Z3(1) + 0,000*Z3(2) + 0,000*Z3(1)*Z3(2)
z4(3) = 0,784*Z3(2) + 0,993*Z3(4) - 1,172*Z3(2)*Z3(4)
z4(4) = 0,736*Z3(1) + 1,110*Z3(4) - 1,440*Z3(1)*Z3(4)

      5 row results
              Error     Coef-t-1     Coef-t-2     Coef-t-3
  1( 1- 2) =     11.940320     1.000000     0.000000     0.000000
  4( 1- 4) =     14.851743     1.078460     1.163604    -3.101076
  2( 1- 2) =     14.876404     1.179823     1.011738    -2.886278
  3( 1- 3) =     14.884729     1.115208     1.117868    -3.023926
  6( 2- 4) =     15.421893     1.077140     1.047707    -2.720931
  5( 2- 3) =     15.452147     1.139382     0.944632    -2.586491
  7( 3- 4) =     16.021664     1.080676     0.910812    -2.341144

z5(1) = 0,909*Z4(3) + 0,999*Z4(4) - 1,490*Z4(3)*Z4(4)
z5(2) = 1,000*Z4(1) + 0,000*Z4(2) + 0,000*Z4(1)*Z4(2)
z5(3) = 1,000*Z4(1) + 0,000*Z4(2) + 0,000*Z4(1)*Z4(2)
z5(4) = 0,784*Z4(2) + 0,993*Z4(4) - 1,172*Z4(2)*Z4(4)

             Row    Error     Coef-t-1     Coef-t-2     Coef-t-3
7( 3- 4) =  3    11.940320     0.909004     0.998533    -1.489588
1( 1- 2) =  4    11.940320     1.000000     0.000000     0.000000
1( 1- 2) =  5    11.940320     1.000000     0.000000     0.000000
6( 2- 4) =  3    12.143334     0.783708     0.993105    -1.172293

Y(1) = 0,909*Z2(3) + 0,999*Z2(4) - 1,490*Z2(3)*Z2(4)
Y(2) = 1,000*Z3(1) + 0,000*Z3(2) + 0,000*Z3(1)*Z3(2)
Y(3) = 1,000*Z4(1) + 0,000*Z4(2) + 0,000*Z4(1)*Z4(2)
Y(4) = 0,784*Z2(2) + 0,993*Z2(4) - 1,172*Z2(2)*Z2(4)
End of breeding
```

At each iteration step, 4 optimal candidate models were selected. The best model for predicting biomass growth was obtained on the 3rd row of selection and is based on 3 initial parameters from 4. The parameter X4 – Temperature (Temperature) is excluded.

As you can see the optimal result obtained in the third row of selection:

$$Y(1) = 0{,}909 * Z2(3) + 0{,}999 * Z2(4) - 1{,}490 * Z2(3) * Z2(4). \qquad (9)$$

The analysis showed that the equation describing the dependence of biomass growth on evaporation of soil moisture, photosynthetic active radiation, precipitation, and temperature has a non-linear form. At the same time, air temperature does not significantly affect the process under study.

We rewrite equation (9) in the variables Y, X1, X2, X3, X4:

$$
\begin{aligned}
y = {} & 1.804x_1 + 2.355x_2 + 2.319x_3 - 7.927x_1x_2 - 8.257x_1x_3 - \\
& 10.77x_2x_3 - 1.298x_1^2 - 3.452x_2^2 - 3.469x_3^2 + 38.77x_1x_2x_3 + \\
& 12.017x_1x_2^2 + 13.283x_1x_3^2 + 17.624x_2x_3^2 + 6.423x_1^2x_2 + \\
& 7.043x_1^2x_3 + 17.51x_2^2x_3 - 1.056x_1^2x_2^2 - 12.601x_1^2x_3^2 - \\
& 31.067x_2^2x_3^2 - 51.256x_1x_2^2x_3 - 66.02x_1x_2x_3^2 - 34.498x_1^2x_2x_3 + \\
& 37.15x_1^2x_2x_3^2 + 111.269x_1x_2^2x_3^2 + 56.245x_1^2x_2^2x_3 - \\
& 99.157x_1^2x_2^2x_3^2 + 1.794x_2^3 + 1.886x_3^3 - 7.714x_1x_3^3 - \\
& 10.624x_2x_3^3 - 10.052x_2^3x_3 - 6.443x_1x_2^3 + 39.674x_1x_2x_3^3 - \\
& 11.814x_1^2x_2x_3^3 + 7.428x_1^2x_3^3 + 19.486x_2^2x_3^3 - 67.499x_1x_2^2x_3^3 + \\
& 18.784x_2^3x_3^2 - 63.404x_1x_2^3x_3^2 + 23.597x_1x_2^3x_3 + 38.159x_1x_2^3x_3^3 - \\
& 11.707x_2^3x_3^3 + 5.772x_1^2x_2^3 + 53.491x_1^2x_2^3x_3^2 - 30.527x_1^2x_2^3x_3 + \\
& 57.817x_1^2x_2^2x_3^3 - 31.02x_1^2x_2^3x_3^3
\end{aligned}
\tag{10}
$$

As can be seen from formula (10), the relationship between the dynamics of the biomass plant Miscanthus (Biomass) and the following parameters: 1 is evaporation of soil moisture, 2 is photosynthetic active radiation, 3 - precipitation is described by an 8th degree polynomial.

Analyzing it should be noted that the predictors that have the greatest influence on the biomass on the contaminated soil by the plant depend on the evaporation of soil moisture, photosynthetic active radiation and precipitation.

Conclusion

This article developed a construction program for constructing regression equations based on experimental data processing, which is used to derive equations of a mathematical model of the influence of various environmental factors on the biomass plant on the contaminated soil with toxic elements.

In the method of self-organization, the construction of mathematical models from experimental data is carried out automatically. In the process of modeling, the task is to identify the form of dependence and the choice of factors that have a significant impact on the dependent variable. Moreover, the structure of the model is not fixed in advance but is selected from a variety of options according to specified criteria.

The fundamental difference between the method of self-organization from regression analysis is that its goal is to achieve the minimum of the selection criterion that is appropriately chosen, and the goal of the regression analysis is to achieve the minimum mean-square error at all extreme points for a given type of regression equation. From here various results follow. The accuracy of the self-organization method implemented in the article, in contrast to the regression analysis, is higher due to the optimization of the complexity of the model. Self-organization algorithms are used to solve the problems of pattern recognition, predicting random processes, identifying multi-extreme static and dynamic characteristics and optimal control of complex objects.

The method of self-organization has great prospects in contrast to regression analysis and artificial neural networks and therefore is widely used in the world. The practical value of the obtained results lies in the creation of a set of tools applicable for the construction of operational decision support systems based on the trend analysis of time series. The task of analyzing and forecasting time series is in demand in many developing areas, such as: data mining, analysis of the relationship between economic data, forecasting of environmental processes, etc.

Using a self-organization program, we estimated the dependence of plant biomass growing on soil contaminated with heavy metals on temperature, acidity of soil moisture evaporation, photosynthetic active radiation, and precipitation, taking into account the approximation confidence value. Mathematical models of plant biomass are obtained depending on environmental factors. The analysis showed that the predictors that have the greatest effect on the biomass of plants growing on a medium contaminated with heavy metals are evaporation of soil moisture, photosynthetic active radiation and precipitation.

The data obtained are relevant for predicting the processes of cleaning technogenic contaminated soil using plants. Models can be used to make a quantitative forecast when deciding how to clean contaminated soils with plants.

Acknowledgements

The work was carried out at the expense of grant funding for scientific research for 2018-2020 under the project AR05132044 “Development of a hardware-medical complex for assessing the psychophysiological parameters of a person”.

Funding

The authors received no direct funding for this research.

References

[1] Kane M., Brewer J. An information technology emphasis in biomedical informatics education. J Biomed Inform. 2007. 40. 67–72.

[2] Patel S., Sayyed I. U Impact of information technology in agriculture sector. JFAV. 2014. 4(2). 17–22.

[3] Zhu H., Research on Agricultural Ecological Factors Information Technology based on Internet +, Advances in Computer Science Research. 3rd International Workshop on Materials Engineering and Computer Sciences (IWMECS 2018), 2018. 78. 19–22.

[4] Rodhain F., Fallery B. ICT and Ecology: In favour of research based on the Responsibility principle. AIS Electronic Library (AISeL). Mediterranean Conference of Information Systems, 2009. Athens, Greece. 173–186.

[5] Khomyakov D. M., Iskandaryan R. A. Information technology and mathematical modeling in environmental management when implementing the concept of sustainable development, In the book: Ecological and socio-economic aspects of Russia's development in the context of global changes in the natural environment and climate. M.: Geos. 1997. 102–119.

[6] Eryomin A.L. Information Ecology – a Viewpoint. Int. J. Environ. Sci., Sections A & B. 1998. 3(4). 241–253.

[7]Panin M. S. The influence of technogenic factors and human agrochemical activity on the content of heavy metal migration in the soil-plant system. State and rational use of soils of the Republic of Kazakhstan,1998. 76–79.

[8] Kabata-Pendias A., Pendias H. Trace elements in soils and plants, M.: Mir, 1989. 439 p.

[9] Damalas C. A. Understanding benefits and risks of pesticide use, Science Research Essay, 2009. 4 (10). 945–949.

[10] Issimov N., Mazakov T., Mamyrbayev O., Ziyatbekova G. Application of fuzzy and interval analysis to the study of the prediction and control model of the epidemiologic situation, Journal of Theoretical and Applied Information Technology, 2018. 96(14). 4358–4368.

[11] Mamyrbayev O., Alimhan K., Zhumazhanov B., Turdalykyzy T., Gusmanova F. End-to-End Speech Recognition in Agglutinative Languages, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020.

[12] Ivakhnenko A. G., Zaichenko Yu. P., Dimitrov V. D. Decision making on the basis of

self-organization. M.: Soviet Radio, 1976. 280 p.

[13] Ivakhnenko V. I., Labkovsky V. A. The problem of uncertainty in decision making tasks, Kiev: Naukova Dumka, 1990, 132 p.

[14] Seber J. Linear Regression Analysis, M.: Mir, 1980, 456 p.

[15] Belov V. V., Chistyakova V. I. Modeling and forecasting of business processes using self-organization algorithms of formal descriptions, Business Informatics, 2008. 4 (06). 37–48.

[16] Ayvazyan S. A. and others. Classification and reduction of dimension, M.: Finance and Statistics, 1989, 607 p.

[17] Ilyin V. A., Sadovnichy V. A., Sendov Bl. Kh. Mathematical analysis. Beginner course. M.: Moscow State University, 1985. 662 p.

[18] Fichtenholtz G. M. Fundamentals of mathematical analysis. M.: Nauka, 1968. 1. 440 p.

[19] Kalitkin N. N. Numerical methods, M.: Nauka, 1978. 512 p.

[20] Bakhvalov N. S. Numerical methods. M.: Science, 1975. 1. 632 p.

[21] Tyukachev N. A., Rybak K. S., Mikhailova E. E. Programming in Delphi for beginners, SPb: BHV-Petersburg, 2007. 672 p.

[22] Bucknell J. M. Fundamental algorithms and data structures in Delphi, St. Petersburg: DiaSoftUP LLC, 2003. 560 p.

IRSTI 81.93.29

UDC 004.056

AUTOMATED WATER LEVEL MONITORING SYSTEM IN WATER BODIES

T. Mazakov[1], P. Kisala[2], G. Ziyatbekova[3], and M. Aliaskar[3]
Institute of Information and Computational Technologies,Almaty, Kazakhstan
Al-Farabi Kazakh National University, Almaty, Kazakhstan
[1]tmazakov@mail.ru, [2]p.kisala@pollub.pl , [3]ziyatbekova@mail.ru, [3]87019931011@mail.ru
[1]ORCID ID: https://orcid.org/0000-0001-9345-5167
[3]ORCID ID: https://orcid.org/0000-0002-9290-6074

Abstract. The article is devoted to the creation of an automated system for monitoring the water level in reservoirs to prevent the breakthrough of weirs and dams. The paper offers hardware and software for monitoring the reservoir occupancy with prompt notification of interested organizations (local administrations) and local emergency departments.

The article describes a developed automated system for monitoring the water level in a reservoir, which allows to get real-time information about the relative humidity and air temperature, the distance from the dam crest to the water surface in the reservoir. Based on the information received, the system allows to estimate the forecast time of increasing the volume of water level from the current to the critical level and inform the population about the state of the reservoir.

The general characteristic of the problem and the formulation of the research objectives are given. Based on microprocessor technology and sensor sensors, an autonomous microcomputer climate data transmission system has been developed.

Keywords: flood, dam, water level monitoring, microprocessor system, water level sensor, Raspberry microcomputer, Arduino UNO platform.

Introduction

It is necessary to analyze large volumes of heterogeneous information, inconsistency of goals of various state bodies to assess the environmental safety of the region. The solution of such tasks is impossible without the use of modern information systems for decision support.

Currently, there are 1665 hydraulic structures on the territory of the Republic, including reservoirs with volume greater than 1.0 m^3 – 319 units (including in the Republican ownership – 83, in municipal ownership – 200 in private ownership – 34 and ownerless – 60); dams – 443 units (including in the Republican ownership – 32, in municipal ownership – 346, private ownership – 45 and ownerless – 20); dams – 125 units and other hydraulic structures – 778 units.

As of May 1, 2017, a total of 1212 hydraulic structures were examined, 865 hydraulic structures of them are in satisfactory condition, and 347 hydraulic structures are in unsatisfactory condition and require repair.

The necessity of legal regulation of the safety issues of hydraulic structures is determined by the large-scaled social and economic consequences of their damage and destruction. At the same time, human losses and material damage are comparable to the consequences of devastating natural disasters.

In Kazakhstan the construction of many hydraulic structures was carried out in the 60-80s of the last century [1]. Their survey today shows that the actual depreciation is more than 60%, the reliability and safety of strategically important hydraulic structures are sharply reduced.

In accordance with the Water code, Presidential decree of the Republic of Kazakhstan dated November 1, 2004, no. 1466, a list of water facilities (hereinafter referred to as the List) of particular strategic importance was defined, which includes 57 reservoirs and 29 retaining hydraulic structures. In accordance with Article 25 of the Water Code, these water facilities cannot be leased, trust and cannot be privatized.

The long service life and reduction in the last 20 years of funding for operating expenses, current and capital repairs, as well as the influence of climatic and seismic factors gradually lead to moral and physical deterioration of the entire complex of hydraulic structures. There are also objects located close to hazardous industries.

A dam is a blocking river (or other watercourse) for raising the water level in front of it, concentrating the pressure at the location of the structure and creating a reservoir. The dam and reservoir significantly affect the river and adjacent territories: the regime of river flow, water temperature, and the duration of freezing change; migration of fish is difficult; the banks of the river in the upstream are flooded; the microclimate of coastal territories is changing.

The damage caused by natural disasters is, in particular, slightly worn, while the loss of life is slightly higher. To select a set of measures to minimize the damage, it is advisable to carry out a forecast of the main characteristics of floods affecting the magnitude of the damage. Their size influences the severity of the consequences of floods for population, economy, agriculture, etc.

At the present time, it is possible to have many times of flooding, flooding and erosion of the land caused by such an extraordinary accident, like breaking the plate.

Emergencies that arise as a result of the destruction of pressure front facilities and are characterized by a major damaging factor: a breakthrough wave and, accordingly, a catastrophic flooding of the area are often accompanied by secondary damaging factors:

- fires: due to breaks and short circuits of electric cables and wires;
- landslides, landslides: due to erosion of the soil;
- infectious diseases: due to contamination of drinking water, food, etc.

The causes of accidents, accompanied by a breakthrough of hydraulic structures of the pressure front and the formation of a breakthrough wave, can be different. Most often, such accidents occur due to the destruction of the foundation of the structure and the lack of spillways. The percentage ratio of their various causes is shown in table 1.

Table 1 – Frequency of various causes of accidents in hydraulic structures, accompanied by the formation of a breakthrough wave.

Reason for destruction	**Frequency %**
Destruction of the foundation	40
Spillway Insufficiency	23
Design flaws	12
Uneven draft	10
High threshold (capillary) pressure in the washed dam	5
Military actions	3
Creep slopes	2
Material defects	2
Earthquakes	1
Improper operation	2
TOTAL:	100

One of the main reasons leading to accidents at hydrotherapy consoles is as natural, and so are the factors:

– if you are in extreme use, it may be possible to overfill the water and to avoid regular waste of work, which will result in a loss of water;

– due to the long service life, it is possible to wear out the main weir and hydraulic equipment, which may result in loss of life;

– due to a malfunction of the personnel associated with the lack of monitoring of dangerous situations and the inadequacy of the information provided on the product;

– consequence of the possible terrorist act leading to the destruction of the dam.

Today, such large reservoirs are operated as the Astana reservoir built in 1970 with a capacity of 410.9 million cubic meters, the Seletinsky reservoir – 1965 (230 million cubic meters), the

Kargalinsky reservoir – 1975 (280 million cubic meters), the Bartogaysky reservoir – 1982 (320 million cubic meters), the Kapshagai reservoir – 1970 (18560 million cubic meters), the Ters–Ashibulak reservoir – 1963 (158,6 million cubic meters), the Tasotkelsky reservoir – 1974 (620 million cubic meters), the Samarkandsky – 1939 (253,7 million cubic meters), the Upper Tobol – 1972 (816,6 million cubic meters), the Karatomarskoy – 1965 (586 Bugunskoybuilt in 1967 (370 million cubic meters) and others.

Monitoring systems should ensure constant monitoring of phenomena and processes occurring in nature and the technosphere, in order to anticipate increasing threats to humans and their environment. The main purpose of monitoring is to provide data for an accurate and reliable forecast of emergencies based on the combination of intellectual, informational and technological capabilities of various departments and organizations involved in monitoring certain types of hazards. Monitoring information serves as the basis for forecasting.

Microprocessor technology has now actively entered our lives. Versatility, flexibility, simplicity of hardware design, almost unlimited possibilities for complicating information processing algorithms – all this promises a great future for microprocessor technology. Microprocessors are used both in household appliances for the simplest signal processing and command generation, as well as in the most complex measuring systems for digital signal processing.

Modern opportunities for the development of various sensors [2, 3] and the cheapening of microprocessors have also opened up a wide opportunity to implement hardware-software tools for monitoring climate parameters.

In particular, the relatively cheap Arduino controller, which has a large database of developed sensors and their means of communication with a computer, has found wide application in applied problems [4, 5].

In this regard, the research in this work on the development and research of a mathematical model of a dam breakthrough and information security tools is relevant.

Implementation

The following system is proposed for monitoring the threat of a breakthrough of hydroelectric facilities, consisting of two blocks:

1) block for receiving and transmitting current information about water level, humidity and temperature on the dam crest;

2) block for processing constant and operational information about the threat of a dam break (server).

There are two options for connecting blocks.

In the first case, the Arduino microprocessor is directly connected to the server. This option requires a permanent power supply system and the presence of processing personnel at the waterworks.

In the second case, the Arduino microprocessor is connected to the Raspberry Pi microcomputer, which transmits current information to the server via satellite communication. This option does not require the constant presence of processing personnel at the waterworks. And due to its small size and low power consumption, it can be provided with small-sized solar energy.

Block for receiving and transmitting current information

The block for receiving and transmitting current information is implemented in the form of the water level sensors, humidity and temperature and is located on the crest of the dam. The sensors are connected to an Arduino microprocessor, which provides pre-processing of data received from the sensors and transmits them for further processing.

To create an autonomous microprocessor system of transmitting climate data, we used a single-board Raspberry Pi 3 B+ microcomputer [6, 7]. Power is provided by a solar panel.

The system includes a set of necessary sensors and software. онThe measurement modules are connected to the computer via a USB adapter.онThe software presents the results of measurements in tabular and graphical form, and also allows to view and print the archive of measurements accumulated in the database for any period of time. онIt is possible to view data from sensors both on other computers of the local network and via the Internet.

The Arduino is a device based on the ATmega microcontroller 328 [8-10]. It includes everything necessary for convenient work with the microcontroller. To start working with the device, simply supply power from an AC/DC adapter or battery, or connect it to a computer using a USB cable.

The Raspberry Pi is a single-board computer with the size of a bank card, that is, the various parts of the computer that are usually located on separate boards are presented here on one. Raspberry Pi runs mainly on Linux and Windows operating systems.

The PC-oriented experiment is configured using ISES relay modules (1) for the control pump and sensors (2) for measuring the water level (Figure 1). Modules work using the ISES panel (3) and the server. Based on the above blocks, an autonomous microprocessor data transmission system is implemented (Figure 2).

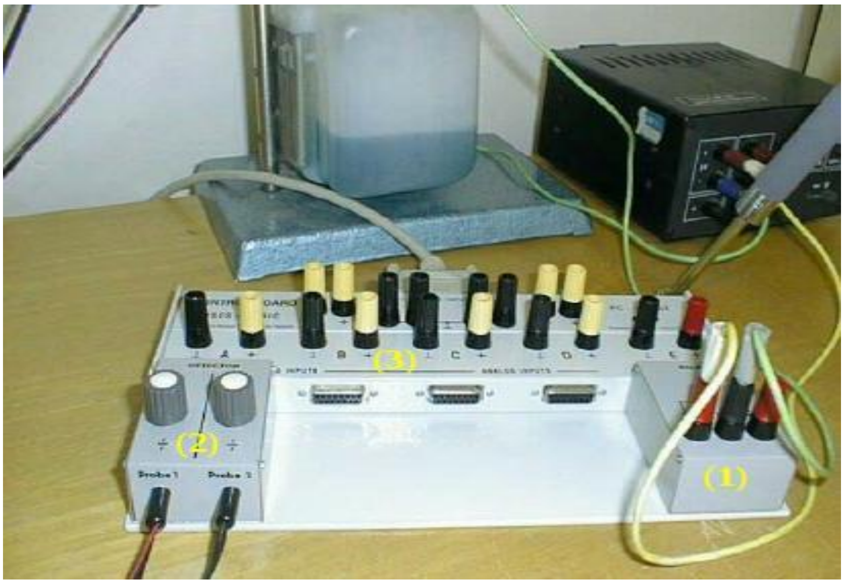


Figure 1 – Location of the control remote experiment controlling the water level Relays (1), Sensors (2) and ISES Panel (3)

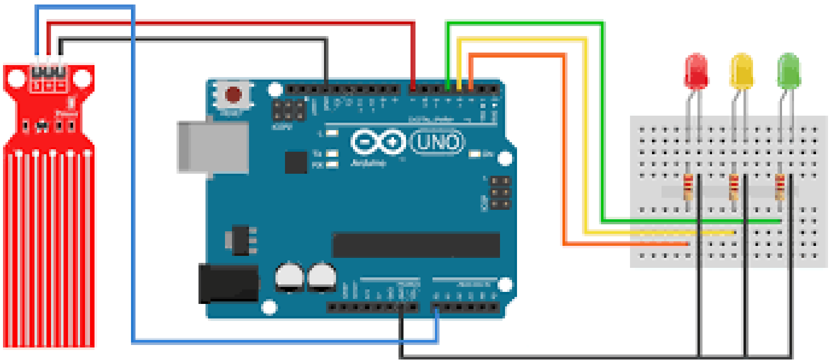


Figure 2 – Water level sensor and its interaction with Arduino

The processing block of constant and operational information

The 2nd block contains constant information about the characteristics of the reservoir and dam, and also quickly receives current information. Based on the processing by which it calculates the level of safety, anxiety or disaster of the hydraulic unit. In the latter case, it automatically informs state authorities (emergencies, akimats, etc.) about the possible threat of a dam break.

Due to the specifics of the studied hydrological processes [11-13], fuzzy and interval mathematics are used in the work [14-16].

To assess the threat of a breakthrough, the mathematical model is proposed with the following interval linguistic variables:

1) low level;
2) safe level,
3) alarming level, and
4) catastrophic level of reservoir occupancy [17-18].

The values of the entered linguistic variables are predefined in the following percentages of the dam height:

1) low level – 40%;
2) safe level – 30%,
3) alarming level – 20%;
4) catastrophic level – 10%.

For underground dams, the catastrophic level decreases by 3%. In the presence of precipitation for underground hydraulic structures, the catastrophic rate from the reference book "Name of dams or weirs” is reduced by another 2%, to take into account the possibility of dam weakening due to external precipitation.

On the night of March 10, the water level reached 30 million cubic meters. The next day, in the afternoon or in the evening, exact time unknown, the water level exceeded 40 million cubic meters. In other words, 15-16 million cubic meters of water was added to the Kyzylagash reservoir in 15-16 hours. The dam broke on March 11 at 10.30 p.m. Two hours later, the water gushed towards the village of Kyzylagash. The wave width of the mudflow was 1.6 kilometers, and the height was about 3-4 meters. According to official data, most of the village was severely damaged. 70% of the village of Kyzylagash was destroyed. The tragedy in Kyzylagash claimed the lives of 44 people.

Conclusion

This article analyzes the characteristics of dams, the capabilities of modern control systems based on the use of microprocessor technology.

The tragic occurrences of spring 2010 in the Almaty region and 2014 in the Karaganda region with human casualties and destruction, as well as floods in other regions of Kazakhstan, served as a serious lesson to prevent similar situations in the future. It is necessary to develop recommendations for equipping hydraulic structures with modern control and measuring devices, equipment and means to improve operational safety.

Acknowledgements. The work was carried out at the expense of grant funding of scientific research for 2018–2020 under the project AP05131027 "Development of biometric methods and information protection tools".

Funding

The authors received no direct funding for this research.

References

[1] Chokin Sh. Ch., Baishev B. B., Grigoriev V. A. Calculations of multi-purpose reservoirs. Alma-Ata: Nauka Kaz SSR, 1983. 208 p.

[2] Platt Ch. Electronics: logic circuits, amplifiers and sensors for beginners. SPb.: BHV-

Petersburg, 2015. 448 p.

[3] Igo T. Arduino, sensors and networks for communication devices. SPb.: BHV-Petersburg, 2015. 544 p.

[4] Boxall J. We study Arduino. 65 do-it-yourself projects. SPb.: Piter, 2017. 400 p.

[5] Petin V.A. Projects using the Arduino controller. - SPb.: BHV-Petersburg, 2016. 464 p.

[6] Karvinen T., Karvinen K., Valtokari V. We make sensors: projects of sensor devices based on Arduino and Raspberry Pi. M.: LLC “I.D. Williams", 2017. 432 p.

[7] Petin V.A. Arduino and Raspberry Pi in Internet of Things projects. SPb.: BHV-Petersburg, 2017. 320 p.

[8] Belov A. V. Arduino: from the basics of programming to creating practical devices. SPb.: Science and technology, 2018. 480 p.

[9] Simon Monk. Makering. Arduino and Raspberry Pi. Motion control, light and sound. SPb.: BHV-Petersburg, 2017. 336 p.

[10] Yazenkov V.S. From Arduino to Omega: platforms for makers step by step. SPb.: BHV-Petersburg, 2018. 304 p.

[11] Gelfan A.N. Dynamic stochastic modeling of melt runoff formation. M.: Nauka, 2007. 279 p.

[12] Vinogradov Yu. B. Mathematical modeling of runoff formation processes. L.: Hydrometeoizdat, 1988. 312 p.

[13] Alexandrov D.V., Zubarev A.Yu., Iskakova L.Yu. Applied hydrodynamics. URIGHT, 2018. 110 p.

[14] Kalmykov S.A., Shokin Yu.I., Yuldashev Z.Kh. Methods of Interval Analysis. Novosibirsk: Nauka, 1986. 224 p.

[15] Zadeh L. The concept of a linguistic variable and its application to making approximate decisions. M.: Mir, 1976. 167 p.

[16] Dzhomartova Sh.A. Practical interval computing, Bulletin of the NAS RK. 2002. 2. 41–46.

[17] Issimov N., Mazakov T., Mamyrbayev O., Ziyatbekova G. Application of fuzzy and interval analysis to the study of the prediction and control model of the epidemiologic situation. Journal of Theoretical and Applied Information Technology. 2018. 96(14). 4358–4368.

[18] Dzhomartova Sh.A., Mazakov T.Zh., Isimov N.T. Mazakova A.T. Real-time forecasting program. Bulletin of the National Engineering Academy of the Republic of Kazakhstan. 2017. 4(66). 27-32.

IRSTI 20.23.17; 20.23.21
UDC 004.421, 004.912

THE DECISION AMBIGUITY PROBLEMS FOR THE KAZAKH LANGUAGE

D. Rakhimova[1], Waldemar Wójcik[2], A. Karibayeva[3]
Al-Farabi Kazakh National University, av. 71, Almaty, Kazakhstan
Lublin Technical University, Poland
[1]di.diva@mail.ru, [2]waldemar.wojcik@pollub.pl, [3]a.s.karibayeva@gmail.com
[1]ORCHID ID: 0000-0003-1427-198X

Abstract. For the Turkic languages, including the Kazakh language, there are many applications that are not allowed. Recently, the study of the problems of ambiguity in natural language processing is relevant. This problem occurs in various intelligent information systems, such as information search, machine translation, text and speech analysis, etc. There are various approaches to solving this problem. But for the application of the Kazakh language, taking into account its features, there were difficulties. This article discusses the problems of ambiguity of the word for the Kazakh language. Based on the proposed technology, the process of solving the problem of ambiguity for the Kazakh language in the machine translation system for Kazakh-English and Kazakh - Russian pairs (and Vice versa) is described. The proposed technology includes a restriction grammar model and a maximum entropy model for a more effective solution to the problem of lexical selection for the Kazakh language. The results are presented by comparing the two technologies separately and together.

Keywords: ambiguity, technology, Kazakh language.

Introduction

Automatic detection of the correct translation of words that depend on the context is a very difficult task. The solution of lexical polysemy is perceived as the main task, the solution of which will allow achieving an almost perfect machine translation. Machine translation has two main problems in text processing. First of it is lexical selection, which connected with problem of choosing corresponded translation by context. Second problem of a lexical words is order. Later consider the task of words order in sentence of target language. In word processing the main problem is lexical selection, which leads to disambiguation task.

The work of the translator's program is carried out in several stages. The development of algorithms that allow us to recreate the human ability to understand and choose the right meaning of a word is a difficult task. It consists in choosing the most suitable translation in the context under consideration from all possible translations. The solution of the problem involves finding the probable meanings of words, determining the relationships between these values and the context in which the words were used.

The scientific novelty of paper is in developing combined technology of lexical polysemy (selection) based on the constraint grammar model and maximum entropy model and them sequence applying to Kazakh language as source and target language in translation pairs Kazakh-English and Kazakh-Russian (and vice versa).

Related works

Scientific research on a word sense disambiguation has long-term history. With a current of years the number of the proposed solutions and their efficiency steadily grew, but the task hasn't received the full decision yet. A large number of methods have been investigated: from the methods based on knowledge, rules, lexicographic sources, training with the teacher at corpus to training methods without teacher, clustering words on the basis of meaning. Among listed, today, training methods with the teacher have shown the best efficiency.

The problem of resolving ambiguity as a separate problem was formulated in the middle of 20th century, almost simultaneously with the advent of machine translation. Since that time, many methods of solving this problem have been developed, but it is still actual.

M. Tyers, Felipe S´anchez-Mart´ınez, Mikel L. Forcada in [1] uses the maximum entropy model for performing lexical selection in machine-based translation systems based on rules to English-Basque, English-Catalan. As a learning method, the method of teaching without the teacher (unsupervised methods) is used, which does not require an annotated corpus. The system uses the maximum entropy formalism for lexical selection as in Berger [2] and Marecˇek (2010) [3], but instead of counting the actual lexical selection event in the annotated corpus, they consider fractional occurrences of these events according to the model of the target language. David Marecˇek, Martin Popel, Zdeneˇk Zˇabokrtsky in [3] showed the model of the maximum entropy of translation in machine translation based on dependencies, which allows to develop a large number of features of functions in order to obtain more accurate translations. Francis Morton Tyers in [4] the general method of teaching rules for the module is described. Monolingual and bilingual corpora can be used for the method. For learning a monolingual corpus the method without the teacher (unsupervised method) is used. Also weighting method is described, based on the principle of maximum entropy. This method allows to take into account all the rules without having to choose between conflicting and overlapping rules.

Rule-based approaches in lexical selection not cover all possible translation of source language. Since it is not always possible to take into account all rules. In the statistical approach, lexical choice also does not guarantee for the correctness of the translation in context. Analyzing both approaches, we came to the conclusion about the development of technology combining the two approaches mentioned above with solutions to lexical selection problems for the Kazakh language in Apertium systems.

Technology of lexical selection

The technology for solving the problem of lexical selection we propose consists of two models: the model of production rules (constraint grammar) and the maximum entropy models (fig. 1).

In the process of translating a lexically ambiguous word, the system first solves the problem using the model of production rules. The model of production rules gives a good result, but does not completely solve the task. Because rules are not cover all cases of ambiguity of a certain word in any context. A word can have a different senses, and can be used with different other words in a context. And there can be many such combinations. And writing the rules can take a lot of time. Therefore, for the case when there is no written rule for some case, the selection of translation is solved by using data in semantic cube based on parallel corpora processed by statistical model. So, the ambiguous word is processed using the statistical part of the module.

The models of combined technology performed sequentially. Firstly is used the model of production rules and then maximum entropy model.

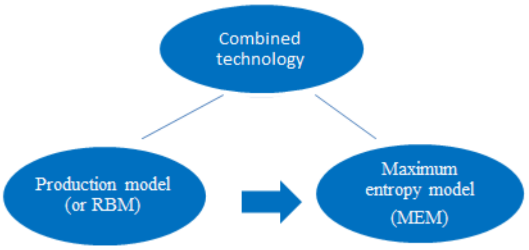


Figure 1 – The models of combined technology

The technology of combined technology was introduced in the Apertium [5]. In the aperture, divide into two parts the lexical selection and lexical choice based (fig. 2) on the semantic cube.

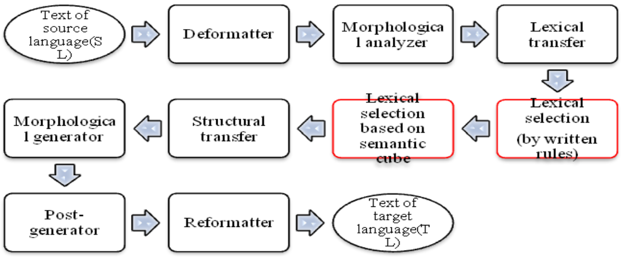

Text of source language(SL)
Deformatter
Morphological analyzer
Lexical transfer
Lexical selection (by written rules)
Lexical selection based on semantic cube
Structural transfer
Morphological generator
Post-generator
Reformatter
Text of target language(TL)

Figure 2 – The technology in Apertium work pipeline

Model of production rules

In the Apertium system, the lexical selection (polysemy) problem is solved by writing the rules manually in the lexical selection module using the Constraints Grammar, which has the form of product rules.

The model of product rules is a rule-based model, where knowledge is shown in the form "if (condition), then (action)". The product model can be represented in the following form:

$$i = \langle S, L, A \rightarrow B, T \rangle$$

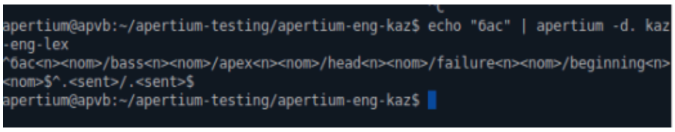
where S is a case, L is translation, $A \rightarrow B$ is productivity kernel, T is a postcondition of the production rule.

The rules of lexical selection are written in the file "apertium-eng-kaz.eng-kaz.lrx" for the English-Kazakh and in the file "apertium-eng-kaz.eng-kaz.lrx" for Kazakh-English pair of languages. All rules are written in XML format. The format of lexical rules is based on the technology of patterns. This module is used to select a translation when the meaning of ambiguous word refers to one part of the speech. For example, the word "year" is more often translated as "жыл"[zhyl], and the phrase "five years old" is translated as "бес жаста"[bes zhasta], that is, it indicates the age of a person. For this case the following rule is written:

```
<rule> <match lemma="year" tags="n.pl.*">
          <select lemma="жас" tags="n.*"/>
</match> <match lemma="old" tags="adj.*"/></rule>
```

The meaning of the rule: if after the noun "year" is an adjective "old", then the translation "жас"[zhas] is chosen.

For example word “бас”[bas] has various translation(fig. 3):



```
apertium@apvb:~/apertium-testing/apertium-eng-kaz$ echo "бас" | apertium -d. kaz
-eng-lex
^бас<n><nom>/bass<n><nom>/apex<n><nom>/head<n><nom>/failure<n><nom>/beginning<n>
<nom>$^.<sent>/.<sent>$
apertium@apvb:~/apertium-testing/apertium-eng-kaz$
```

Figure 3 – The translations of word “Бас” in bilingual dictionary of English

The figures below show lexical units to word “бас(bas)”. As you see given words have 4 translations.

```
<rule><match lemma= “адам” tags=n.*”/>
  <match lemma= “бас” tags=”n.*”>
<select lemma=”head” tags=”n.*”/></match></rule>
<rule>
  <match lemma="фильм” tags= “n.*”/>
```

<match lemma= “бас” tags= “n.*”><select lemma= “beginning” tags= “n.*”/></match></rule>

The meaning of the rule: if word “бас”[bas] comes after word “адам”[adam], the translation will be as “head”, whereas when it comes after with “фильм” [film] it translated as “beginning”

The model of production rules gives a good result, but does not completely solve the task. Because rules are not cover all cases of ambiguity of a certain word in any context. Kazakh-English language pairs in Apertium currently have 98 and 75 lexical selection rules respectively. And Kazakh-Russian language pairs in Apertium currently have 76 and 59 lexical selection rules respectively. A word can have a different senses, and can be used with different other words in a context; and there can be many such combinations. And writing the rules also depending of knowledge level of developer. The lexical selection rules are written to the various part of speech, namely to noun, verb, adjective, adverb, preposition and etc. Therefore, for the case when there is no written rule for some case, the selection of translation is solved by using statistics the maximum entropy model.

Maximum entropy model

Lexical selection maximum entropy model includes a set of binary functions and appropriate weights for each function [4]. The feature is defined as $h^s(t,c)$ in equation (1), where t is a translation and c is a source language context for each source word s:

$$h^s(t,c) = \begin{cases} 1, c\_the\_value\_of\_tunder\_the\_condition\_c \\ 0, other \end{cases} \qquad (1)$$

During the learning process each function is assigned a weight λ^s, and combining the weights as in Equation (2) gives the probability of a translation t for word s in context c.

$$p_s(t \mid c) = \frac{1}{Z} \exp \sum_{k=1}^{n_F} \lambda_k^s h_k^s(t,c) \qquad (2)$$

where Z is a normalizing constant, n_F – numbers of features for s. Thus, the most probable translation can be found using equation (3)

$$\hat{t} = \arg\max\ p_s(t \mid c) = \arg\max \sum_{k=1}^{n_F} \lambda_k^s h_k^s(t,c) \qquad (3)$$

where $t \in T_s$, T_s are all possible translations for s source word, λ is weight of known translation.

It is important to note that the rules for the function $h^s(t,c)$ will be different depending on the language pair. Consider an example: 'Mark is a bass player. He fried the bass'. In these sentences the word bass is ambiguous. In the first sentence, the sentence is translated into Kazakh as a "bass guitar", and in the second as a type of fish "алабұға" [alabuga]. Then the function has the form

$$h^s(t,c) = \begin{cases} 1, if\_t = 'бас - гитара'\_and\_'player'\_after\_bass \\ 0, in\_other\_clases \end{cases}$$

In combined technology the maximum entropy model is realized through the construction of a semantic cube.

Algorithms of semantic cube

The algorithm for implementation consists of the following steps:

Step 1. Create a frequency list of words.

First, to build a cube, we create a frequency list and a list of ambiguous words. The frequency list is a list of the most often met words in the corpus.

Step 2. Create a list of ambiguous words.

After composing the frequency list, it is necessary to find among them ambiguous ones, namely, the lexical ambiguity is taken into account, that is, when possible translations of the required word belong to one part of the speech (tab. 1).

Table 1. Example of ambiguous words.

Ambiguous words	Part of speech	Sense 1	Sense 2	Sense 3
String	Noun	жол	Жіп	ішек
Order	Noun	рет	жарлық	орден
Part	Noun	бөлік	партия	дене
Small	Adjective	кішкентай	Ұсақ	шағын
Thing	Noun	зат	Нәрсе	дүние
Discover	Verb	байқау	Ашу	табу
Information	Noun	ақпарат	Хабар	мәлімет

Step 3. Prepare bilingual parallel corpus.

Next, we prepare a bilingual corpus. Each statistic machine translation consist the greater number of text array. These arrays named as corpora. The word "prepare" means that "unnecessary" words or stop words are deleted, which have a special effect in calculating probabilities, as well as choosing the sense of the word. Such words, for example, articles like the, a, an, numbers, punctuation marks, etc. It is also necessary to bring all the words from the context to the basics, that is, to the initial form of the word. This completes the preparation of the corpus.

Step 4. "Training" / "Machine learning" stage. Semantic cube building.

The result of applying the maximum entropy model to solving the problem of lexical selection is the construction of a "semantic cube". At first the statistical system passes a stage of "training" at which statistical data on the translation of separate words and phrases from source language on target language are taken. The maximum entropy model is trained on a pre-prepared parallel bilingual corpus [8]. On the basis of this stage the cube is formed. The cube represents a three-dimensional set of tables. The tables contain the senses of ambiguous words, the context, the meanings of the probabilities of the senses (translations) of ambiguous words that depend on words from the context.

There two variables can be denote as a linear releationship:

$$amb_{word} : t_i \rightarrow f_i \in C$$

Choosing or finding word’s translation basically depends on two main variables - translation and context.

The semantic cube model determines the translation by weight of word’s context in parallel corpora. The model gives information about frequency and probability to determined words and his neighbourhood words. The frequency of a translation is determined by the number of occurrences of the features, and when the known features are repeated, the number increases to 1.

The probability of translation calculated by next formula (4):

$$P_{amb\_word}\left(v_{f_{ij}} \mid f_n\right) = \frac{v_{f_{ij}}}{\sum_{i=1}^{n} f_n} \qquad (4)$$

where, $v_{f_{ij}}$ is the frequence of a word with a certain translation, $\sum_{i=1}^{n} f_n$ is total amount of features in corpora for a particular translation.

The found probability gives the weight of a specific word translation. The model of semantic cube chooses the translation from the corresponding set of tables of multivalued words maximizing their frequency and probability (formula 5):

$$\tilde{t} = \arg\max_{t \in c} \sum_{i=1}^{n} P_{amb_{word}}\left(t_i \mid f_n\right) \qquad (5)$$

The advantages of this model it works with parallel corpora.

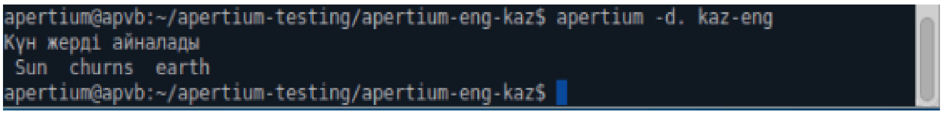
Step 5. "Testing" stage.

Next is the "testing" stage which is made on another separate text, consisted of sentences, in which there are ambiguous words. During the translation process, this system calculates the most likely translation of the source sentence based on the data obtained during training. By comparing probabilities, as a translation for ambiguous word is selected the sense which has higher probability. It should be noted that the larger the volume of the corpus, the higher the quality of translation.

There are three outputs of the result for this technology. In the first case, when the system recognizes that there is an (lexical) ambiguous word in the sentence, then the word goes into the lexical selection module. First, the ambiguity is solved by using the model of product rules. If there is a written rule in the file for the required ambiguous word, then the translation for the word is selected based on this rule. In the second case, if there is no such rule, then a transition to the next stage occurs, where the problem is solved using the maximum entropy model. In this case, the translation for the required word will be that sense, which probability is higher. In the third case, when the translation cannot be found using the product rules model, or the maximum entropy model, as the default translation selects the meaning that is first specified in the bilingual dictionary.

Compared with other works of Berger(1996)[2] and Dell Pietra(1997) [6], our model differs in that the whole sentence is used as a context. Sanchez-Martinez and Tyers (2015) [7] use monolingual corpora; we use bilingual parallel corpora and a frequency list of ambiguous words with certain senses. Tyers uses a monolingual corpus of the source language and a statistical model of the target language. In his model the volume of context for an ambiguous word is four words, two on each side. We have the context of all the words included in the sentence, with the ambiguous word.

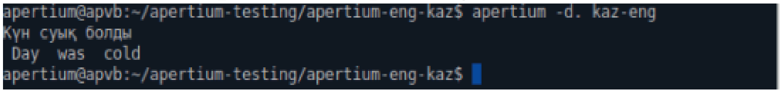
In the below figures the results of combined technology are presented (fig. 4):



```
apertium@apvb:~/apertium-testing/apertium-eng-kaz$ apertium -d. kaz-eng
Күн жерді айналады
 Sun  churns  earth
apertium@apvb:~/apertium-testing/apertium-eng-kaz$
```

Figure 4 – The results to word “күн”[kun] with combined technology

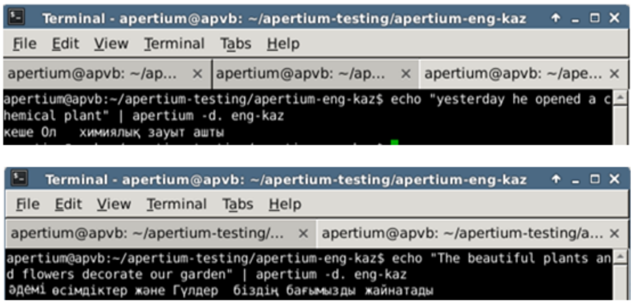
At the fig. 4 Kazakh word “күн”[kun] depending of context is translated as “sun”.



```
apertium@apvb:~/apertium-testing/apertium-eng-kaz$ apertium -d. kaz-eng
Күн суық болды
 Day  was  cold
apertium@apvb:~/apertium-testing/apertium-eng-kaz$
```

Figure 5 – The translation’s results to word “күн”[kun] with combined technology

At the fig. 5 Kazakh word “күн”[kun] depending of context is translated as “day”.


Terminal - apertium@apvb: ~/apertium-testing/apertium-eng-kaz
File
Edit
View
Terminal
Tabs
Help
apertium@apvb: ~/ap...
apertium@apvb: ~/ap...
apertium@apvb: ~/ape...

```
apertium@apvb:~/apertium-testing/apertium-eng-kaz$ echo "yesterday he opened a c
hemical plant" | apertium -d. eng-kaz
кеше Ол   химиялық зауыт ашты
```

Terminal - apertium@apvb: ~/apertium-testing/apertium-eng-kaz
File
Edit
View
Terminal
Tabs
Help
apertium@apvb: ~/apertium-testing/...
apertium@apvb: ~/apertium-testing/a...

```
apertium@apvb:~/apertium-testing/apertium-eng-kaz$ echo "The beautiful plants an
d flowers decorate our garden" | apertium -d. eng-kaz
әдемі өсімдіктер және Гүлдер  біздің бағымызды жайнатады
```

Figure 6 – The translation results for word “plant” with combined technology

In fig. 6 is given examples for English-Kazakh language pair. The word “plant” is ambiguous word can be translated into Kazakh as “зауыт”[zauyt]-“factory” and “өсімдіктер”[osimdikter]-“herb”. In both situations combined technology model chose right senses, for the first “зауыт”[zauyt] and for the second “өсімдіктер”[osimdikter].

For realization of the combined technology the rules and parallel corpora are used to determine the context. By using corpora for training and testing we distinguish context for ambiguous word. To resolving the task of lexical selection we collect and use parallel corpora [8, 9] taken from different sources such as texts in electronic form from various famous literary novels, fairy tales, open Internet resources, news portals.

We have collected the English-Kazakh parallel corpus of ~30000 sentences, the Kazakh-Russian parallel corpus of ~26 000 sentences.

Experiments results

In experiment for determining the translation were estimated in modes of checking productive rules, maximum entropy model and proposed combined technology in two direction of translation: Kazakh-English and Kazakh-Russian. Results are given in tab. 2.

Table 2. Results of experiments on corpora

Language pair	Total number of sentences	Produc	Maximum en	Combined technolog

		tive rule model, %	tropy model, %	y, %
Kazakh-English	26078	52	72	87
English-Kazakh	26078	48	65	78
Kazakh-Russian	22053	53	64	68
Russian-Kazakh	22053	62	68	70

By the results we can see that the proposed combined technology works better.

Conclusion and future work

In this paper was performed a combined technology uniting the model of productive rules and maximum entropy model for solving the problem of lexical selection for English-Kazakh and Kazakh-English language pairs. Also parallel bilingual English-Kazakh corpus has been developed with ~30000 sentences, Kazakh-Russian corpus has been developed with ~26 000 sentences. Experiments of checking productive rules, maximum entropy model and proposed combined technology in two direction of translation: Kazakh-English and Russian-Kazakh shows better results of proposed combined technology of lexical selection.

In future work we plan to increase the volume of parallel corpora for receiving more exactly results in solving of lexical selection.

References

[1] Tyers M., Felipe S´anchez-Mart´ınez, Mikel L. Forcada. Unsupervised training of maximum-entropy models for lexical selection in rule-based machine translation. Proceedings of the 18th Annual Conference of the European Association for Machine Translation, Antalya, Turkey (2015). 145-153.

[2] Berger, A., Pietra, S. D., and Pietra, V. D. A maximum entropy approach to natural language processing. Computational Linguistics, 1996. 22(1). 39–71.

[3] Marechek D., Popel M., Zabokrtsky Z. Maximum Entropy Translation Model in Dependency-Based MT Framework. Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, (2010). 201–206.

[4] Tyers F. M. Feasible lexical selection for rule-based machine translation. Ph.D. Thesis – Universitat d'Alicante. 2013. 110 p.

[5] Otkry`taya sistema mashinnogo perevoda [Open machine translation system], URL: https://www.apertium.org/ (accessed 04.15.2019). (In Russian)

[6] Francis M. T., Martínez F. S., Forcada M. L. Unsupervised training of maximum-entropy models for lexical selection in rule-based machine translation. Proceedings of the European Association for Machine Translation EAMT, (2015). 145–152.

[7] Rakhimova D., Abakan M. "Lexical selection in machine translation of russian-to-kazakh". Proceedings of the II International Conf. Computer processing of Turkic languages, Turkey, (2014). 97–102.

[8] Rakhimova D., Zhumanov Zh. Complex technology of machine translation resources extension for the Kazakh language. Studies in Computational Intelligence. Springer, 2017. 710(307). 297.

[9] Sánchez-Cartagenaa V.M., Pérez-Ortiza J.A., Sánchez-Martínez F. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. Computer Speech & Language, 2015. 32(1). 46–90.

IRSTI 20.19.21
UDC 004.912

AN ANALYTICAL STUDY OF MODERN INFORMATION EXTRACTION TECHNOLOGIES AND APPROACHES

A. Y. Nuraliyeva
Kazakh-British Technical University, Almaty, Kazakhstan
nuraliyevaassel@gmail.com
ORCID ID: 0000-0001-6451-3743

Abstract. Due to the massive use of electronic media the amount of unstructured text is increasing tremendously day by day. Many of researchers in machine learning work with that data in order to extract relevant and succinct information for such application areas like biomedical natural language processing, maintaining clinical inventories, providing speech aid to challenged children and machine translation where one convert semantic features of one language to another language. NLP technologies help us to improve our communication, achieve our goals and get results from every interaction. They also help us to overcome personal obstacles and psychological problems. By studying NLP methods correctly, we can achieve our goals in a very satisfactory way and overcome the obstacles we face. This paper covers three scientific papers and aims to provide their approach, main idea, techniques and usefulness. Article can be extremely helpful in academics, researches in natural language processing and also to novice specialists.

Keywords: unstructured text, machine learning, natural language processing, feature extraction, information extraction.

Introduction

A large number of unstructured electronic texts are available online, including news feeds, blogs, email messages, government documents, discussion journals, etc., which can be used to create a variety of content. Natural Language Processing (NLP) is a sub-section of computer science and AI that deals with how computers analyze natural (human) languages. NLP allows the application of machine learning algorithms for text and speech. For example, we can use NLP to create systems such as speech recognition, document summarization, machine translation, spam detection, named entity recognition, question answering, autocomputer, predictive text input, etc. The ultimate goal of NLP is to help computers understand language like we do. It is the driving force behind things like virtual assistants, speech recognition, emotion analysis, automatic text summary, machine translation and much more. Natural language processing (NLP) is the intersection of computing, linguistics and machine learning. The field focuses on communication between computers and humans in natural language, and NLP is about making computers understand and generate human language [1,2].

Information extraction means the automatic extraction of structured information such as entities, relationships between entities, and attributes that describe entities from unstructured sources. Information extraction (IE) is the task of automatically retrieving structured information from unstructured and/or semi-structured machine-readable documents. In most cases, this activity is related to the processing of human language text using the Natural Language Processing (NLP) method [3,4,5].

In addition, extracting the semantic relationships between objects in natural language text is a crucial step in understanding natural language applications that recognize the relationships between objects in unstructured text. In this paper, we will review three scientific papers dealing with important tasks in the field of natural language processing [5,6,7].

In this paper, three scientific works are covered: A Detailed Analysis of Core NLP for Information Extraction by Simran Kaur et al., Large Scaled Relation Extraction with Reinforcement Learning by Shizhu He et al, Attention-Based Convolutional Neural Network for Semantic Relation Extraction by Yatian Shen et al.

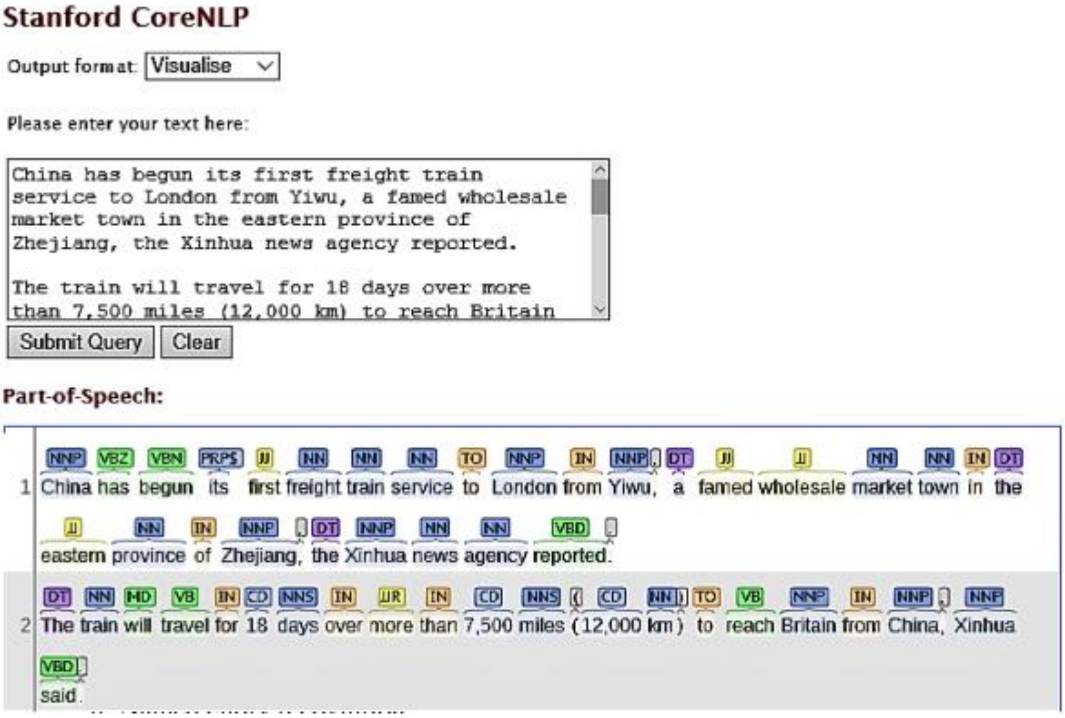
Methodology

In the first article, Simran Kaur and Rashmi Agrawal provided a detailed analysis of the

Stanford Core GNP which provides a range of natural language analysis tools and also examined a wide variety of techniques involved in information extraction and the problems they solve. Information extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. For information extraction, the bootstrapping approach is mainly used. It extracts a large amount of information from unrated seeds and data from texts or other word corpora. Bootstrapping is an extremely powerful approach that extracts good patterns from unstructured language. However, its weak point is its tendency to decrease accuracy over time, since initially there was no tagging. For this algorithm as well, the choice of seeds is crucial for success. In this paper, three approaches to bootstrapping algorithms such as Nomen, Basilisk and Snowball were discussed [8,9].

The first approach essentially consists in checking whether two different algorithms have the same result on the problem or not, which is used to solve the problem of generalized name learning in a biomedical context. Second, the Basilisk algorithm was originally designed to extract semantic lexicons automatically or semi-automatically, including information extraction, answering questions and adding prepositional sentences. Thirdly, in Snowball, he works on the idea of the duality of pattern relations, according to which a good pattern will have good tuples present in it and vice versa by following an alternative approach.

Bootstrapping systems are better suited to natural language processing tasks because of their ability to learn and navigate the syntactically rich, unstructured and extremely complex nature of unstructured natural languages [10].



Stanford CoreNLP

Output format: Visualise

Please enter your text here:

```
China has begun its first freight train
service to London from Yiwu, a famed wholesale
market town in the eastern province of
Zhejiang, the Xinhua news agency reported.

The train will travel for 18 days over more
than 7,500 miles (12,000 km) to reach Britain
```

Part-of-Speech:

1 China has begun its first freight train service to London from Yiwu, a famed wholesale market town in the eastern province of Zhejiang, the Xinhua news agency reported.

2 The train will travel for 18 days over more than 7,500 miles (12,000 km) to reach Britain from China, Xinhua said.

NNP VBZ VBN PRP$ JJ NN NN NN TO NNP IN NNP , DT JJ JJ NN NN IN DT
JJ NN IN NNP , DT NNP NN NN VBD .
DT NN MD VB IN CD NNS IN JJR IN CD NNS (CD NN) TO VB NNP IN NNP , NNP
VBD .

Figure 1 – Standford CoreNLP interface

For basic GNP cases, the Stanford Core GNP tool is excellent because it provides the basic forms of words, their parts of speech, indicates which names refer to the same entities, indicates feeling, extracts special or open class relationships between entity mentions, etc. fig. 1 shows Standford CoreNLP interface. A tool can be run with only two lines of code and can be used for pos tagging, recognizing names, numeric and temporal entities, generating lemmas and relationships between entities using relationship annotators [11].

In the next one, the purpose of sentence relation extraction is to extract facts relevant to the case from the verdicts. In order to extract a largely scalar relationship, the authors used the existing knowledge base to heuristically line up with the texts. In their research using remotely controlled data, they encountered a problem: sentences in a remotely controlled dataset are not marked directly and not all sentences mentioning a pair of entities may represent a relationship between them. To solve this problem, they proposed a new model with enhanced learning. Model is trying to extract relations from every single sentence while the DSRE models aim at extracting relation of an entity pair from all sentences that mention these two entities (the bag). They dealt with two types of experiments on a publicly published dataset. As a result, the method significantly exceeded the comparative baseline, leading to a 13.36% improvement [12, 13, 14, 15].

As an example of report extraction, the authors presented several suggestions and their report. They supplemented the expected reports with sentences to predict the scholarship report, which, compared to the Golden Scholarship report, was used to determine long-term compensation, and then used it to train the report writer. In our model, we need to integrate the expected sentencing reports into the sack report so that we can compare them with the gold sack report to determine the long-term compensation. They followed the last-minute hypothesis to predict the bag report.

In Shizhu He, Kang Liu, Kang Liu, Jun Zhao, Xiangrong Zeng' novel model, first extracted the report of each sentence independently, then predicted the scholarship report based on the extracted reports and compared it with the gold scholarship report. Finally, they used the result of the comparison to guide the formation of the report extractor. For the report extractor they used the PCNN to represent the sentences because it is easier to implement and more efficient for the calculation. The process is: in the raw sentence input, first it is divided into tokens. Then, each token splits into dense vectors, which will be used as input for convolutional neural networks. Finally, a multi-layer perceptron with softmax is applied to produce the probability of each relation. To reduce the variance and make the training faster and more stable they used Williams' simple algorithm that follows the statistical gradient. In other words, or a batch of data with N bags, the baseline is calculated as the average of all the advantages in batch. While conducting the experiment they noted that the bag prediction relies heavily on the relationship extractor, therefore, the results of the evaluation in a remote supervised relationship extraction task can demonstrate the effectiveness of the model [16, 17].

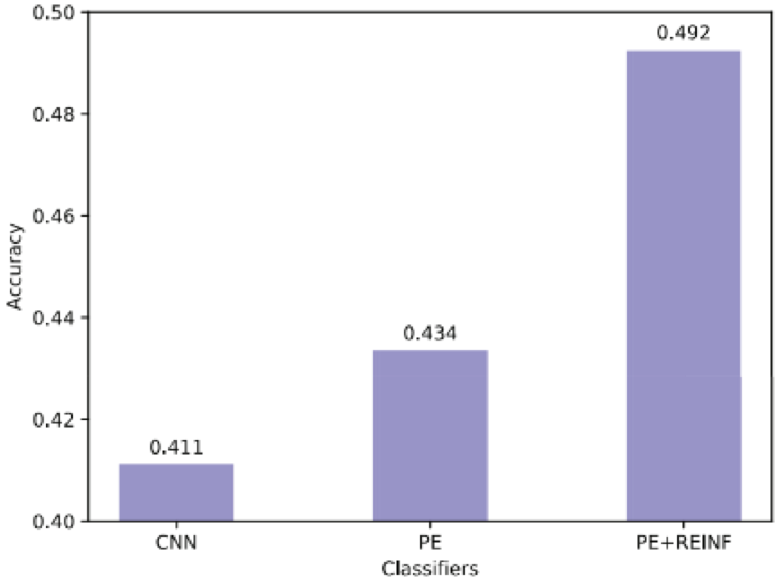


Figure 2 – The accuracy of relation extractor based on CNN, PE and PE+REINF

In fig. 2 one can see, that there is no doubt that their model leans a better relation extractor. As a result, the model achieved an improvement of 19.71% and 13.36% compared to CNN and PE. The improvement was achieved by PE to PE+REINF, which shows that applying reinforcement learning and using remote supervision to guide training can lead to better results.

In the last paper, Yatian Shen, Xuanjing Huang propose a new convolutional neural network model based on attention that makes full use of word embedding, language part tag embedding and position information embedding. Their word-level attention mechanism is better able to determine which parts of the sentence are more influential than the two entities of interest. The architecture makes it possible to learn some important features from task-specific tagged data, while renouncing the need for external knowledge such as explicit dependency structures. The experiments covered the reference data set of SemEval-2010 Task 8, which showed that the new model performs better than several state-of-the-art neural network models [18, 19, 20].

The authors had a hypothesis if the relevance of words with respect to the target entities is effectively captured, if critical words that determine semantic information can be found. Therefore, they proposed to introduce the mechanism of attention in a neural convolution network (CNN) to extract the words important for the meaning of the sentence and aggregate the representation of those informative words to form a sentence vector. Their process is as follows: given a set of x1, x2, ...xn sentences and two corresponding entities, the model measures the probability of each r relation. In their architecture the extraction of the characteristics is the main component, which is composed of the convolution of the sentences and the selection of the context based on attention. After the feature extraction two types of vectors are generated - the vector of the sentence convolution and the attention-based context vector - for the classification of semantic relations. To obtain the conditional probability they applied a softmax operation on all relation types [21, 22].

Table 1. Score obtained for various sets of features on the test set.

Feature Sets	F1, %
WF	74.5
+pF	80.7
+POSF	82.6
+WA	84.3
+WA+(Lexical Feature)	85.9

Authors performed ablation tests on the four sets of features in tab. 1 to determine which type of features contributed the most. From the results one can observe that their learned position embedding features are effective for relation classification. The F1-score is improved remarkably when position embedding features are added. To evaluate the effectiveness of automatically learned features, authors have choosen six approaches as competitors to their method. All models have adopted the word embedding as a representation, except for SVM. Their network model mainly contains four sets of functions: Word Embedding (WF), Position Embeddings (pF), Part-of-speech tag Embeddings (POSF) and Word Attention (WA). Investments in POS increased F1 by 1.9%, the system achieved an improvement of about 2.3% with the addition of Word Attention. When all the features were combined, they achieved the best result of 85.9%. The bottom portion of the table shows the best combination of all the features.

Conclusion

In conclusion, there is no doubt that natural language processing is an area that covers various issues such as speech recognition, natural language understanding and natural language generation. NLP technologies help us to improve our communication, achieve our goals and get results from every interaction. They also help us to overcome personal obstacles and psychological problems. By studying NLP methods correctly, we can achieve our goals in a very satisfactory way and overcome the obstacles we face.

In the first article, the result generated by bootstrapping algorithm are tokens and enhanced relations between them using Basilisk algorithm which provides gold summary which has high confidence and accuracy in order to improve the results of tagged patterns which would further help in other application areas like biomedical natural language processing, maintaining clinical inventories, providing speech aid to challenged children and machine translation where we convert semantic features of one language to another language.

In the second scientific work, authors’ novel method with reinforcement learning results that model outperforms comparative baselines significantly. There are many directions of future work.

Most neural models in relation extraction task are based on convolution neural network and utilize position embeddings as the feature.

In the last paper, authors' an attention-based convolutional neural network architecture for semantic relation extraction model made full use of word embedding, part-of-speech tag embedding and position embedding information. Meanwhile, their word level attention mechanism is able to better determine which parts of the sentence are most influential with respect to the two entities of interest.

Reference:

[1] Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., and Kuksa P. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 2011. 12. 2493–2537.

[2] Narasimhan, K., Yala, A., and Barzilay, R. Improving information extraction by acquiring external evidence with reinforcement learning. In Proceedings of EMNLP, (2016). 2355– 2365.

[3] Hoffmann R., Zhang C., Ling X., Zettlemoyer L., and Weld D. S. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, (2011). 1. 541–550.

[4] Agichtein, E. and Gravano L.. Snowball: Extracting relations from large plain-text collections. In Proceedings of the 5th ACM International Conference on Digital Libraries (DL‟00), San Antonio, (2000), Jun., TX.

[5] Agichtein, E., Eskin E., and Gravano L. Combining strategies for extracting relations from text collections. In Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2000), 2000.

[6] Agichtein, E. and Gravano L. Snowball: Extracting relations from large plain-text collections. In Proceedings of the 5th ACM International Conference on Digital Libraries (DL‟00), San Antonio, Jun., (2000), TX.

[7] Thelen M. and Riloff E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics, (2002). 10. 214–221.

[8] Kaur, S., and Agarwal, R. A Detailed Analysis of Core NLP for Information Extraction. International Journal of Machine Learning and Networked Collaborative Engineering, 2018. 1(01). 33-47.

[9] Lin W., Yangarber R., and Grishman R. Bootstrapped learning of semantic classes from positive and negative examples. In Proceedings of ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data, (2003). 4(4).

[10] Manning C. D., Surdeanu M., Bauer J., Finkel J., Bethard S. J., and McClosky D. The stanford corenlp natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, (2014). 55–60.

[11] Brin S. Extracting patterns and relations from the World Wide Web. In The World Wide Weband Databases, Berlin: Springer Berlin Heidelberg, (1999). 172–183.

[12] Hoffmann R., Zhang C., Ling X., Zettlemoyer L., and Weld D. S. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of ACL, (2011). 541–550.

[13] Jiang, X., Wang Q., Li, P., and Wang B. Relation Extraction with Multi-instance Multi-label Convolutional Neural Networks. In Proceedings of COLING, (2016). 1471–1480.

[14] Zeng X., He Sh., Liu K., and Zhao J. Large scaled relation extraction with rein-forcement learning. In Proceedings of AAAI. (2018).

[15] Turian J., Ratinov L., and Bengio Y. Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, (2010). 384–394.

[16] Wang L., Cao Zh., Melo G. d., and Liu Zh. Relation classification via multi-level attention

CNNs. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, (2016).

[17] Shen Y. and Huang X. Attention-based convolutional neural network for semantic relation extraction. In COLING, 2016. 2526–2536.

[18] Yin W., Schutze H., Xiang B., and Zhou B. Attention-based convolutional neural network for modeling sentence pairs, 2015.

[19] Ji G., Liu, K., He S., and Zhao J. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In Proceedings of AAAI, (2017). 3060–3066.

[20] Li J., Monroe W., Ritter A., Galley M., Gao J., and Jurafsky D. Deep Reinforcement Learning for Dialogue Generation. In Proceedings of EMNLP, (2016). 1192–1202.

[21] Zeng D., Liu K., Chen Y., and Zhao J. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In Proceedings of EMNLP, (2015). 1753–1762.

[22] Nguyen B. and Sameer B. A review of relation extraction. Literature review for Language and Statistics II, 2007.

IRSTI 81.93.29
UDC 004.056

PIPELINE MODULAR MULTIPLIER

S. Tynymbayev[1], R. Sh. Berdibayev[2], A. A. Shaikulova[3], S. Adilbekkyzy[4], T. A. Namazbayev[5]

[1,2,3]Almaty University of Power Engineering and Telecommunication, Almaty, Kazakhstan
[4]L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan
[5]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[1]s.tynym@mail.ru, [2]r.berdybaev@aues.kz, [3]shaikulova_ak_al@mail.ru,
[4]sairan.02.95@mail.ru, [5]tirnagog@mail.ru
[1]ORCID ID: 0000-0002-9326-9476
[2]ORCID ID: 0000-0002-8341-9645
[3]ORCID ID: 0000-0001-9634-143X
[4]ORCID ID: 0000-0002-3929-7070
[5]ORCID ID: 0000-0002-2389-2262

Abstract. Various approaches of multiplying multi-bit numbers modulo are considered. The algorithm of multiplication of numbers is presented, where the process of multiplication modulo is divided into steps and at each step the operation of multiplication is combined with the operation of reducing numbers modulo forms a partial remainder. The circuit solutions for pipeline multiplication of numbers modulo with the analysis of the least significant bits of the multiplier are considered. The proposed modular multiplier does not require preliminary calculations and the calculation results do not go beyond the bit grid of the module. To evaluate the effectiveness, the ratios are used, by which the timing parameters of the multipliers are determined without a pipeline and using a pipeline. Algorithmic validation and verification of the pipeline modular multiplier was carried out on a Nexys 4 board based on the FPGA Artix-7 from Xilinx. Verilog HDL is used to describe the circuit of the pipeline multiplier. The results of a timing simulation of the device are presented in the form of time diagrams, confirming the correct operation of the device.

Keywords: public key cryptosystem, hardware encryption, modular multiplication, remainder former, pipelined multiplier.

Introduction

In asymmetric cryptosystems, the encryption and decryption of data is carried out by exponentiation the numbers a to the power x modulo P ($a^x \bmod P$), which can be implemented in software, hardware-software and hardware [1, 2]. Hardware encryption has a number of significant advantages over software, which has a higher speed and guarantees its integrity [3]. At the same time, the generation and storage of keys, as well as encryption, are carried out in the encoder board itself, and not from a portion of the computer's RAM, it provides security for the implementation of the algorithm itself. Thus, the implementation of the algorithm is protected, which is also an important advantage. Therefore, the development of high-speed operating units of hardware cryptoprocessors for asymmetric encryption, despite their high cost, is an urgent task.

Modular multiplication approaches

The multiplication of numbers modulo can be done in two ways. In the first method, the operation is divided into two stages. At the first stage, n-bit numbers A and B are multiplied and a 2n-bit number C is formed. At the second stage, the product $C = A*B$ is reducing modulo P.

Nowadays, much experience has been gained in the development of high-speed integer multipliers and a squaring device. These include Braun and Wallace multipliers, Dadd multipliers, systolic and Vedic multipliers and quadrants, where the computational complexity is $O(n^2)$ bit operations. But these multipliers are very effective in computing "low-bit" numbers, which are widely used in the construction of operating units of computers of various classes [4].

In cryptography for multiplying multidigit numbers, which make it possible to calculate the required product faster than in $O\left(n^2\right)$ steps (bitwise operations), the Karatsuba method [5] is widely used, the complexity of which is $O\left(n^{\log_2 3}\right)$, Toom-Cook algorithm [6] with complexity of order $O\left(n2^{\sqrt{2\log_2 n}}\right)$ bit operations. Besides the Schönhage-Strassen algorithm [7] allows you to multiply two n-bit numbers in $O(n\log n\log n)$ bit operations.

The modular reduction operation, which is performed in the second stage, is to obtain the remainder of dividing the product $C = A * B$ by the module P. In [8], various modular reduction methods were analyzed. It is shown that the most effective construction tool is a modular reduction device based on a dividing device. The composition of such a dividing device includes a shaper of partial remainders. High-performance matrix and pipeline units for modulating numbers of numbers are easily implemented on the basis of shapers of partial remainders [9, 10, 11, 12, 13].

In the second method, by using the algorithms of Barrett or Montgomery [14, 15, 16] accelerates the process of multiplying large numbers modulo. However, these algorithms require preliminary constant calculations related to the need to use the algorithm for dividing large numbers. Barrett's algorithm requires precomputation constant

$$\mu = \left\lfloor \frac{d^{2m}}{N} \right\rfloor,$$

where $d = 2^k$, k is word size in bits, m is number of words in module N.

The Montgomery algorithm requires precomputing the constant "$r^2(\bmod N)$", using division with remainder.

In the third method, the process of multiplying numbers modulo is performed in many steps, where at each step the multiplication operation is combined with the operation of casting numbers modulo, forming partial remainders. The number of steps is determined by the bit depth of the multiplier.

Multipliers of numbers modulo sequential actions with the analysis of the least significant and highest bits of the multipliers are considered in [17, 18]. Matrix multipliers of numbers modulo were considered in [19, 20].

The main disadvantages of these multipliers are their low speed.

In this paper we consider pipelined device for modular multiplication with analysis of low-order bits, which is an architectural technique for increasing productivity for multiplying of numbers modulo.

Pipeline modular multiplier

With pipeline multiplication of numbers, the entire process of multiplying numbers modulo breaks down into a sequence of completed stages. Each of the stages of the multiplication procedure is performed by the logic gate blocks of its stage, and these logic gate blocks work in parallel.

The results obtained at the i-th stage are transferred to the (i+1)-th stage for further processing. Information is transferred from one stage to another through a buffer memory, which is located between them.

The synchronization of the pipeline is provided by clock pulses, the period of which is determined by the slowest stage of the pipeline and the delay in the buffer memory element.

For simplicity of explanation, we consider the operation of the pipeline using the example of a 4-stage pipeline. The number of pipeline steps is determined by the bit depth of the multiplier. Its structure is shown in fig. 1. The composition of the pipeline is the register PrA, PrB and PrP, where, before the start of multiplication, the multiplier A and factor B and module P are respectively taken.

Each pipeline stage consists of blocks of logic circuits and buffer registers. The number of pipeline steps is determined by the bit depth of the multiplier.

The logic gate blocks of the 1-st stage of the pipeline include the block of the And1 circuit and the former of the partial residuals of the PRF.1. Buffer registers of the 1-st stage of the pipeline are RegB.1, Regr.1, RegR.0, RegP.1.

The logic gate blocks of the 2-nd stage include the block of logic gate circuits And2, PRF.2 and the adder modulo P AddMP.1 registers RegB.2, Regr.2, Regr.1 RegP.2 are buffer registers of the 2-nd stage pipeline.

The block of logic gate And3. PRF.3 and AddMP.2 comprise logic gate blocks of the 3-rd stage pipeline. The registers RegB.3, Regr.3, Regr.2 RegP.3 are buffer registers of the 3-rd stage.

The logic gate blocks of the 4-th stage include the block of logic circuits And4, AddMP.3, the buffer register of the 4-th stage is the register R.

When transferring bits, the multiplier B from the buffer register of the i-th stage of RegB.i+1 to the register (i+1)-th stage of RegB.i+1 only those bits that have not yet entered the operation are transmitted.

When each clock pulse is applied, the corresponding modules P are moved from the buffer register of the i-th stage RegP.i to the buffer register of the i+1-th stage.

After the supply of the first clock pulse CP1 along its leading edge, the contents of the input registers that make up the first three numbers A1, B1, P1 are transferred to the buffer registers of the first stage. In this case, the logical operation A1 & b3 is performed in the block of circuits And.1 and, with b3 = 1, the remainder R0 = A1 is formed at the outputs of the block of circuits And1, which is received in the RegR.0 buffer register.

At the same time, the partial remainder r1 is formed at the outputs of the PRF.1 by performing the operations $r_1 = 2A_1 \bmod P = 2A_1 + \overline{P}_1 + 1$. For this, the multiplier A_1 shifted by one bit to the left is fed to the first input of the PRF.1, and the inverse code of the $\overline{P}_1$ module and signal +1 are fed to the PRF.1 other inputs. The value of r_1 is received in the buffer register Regr.1 of the first stage of the pipeline.

At the same time, the contents of RegB without bit b0 are transferred to the buffer register RegB.1, and the contents of RegP are transferred to the buffer register RegP.1 of the first stage.

On the trailing edge of the pulse CP1 input registers take the second three numbers A_2, B_2, P_2.

After the second clock pulse CP2 has been applied, the contents of the input registers RegA, RegB and RegP, A_2, B_2, P_2 are transferred to the buffer registers of the first stage and during the transfer of the block of circuits And1, the operation $A_2 \& b_0$ is performed and at $b_0 = 1$, the remainder $R_0 = A_2$. R_0 is formed at the outputs of the circuit And1. R_0 is received in the RegR.0 buffer register of the first stage. In addition, a partial remainder $r_1 = 2A_2 \bmod P_2 = 2A_2 + \overline{P}_2 + 1$ is formed in PRF.1, which is stored in the Regr.1 buffer register.

At the same time, the CP2 pulse transfers the contents of the first stage buffer registers to the second stage buffer register. In this case, in the block of circuits And2 and the adder modulo P, an operation is performed to calculate the intermediate remainder R_1 according to the formula

$$R_1 = [(r_1 \& b_1) + R_0] \bmod P_2$$

which is fixed in the buffer register of the second stage Regr.1.

At the same time, the former of the second stage of PRF.2 forms a partial remainder r2, which is fixed in the Regr.2 buffer register.

On the trailing edge of the clock signal CP2, also in the input registers RegA, RegB and

RegP the values of the numbers of the third three A3, B3, P3 are received.

After the arrival of the third clock pulse CP3, the contents of the input registers (A3, B3, P3) are transferred to the buffer registers of the first stage, and the contents of the second stage are transferred to the buffer register of the third stage.

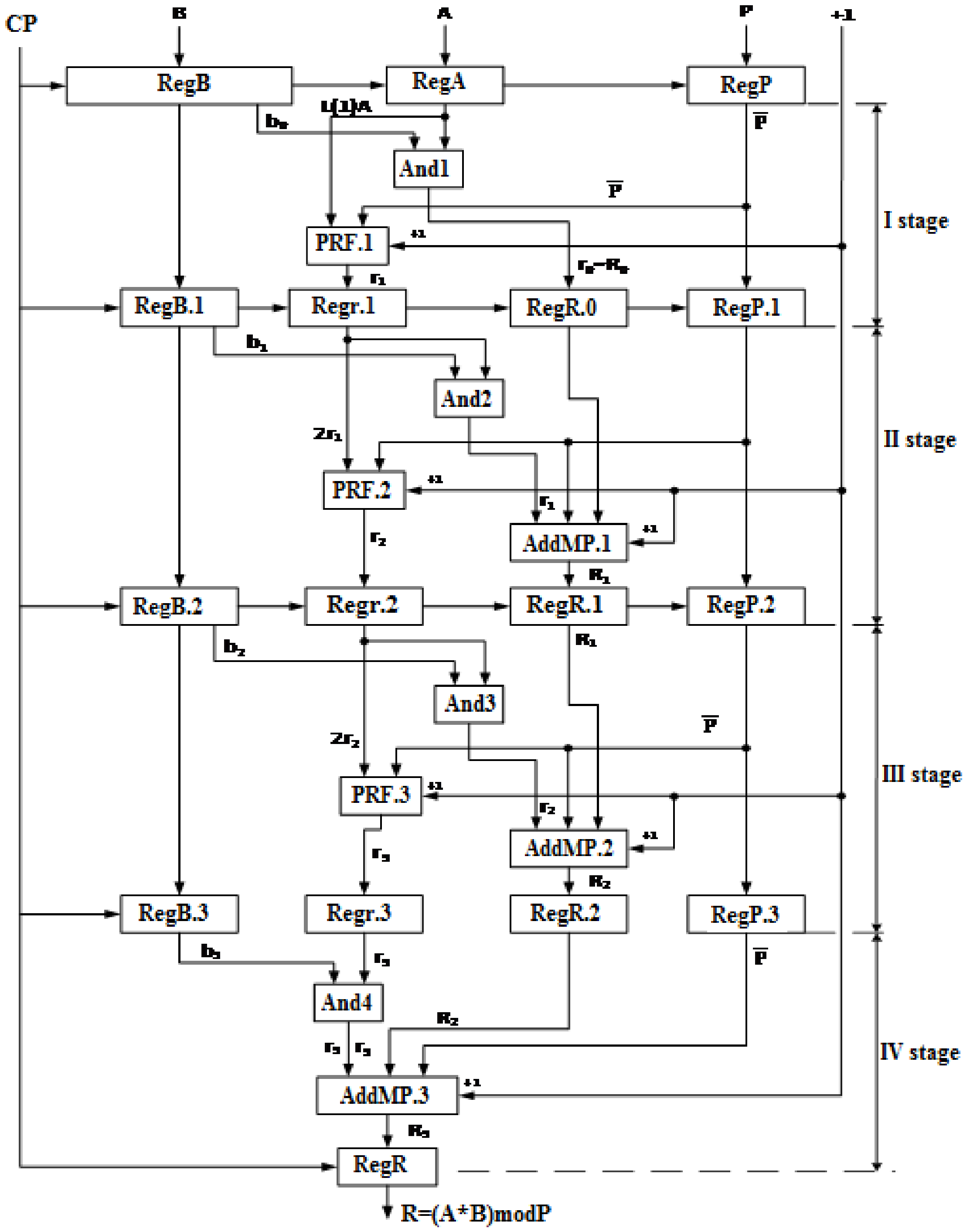


Figure 1 – 4-stage pipelined device for multiplying numbers modulo, where the multiplication begins with the analysis of the least significant bits of the multiplier

In this case, in the first stage the value of the intermediate residue $R_0 = A_3 \& b_0$ is formed, which is fixed in the RegR.0 buffer register and the PRF.1 former computes $r_1 = 2A_3 + \overline{P}_3 + 1$, which is stored in the Regr.1 buffer register.

In the second stage R1, the adder modulo P AddMP.1 calculates where the operation

$R_1 = [(r_1 \& b_1) + R_0] \bmod P_2$ is performed, which is stored in the buffer register Regr.1 of the second stage of the pipeline. In addition, the value $r_2 = 2r_1 + \overline{P}_2 + 1$, which is stored in the Regr.2 buffer register of the second stage of the pipeline, is calculated by the PRF.2.

Under the action of CP3 in the third stage of the pipeline, the value of the partial remainder r3 is calculated by the PRF.3 shaper, where the operation $r_3 = 2r_2 + \overline{P}_1 + 1$ is performed, which is received in the buffer register Regr.3 of the third stage of the pipeline. In addition, an intermediate remainder $R_2 = [(r_2 \& b_2) + R_1] \bmod P_1$ is formed by the adder AddMP.2, which is stored in the buffer register of the third stage Regr.2.

On the trailing edge of CP3, the input registers take the fourth three-number sets A4, B4, P4.

After the fourth CP4 clock pulse is fed into the pipeline, the contents of the input registers (A4, B4, P4) are transferred to the buffer register of the first stage, and the contents of the buffer registers of the first stage are transferred to the buffer registers of the second stage, the contents of the second stage are transferred to the register of the third stage, and the contents of the third step are transferred to the buffer register of the fourth step.

With a clock pulse CP4 in the first stage of the pipeline, the logic gate block And1 calculates the partial remainder R0 by logically multiplying A4 with b0. The value of R0 of outputs And1 is recorded in the buffer register RegR.0 of the first stage, at the same time the partial remainder $r_1 = 2A_4 \bmod P_4$, which is written in the buffer register Regr.1, is calculated by the PRF.1 shaper.

By the fourth clock pulse the PRF.2 shaper calculates the value of the partial remainder $r_2 = 2r_1 + \overline{P}_3 + 1$, which is received in the Regr.2 buffer register. The third clock pulse is also calculated by the adder AddMP.1 intermediate remainder $R_1 = [(r_1 \& b_2) + R_0] \bmod P_3$, which is formed in the buffer register Regr.1 of the second stage of the pipeline.

The fourth clock pulse in the third stage of the pipeline by the PRF.3 former calculates the value of the partial remainder $r_3 = 2r_2 + \overline{P}_2 + 1$, which is received in the Regr.3 buffer register. The intermediate remainder is also calculated by the adder AddMP.2 $R_2 = [(r_2 \& b_2) + R_1] \bmod P_2$, which is written to the Regr.2 buffer register.

At the same time, a fourth remainder in the fourth stage of the pipeline generates a partial remainder by the And4 circuits and the adder AddMP.3 calculates the value $R_3 = [(r_3 \& b_3) + R_2] \bmod P_1$, which is recorded in the Regr.3 buffer register, which is the result of multiplying the number A1 by B1 modulo P1, those. $R = (A_1 \& B_1) \bmod P_1$.

After feeding the inputs of the pipeline CP5, CP6, etc. at the outputs of Regr.3 we get the results of multiplication $(A_2 \& B_2) \bmod P_2$, $(A_3 \& B_3) \bmod P_3$, etc.

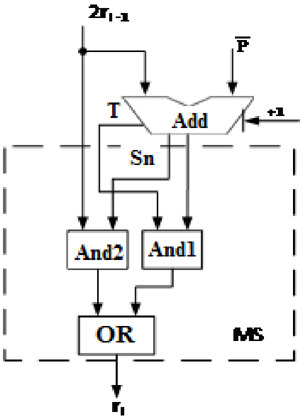


Figure 2 – The structure of the PRF

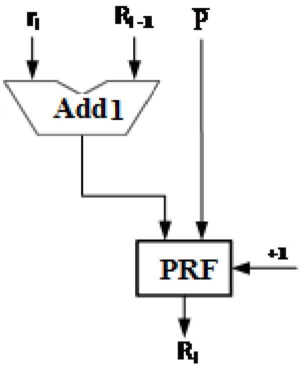


Figure 3 – The structure of the adder modulo P (AddMP)

Fig. 2 shows the functional diagram of the remainder former (PRF), which consists of a binary adder (Add) multiplexer (MS). The multiplexer, in turn, consists of an And1, And2 and OR circuit. The adder has an output P, which captures the transfer from the sign discharge and the sign output $(Sn - Sign)$. If T (transfer) = 1, then Sn = 0 and vice versa. The value of the previous remainder is fed to the left inputs of the adder with a shift of it by one bit in the direction of the highest bit (2ri-1). The right inputs of the Add adder are supplied with bits of the inverse code of the $\overline{P}$ module. To transfer $\overline{P}$ to an additional code, the level +1 is applied to the low order of the adder.

When $2r_{i-1} \geq P$, then the result of addition at the outputs of the adder by transferring T = 1 is output. When $2r_{i-1} < P$, then the value 2ri-1 is given to the output of the circuit with the signal Sn = 1.

In the first way $r_i = 2r_{i-1} + \overline{P} + 1$, and in the second way $r_i = 2r_{i-1}$.

The structure of the adder modulo P (AddMP) is shown in fig. 3, which consists of the adder Add1, where r_i and R_{i-1} are summed and this sum $(R_{i-1} + r_i)$ is modulo P given by the PRF scheme.

Efficiency of pipelined multipliers

To evaluate the effectiveness, it is necessary to determine the ratios by which the time parameters of the multipliers are determined without a pipeline and using a pipeline [4].

The multiplication time of numbers without a pipeline is determined by the formula NKT_k, where K is the number of three-number sets to be processed, N is the number of pipeline steps, T_k is the duration of the clock period, which is determined by the ratio Tk = TAddMP + TBReg, where TAddMP is the summation time by module, TBReg is time of writing the result of processing to the buffer registers.

The execution time of operations on the K input streams of numbers (three-number sets) on the N steps of the pipeline with a clock period Tk can be determined by the relation.

$$T_{NK} = (N + (K - 1))T_K \,.$$

This formula reflects that before the output of the pipeline output of the calculation of the first three polynomials, K cycles must pass, and subsequent results will follow in each cycle.

Then the acceleration of computing S due to pipelining can be described by the formula

$$S = \frac{NKT_k}{(N + (K - 1))T_k} = \frac{NK}{N + (K - 1)}.$$

At $K \to \infty$ the acceleration tends to a value equal to the number of steps K in the pipeline. The gain in time C can be calculated using the formula:

$$C = (NK - (N + K - 1))T_k \,.$$

Consider an example of the execution of multiplication operations modulo on a four-stage pipeline of four three-number sets:

A1 = 14, B1 = 13, P1 = 15; A2 = 11, B2 = 9, P2 = 15;

A3 = 10, B3 = 11, P3 = 15; A4 = 4, B4 = 7, P4 = 15.

The calculation results in each step on the pipeline steps for all four three-number sets are shown in table 2.

In this table, Rij are the numbers of intermediate remainders i (i = 0 ÷ 3) and three-number sets j, where j = 1 ÷ 4.

Table 1. The calculation results

Three-number sets and clock pulses / Stages	A1 = 14, B1 = 13, P1 = 15	A2 = 11, B2 = 9, P2 = 15	A3 = 10, B3 = 11, P3 = 15	A4 = 4, B4 = 7, P4 = 15	–	–	–
	CP1	CP2	CP3	CP4	CP5	CP6	CP7
I	R01 = 14	R02 = 11	R03 = 10	R04 = 4	–	–	–
II		R11 = 14	R12 = 11	R13 = 0	R14 = 12	–	–
III			R21 = 10	R22 = 11	R23 = 0	R24 = 13	–
IV				R31 = 2	R32 = 9	R33 = 5	R34 = 13

Check: R31 = (A1 * B1) mod P1 = (14 * 13) mod 15 = (182) mod 15 = 2;
R32 = (A2 * B2) mod P2 = (11 * 9) mod 15 = (99) mod 15 = 9;
R33 = (A3 * B3) mod P3 = (10 * 11) mod 15 = (110) mod 15 = 5;
R34 = (A4 * B4) mod P4 = (4 * 7) mod 15 = (28) mod 15 = 13.

For our pipeline: $T_{NK} = (N + (K - 1))T_k = 7T_K$.
Multiplication time without pipeline: $NKT_K = 16T_K$.
The time gain is: $C = (NK - (N + K - 1))T_K = 9T_K$.

Fig. 4 shows the timing diagram and the results of modulo multiplication for the above four three-number sets on a four-stage pipeline. Verilog language was used to describe the circuit of the pipeline multiplier.

Artix-7 from Xilinx companies was chosen as the FPGA board.

As can be seen from fig. 4, the first three numbers A1, B1, P1 are fed to the inputs of the i-th stage of the pipeline after the first clock pulse. In this case the intermediate balance R01 = 14 of the first three numbers is calculated.

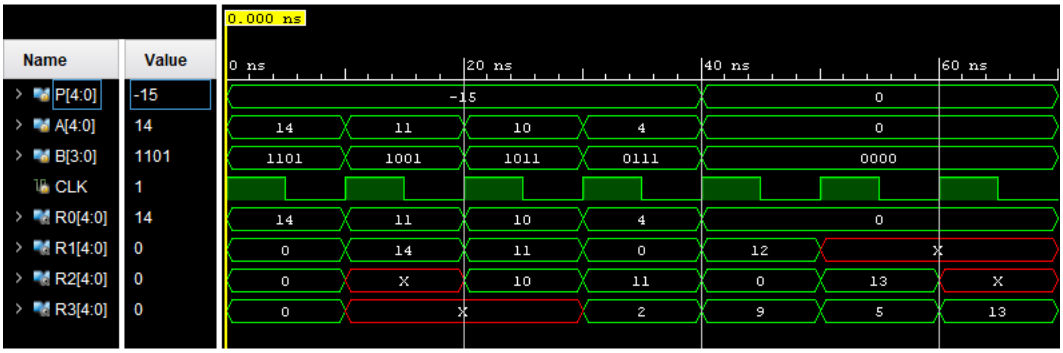


Figure 4 - Timing diagram of the pipeline circuit

During the action of the second clock pulse, the values of the second three-number sets A2, B2, P2 are received at the pipeline input and the intermediate remainder R02 = 11 is calculated, and at the second stage of the pipeline, the intermediate remainder R11 = 14 from the first three numbers is calculated. Similarly, after the third clock pulse is fed to the pipeline input, the value of the third three of numbers A3, B3, P3 is received and an intermediate balance R03 = 10 is formed, and in the second and third stages, R12 = 11, R21 = 10 are formed, respectively.

After applying the fourth clock pulse to the inputs of the first stage of the pipeline, a triple of numbers A4, B4, P4 are received and the intermediate remainder R04= 4 is calculated on the first stage of the pipeline. And on the remaining steps, intermediate residues R13 = 0, R22 = 11 and R31 = 2 are formed. This R31 = 2 is the result of multiplying numbers modulo (A1 * B1) mod P1.

After applying the fifth clock pulse to the II, III, IV stages of the pipeline, R14 = 12, R23 = 0 and R32 = 9 are calculated respectively. R32 = 9 is the result of multiplying the numbers modulo (A2 * B2) mod P2.

After applying the sixth clock pulse to the III and IV stages of the pipeline, the corresponding residues R24 = 13 and R33 = 5 are calculated. R33 = 5 is the result of multiplying the numbers modulo (A3 * B3) mod P3.

After applying the seventh clock pulse to the IV stage of the pipeline, the remainder R34 = 13 is calculated, which is the result of multiplying the numbers modulo (A4 * B4) mod P4.

Conclusion

In the proposed pipeline modular multiplier, preliminary calculations are not required. At each stage of the formation of the intermediate remainder, the operations of multiplication and reduction are combined. All calculations do not go beyond the module bit grid.

Acknowledgement. The work was carried out by the authors within the program-targeted financing of the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (BR053236757 "Development of software and hardware and software for cryptographic protection of information during its transmission and storage in infocommunication systems and general purpose networks").

References

[1] Ryabko B. Y., Fionov AI. Osnovy sovremennoy kriptografii dlya spetsialistov v informatsionnykh tekhnologiyakh [“Fundamentals of modern cryptography for information technology professionals”]. M.: Scientific world, 2014. 173 p. (In Russian)

[2] Akhmetov B. S., Korchenko A. G., Sidenko V. V., Drens Y. A., Seilova N. A. Prikladnaya kriptologiya: metody shifrovaniya [“Applied cryptology: encryption methods”]. Almaty: Satbayev

Univercity, 2015. 496 p. (In Russian)

[3] Aitkhozhayeva E. Zh, Tynymbayev S. T. Aspektyi apparatnogo privedeniya po modulyu v asimmetrichnoy kriptografii [Aspects of hardware reduction modulo in asymmetric cryptography], Bulletin of National Academy of Sciences of the Republic of Kazakhstan, 2014. 5. 88–93. ISSN: 1991-349421. (In Russian)

[4] Orlov S. A., Tsilker B. J. Organizaciya EHVM i system [“Organization of computers and systems”], 3rd ed., SPb.: Peter, 2014. ISBN 978-5-496-01145-7. (in Russian)

[5] Karatsuba A. A., Ofman Y. P. Umnozheniya mnogorazryadnykh chisel na avtomatakh [“Multiplications of multi-digit numbers on automata] . DANSSR, 1962. 145. 293–314. (in Russian)

[6] Cook S. A., Aanderaa S. O. On the minimum computation time of functions. Trans. AMS, 142 (1969). 291–314.

[7] Schonhage A., Strassen V. Bystroye umnozheniye bol'shikh chisel. Kiberneticheskiy sbornik [“Fast multiplication of large numbers”. Cybernetic collection]. 1973. 2.87–98. (in Russian)

[8] Kovtun M., Kovtun V. Obzor i klassifikaciya algoritmov deleniya i privedeniya po modulyu bolshih celyh chisel dlya kriptograficheskih prilozheniy [Review and classification of algorithms for dividing and modulating large integers for cryptographic applications], Cipher Company. URL: http://docplayer.ru/30671408-Obzor-i-klassifikaciya-algoritmov-deleniya-i-privedeniya-po-modulyu-bolshih-celyh-chisel-dlya-kriptograficheskih-prilozheniy.html (date of access: 20.12.2019) (In Russian)

[9] Petrenko V. I, Chipiga A. F. (1995). Umnojitel' po modulu [Modulus multiplexer]. Kombinatsionnyy rekurrentnyy formirovatel' ostatkov [Combination recurrent former of remainders]. Patent of the Russian Federation. No. 2029435 (In Russian)

[10] Petrenko V. I., Sidorchuk A. V., Kuz'minov J. V. (2007) Umnojitel' po modulu [Modulus multiplexer]. Patent of the Russian Federation. No. 2299461 (In Russian)

[11] Kopytov V. V., Petrenko V. I., Sidorchuk A. V. (2009). Ustroystvo dlya formirovaniya ostatka po proizvol’nomu modulu ot chisla [Device for generating remainder from arbitrary modulus of number]. Patent of the Russian Federation. No. 2368942. (In Russian)

[12] Tynymbayev S. T., Aitkhozhayeva Y. Zh., Adilbekkyzy S. High speed device for modular reduction. Bulletin of National Academy of Sciences of the Republic of Kazakhstan, 2018. 6(376). 147–152. ISSN 2518-1467 (Online). ISSN 1991-3494 (Print). doi: https://doi.org/10.32014/2018.2518-1467.38.

[13] Tynymbayev S., Berdibaev R. S., Omar T., Shaikulova A. A., Magauin B. Bystrodeystvuyushchiye ustroystva privedeniya chisla po modulyu [“High-speed devices of reduction”]. Materials of the XIV International Asian School-Seminar "Problems of optimization of complex systems. Almaty, 2018. 2 (In Russian)

[14] Barrett P. Implementing the Rivest Shamir and Adleman Public Key Encryption Algorithm on a Standard Digital Signal Processor. In: Odlyzko A.M. (eds) Advances in Cryptology — CRYPTO’ 86. CRYPTO 1986. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 1986. 263. doi: https://doi.org/10.1007/3-540-47721-7_24.

[15] Montgomery P. L. Modular Multiplication without Trial Division. Mathematics of Computation, 1985. 44(170). 519–521. doi: https://doi.org/10.2307/2007970.

[16] Pisek E., Henige T. M. (2013) Method and apparatus for efficient modulo multiplication. Patent US No.8417756 B2.

[17] Tynymbayev S., Berdibayev R. Sh., Omar T., Gnatyuk S. A., Namazbayev T. A., Adilbekkyzy S. Device for multiplying modulo numbers with analysis of the lower bits of the multiplier. Bulletin of National Academy of Sciences of the Republic of Kazakhstan,2019. 4(380). 38–45.

[18] Tynymbayev S., Aitkhozhayeva Y. Zh., Berdibayev R. Sh., Abilda B. G. A multiplier of numbers modulo with analysis of the two most significant bits of the multiplier per step. Materials of the International scientific-practical conference "Actual problems of information security in Kazakhstan". Almaty, 2020. 236–241.

[19] Tynymbayev S., Berdibayev R. Sh., Aitkhozhayeva Y. Zh., Omar T., Adilbekkyzy S. The matrix matrix of the multiplier of numbers modulo, where the multiplication begins with the analysis of the least significant bits of the multiplier. Certificate of state registration of rights to the object of copyright of the MOJ of the RK, no. 2689, dated: Apr. 8, 2019.

[20] Tynymbayev S., Berdibayev R. Sh., Aitkhozhayeva Y. Zh., Omar T., Shaikulova A. A., Magauin B. The matrix scheme of the multiplier of numbers modulo, where the multiplication begins with the highest digits of the multiplier. Certificate of state registration of rights to the object of copyright of the MOJ of the RK, no. 2690, dated: Apr. 8, 2019.