BASIC PRINCIPLES OF CREATING SOFTWARE SYSTEM TO CONTROL AND CORRECT ERRORS IN TEXT

DJURAEV MUROTALI KARSHIYEVICH
Master Teacher, Department of Applied Mathematics and Informatics, Termez State University,
E-mail: djurayev_mk20@mail.ru

ABSTRACT

This article describes the basic approaches, principles and methods of creating an information processing system for control and correction of text errors in natural languages, as well as the development of methods of assessment and analysis, identification of probabilistic and quantitative indicators of system efficiency. Methods, algorithms and basic approaches of enterprises in the creation of software systems for data processing in YEHA (single electronic document circulation) are studied.

The input text provided to solve information management and processing problems is usually encoded. There are many ways to encode text data

KEYWORDS: Documents, algorithms of morphological analysis, coding of text data, linear, modules, programmable methods of management of digital information.

INTRODUCTION

In world linguistics, since the 50s of the twentieth century, not only linguists, but also scientists in other fields have been working on solving problems related to language and text with the help of computer technology. However, these issues provide new opportunities for the linguistic study of any text, the selection of language material through information technology programs and a certain acceleration of its processing, the development of technologies for linguistic analysis of large texts and information processing programs and their linguistic support.

In this regard, the Government of the Republic of Uzbekistan is taking appropriate measures to develop information systems (IS), computer networks and a single electronic document management system (EDS). However, there are still unresolved issues, in particular, one of them is the reliable transmission and processing of digital, graphical data, as well as the provision of text messages on the computers of enterprises and organizations. This is because there are errors in various texts when entering, transmitting and processing data to a computer.

Therefore, the development of methods, models, algorithms and software systems for word processing in the Uzbek language is an urgent task for the observation and correction of spelling errors, as well as its practical application to the EHA system of enterprises and organizations.

MAIN PART

Observations show that one of the important criteria for the activities of enterprises EHA is reliable data exchange. However, in real terms, the reliability of the data is very low and approximate $3{,}4 \cdot 10^{-2}$ equal to error / character. It was found that 85% of the total number of distortions in the texts are related to the human operator system in data scanning and recognition processes. In addition, the reliability of the processed data for the normal operation of the system 10^{-5}-10^{-6} an error / character level is required, which emphasizes the importance of solving the problem of building a software system for tracking (detecting) and correcting (automatically correcting) errors in texts. [1.B.48-55]

In addition, information input and transmission methods, written texts, documents, graphics and other types of textual information can be transmitted by human operator through scanning devices, including software system recognition, as well as in the form of files on the carrier, e-mail. The text transmission and processing steps in which the error management and correction process is performed are usually performed before the text data is encoded and decoded. Based on effective methods of coding, compression, and decoding in error detection, problems associated with the development of algorithms and error management and correction programs are studied and analyzed.

There is a coding system for detecting errors in texts, and the input text provided to solve information management and processing problems is usually encoded. Various methods can be used to encode textual information, in particular:

Shannon-Fano, Huffman, arithmetic coding;

Dictionary methods: Ziva-Lempel, Lempel-Ziva-Velcha transformation;

Algorithms used in ASCII machine code and others.

The peculiarity of these methods is that the transmitted text can usually be encoded with real numbers or using decimal numbers. This allows you to create an algorithm for adjusting the boundaries for decoding, which should be used to control the reliability of the actual decoded text. In this regard, the tasks related to the development, research and application of software methods based on the use of the features, methods and rules of arithmetic coding recommended for the management and correction of errors in texts are fully explored. [2. B. 25-32]

Based on this, software methods for monitoring information will be developed. Track, detect and automate textual data errors in foreign and domestic practice software methods for correction have not been sufficiently studied and there are no developments that can be effectively applied in practice. The results of the study explored the possibility of using software methods of digital information management using algorithms for generalizing single and double spelling errors, such as linear, modular, and plane, in texts.

As noted above, human operator, scanning, and recognition errors account for a large proportion of the total number of violations. However, such errors are described as multiple (k-gram) errors. Therefore, the tasks of error management and correction in the texts should be solved in a new formula, taking into account the stated conditions of data processing. In addition, software installed in a computer data processing system requires the creation of favorable conditions for the detection and correction of errors using modern computer technology.

In addition to using software methods using artificial reproduction, the use of natural resources to manage and correct errors in text is also effective.

Error correction software in determining the task of managing information based on natural resources can be done on the basis of the development of the following error management methods: along the boundaries of code sets, coded text data, special catalogs of word forms of natural language. In addition, the use of words, letter sequences, or methods that take into account the specific features of the coding system is highly effective and allows the management of the boundaries of existing codes, and so on. Methods that take into account statistical relationships and data interrelationships, semantic methods that take into account the characteristics of language and the structure of word formation, methods of detecting grammatical errors based on morphological analysis, non-morphological methods (dictionary and wordless). Vocabulary methods include methods based on the use of an unstructured list of all existing word forms, while non-word methods include methods of testing some word forms (digrams, trigrams, n-gram methods) and the hash code method, using classification algorithms and determining the text of the language being studied methods and so on.

It should be noted that among the methods mentioned, the main method of checking spelling errors in texts is the morphological analysis of word forms. In this regard, in the framework of this work, the methods of developing data processing programs for spelling control in the Uzbek language (Latin, Cyrillic) are being studied step by step.

It is known that natural languages are divided into three groups according to their structural principles: analytic, inflective and agglutinative. Agglutinative languages that include most Turks, especially Uzbek is distinguished by its intermediate position between analytic and inflective languages. On the one hand, they retain a very rich system of infectious and word-forming affixes, but on the other hand this system is distinguished by considerable constructiveness and simplicity. However, despite the relative simplicity and constructiveness of Turkic languages, linguists are less interested in the problems of developing spelling tools for them, focusing mainly on European languages. There is a list of research groups involved in the development of automatic processing of speech and texts in natural languages.

The next problem is to take into account the Uzbek spelling - it has two different graphics systems (Cyrillic and Latin) at the same time. Uzbek spelling - in both powers of the inspector, the spelling should be checked and, if necessary, translated from one authority to another. Preparing an adequate representative dictionary on the roots of modern Uzbek literary words is the most difficult and time-consuming task, it is interesting and unresolved to study the main aspects of theoretical and practical problems that can be used to create a computer system based on morphological analysis of Uzbek texts.

In addition to the above, in order to assess the text quality of errors in texts, the task is to check the quality of the text, the solution of which requires the development of specific algorithms and programs to compare source and managed text. The following are used to solve this problem:

- ✓ letter, symbolic comparison;
- ✓ comparison of words;
- ✓ compare words with highlighted roots;
- ✓ line comparison;
- ✓ compare the frequency characteristics of words, lines, lines and pages;
- ✓ Compare the times when letters, characters, words, lines appear on the page.

Thus, the basic approaches to the creation of a computer system for tracking and correcting errors in texts are proposed, tasks are developed that define the following areas of theoretical and applied research:

- ✓ Analysis of possible processes occurring in different conditions of transmission and processing of information and development of methods for determining the optimal size of the dictionary of word forms of the Uzbek language;
- ✓ Development of methods and algorithms for text management and correction based on:
- ✓ Optical text detection algorithms;
- ✓ probable error model;
- ✓ arithmetic coding method;

The basis for the development of software for monitoring information on the methods of linear, modular, planned assembly and morphological analysis based on a multi-level model for the expression of word forms in the Uzbek language.

Breaking statistics of operators, scanning devices and recognition systems used in the development of methods and algorithms for tracking and correcting errors in texts in natural languages are studied, the types of occurrence of error flow are identified. [3. B. 10-15]

RESULTS AND DISCUSSION

Currently, many works are devoted to the problems of automatic error management in natural language texts, special systems for error detection have been created and used in practice, however, some complex morphological systems of error management subsystems have not been fully resolved.

However, solving problems aimed at correcting errors identified in the texts is of great theoretical and practical interest. However, few studies have focused on the study of automatic error correction methods, where the problem is mainly considered in the form of a problem statement and is limited to the development of individual procedures of programs.

The task of error correction, as a rule, begins with the study of the mechanisms of errors and the rules of their occurrence in the texts, the types and characteristics of errors in the form of broken letters, symbols and words. Depending on the nature of the errors, mechanisms and algorithms for their correction are constructed. In this regard, great attention should be paid to the study of error statistics in IP activities in order to more accurately study the nature of errors, their classification and determine the probability of their occurrence.

Sources of errors can include a human operator, technical means of data transmission and processing, communication channels, scanning devices, and recognition systems, such as an IP user who makes spelling mistakes because he does not know the language well.

A detailed study of the error statistics of data from the literature sources and data centers of organizations in different fields.

In this case, the study of error statistics of the human operator is a key issue, IP practice shows that the most unreliable connections are the stages of preparation of data and transmission to computer media, where many operations are performed manually by the human operator. For a single error related to the failure of technical means or communication channels, it was proved that there were 3-4 errors due to the fault of the operator. Many errors occur during data preparation and input, and these errors are distributed as follows:

- ✓ Through the fault of the human operator - 94-97%;
- ✓ Due to equipment failure - 1.5-2%;

- ✓ Due to inaccurate entries in the document - 0.5-4.5%.
- ✓ Insufficient attention to error - 23.9%;
- ✓ Insufficient training - 58.6%;
- ✓ Reflex error - 11.9%,
- ✓ Other errors occur for reasons beyond the control of the operators.

The results of the study allow us to conclude that the probability of error depends on the qualifications of the operators. The flow of errors obeys Pousson's law.

A study of operator error statistics in the laboratory and production environment shows that the most common operator errors are error-type errors, errors, overrides, and character repositioning. [4. B. 58-65]

The classification of errors resulting from operator error is given in Table 1. The probability of operators' error when referring to magnetic media on a single computer with different sample data is given in Table 2. It also shows the average probability values of errors for a single character calculated from different literary sources. These distortion statistics can be used to study the effectiveness of software methods to verify the reliability of textual data.

Table 1

Types of operator errors (total percentage)				
Set another character	Skip characters	Put extra characters	Reset characters	Others
58,61	3,92	2,44	5,36	29,67
59,0	3,7	2,2	4,1	31,0
47,8	15,2	5,0	7,1	24,9
50,0	13,0	7,0	2,0	28,0
47,1	13,1	5,8	5,0	29,0
30,57	2,58	0,47	3,41	62,97
55,8	3,7	2,1	4,1	34,3
The average specific error is gravity				
49,84	7,88	3,57	4,43	34,26

Table 2

Test numbers	Probability of operator error	The average probability of error
1.	$2,2.10^{-4}$	$2.16.10^{-3}$
2.	$5,0.10^{-3}$–$5,0.10^{-4}$	
3.	$5,0.10^{-3}$	
4.	$3,0.10^{-3}$–$1,0.10^{-3}$	
5.	$4,34.10^{-4}$	

The quality of the processed data also depends on the accuracy of the scanning devices and recognition systems used in the IP, and Table 3 provides information on the accuracy of the five most common recognition systems.

Table 3

№	System type	Recognition accuracy (in%)	Number of known languages	Probability of recognition error	Source
1	Fine Reader 5 Office	88,9	176	$2,22\cdot10^{-2}$	Journal ChIP №12, 2018 год, Soft Press Publishing House
2	Fine Reader 4 Professional	83,5	53	$3,3\cdot10^{-2}$	
3	Recognita Plus 5	84,8	114	$3,04\cdot10^{-2}$	
4	Cunel Form 2000 Master	67,8	15	$6,44\cdot10^{-2}$	
5	Text Bridge 9 Pro Business Edition	81,5	56	$3,7\cdot10^{-2}$	

Theoretical and experimental studies of error statistics at all stages of information processing have shown that the largest number of malfunctions in information system performance are related to scanning and recognition errors

(10^{-2}) and human operator ($\approx 10^{-3}$) and the developed methods of error correction should take into account the nature, type, and characteristics of the distortions of the processed data identified in this study. [5. P. 88-96] [6. P. 36-41]

To increase the reliability of information, the main basis for creating a software processing system is from artificial and natural resources use is. Artificial abbreviations are formed by adding additional control information to the content of the transmitted message, such as entering a sequence of codes, checking characters, or verifying using software error management methods. A natural redundancy, as a rule, occurs in the processed original text or in its coded version, resulting from the comparison or uneven distribution of letters and symbols on lines or pages.

The increase in the volume of information reproduction, on the one hand, ensures high reliability of information, on the other hand, leads to an increase in the cost of its transmission and processing.

In this regard, there is a problem with the development of such an additional volume detection methodology that provides the required reliability at the lowest cost.

Solving this problem consists of several steps.

The first of them:

- ✓ Synthesis of possible processes that occur in the process of information management and transmission;
- ✓ Second:
- ✓ Determining the amount of data, taking into account the possible processes going on;
- ✓ Determining the amount of information when receiving a regular message;
- ✓ Determining the amount of data when using the correction code and determining the excess amount required for the conditions under consideration for data transmission.

In organizing the synthesis of probabilistic processes in the detection of errors, we define the average probability of errors in the transmission of the P_α-sign to P_α (fractional number, code combination) and n, the conditional probability of obtaining CB_α in the transmission of CB_α or a (a → a) transition probability a we call:

a) $P_\alpha=1$ $j=i$
b) $P_\alpha=0$ $j\neq i$
c) $P_\alpha=1/B$ $0<\alpha<B$

Where B is the range of CB variability determined by the relation B = Xm, where X is the basis of the code; m is the length of the code sequence.

Condition (a) means the error-free transmission of information, (b) the transmission of data with complete distortion, and (c) the transmission of the conditional sign a means that any sign a is equal to a constant conditional probability B-1. The value of CB_α can be denoted by several numbers (singular, decimal, hundredth, etc.) representing the m-sequence of decimal codes. The combination of such m-sequences makes up the length of the processed messages. [7.B.33-36]

Consider the probability of receiving a single-coded message as P (S), with the introduction of information management procedures, the probability of P (S) is divided into two parts: the probability of receiving messages correctly - Q (S) and the probability of incorrect reception - P_H(S), etc.

The ideal mode of information management is P (S) = Q (S). When using a reasonable method of information management, P (S) = 1-Q (S) -min. In an inefficient control procedure - P (S) = P_H(S).

$P(S)< <P_{\alpha,}$ $P_\alpha=10^{-3}/10^{-4}$

Fulfilling this condition requires an increase in data. Suppose the code S_0 is passed. The probability that the correct decision is made is determined by the probability Q (S_0 / S_1) = P (S_0 / S_1), where the code P (S_0 / S_1) – S_1 indicates the conditional probability when S_0 is made. [8.B.22-27] [9.B.45-49]

CONCLUSIONS

In conclusion, the article mainly studied and analyzed the following processes.

1. The main approaches in creating a software system for tracking and correcting errors in texts are based on: corruption statistics; data coding methods and models; software methods of information management based on artificial and natural resources; methods, models, morphological analysis algorithms.

2. Developed methods: detection of excess quantity, which provides the necessary quality of text management; synthesis and evaluation of possible processes that occur in the management process in different conditions

of information transmission; calculation of the memory capacity of the software of the data processing system for tracking and correction of errors in natural languages, in particular in the Uzbek language. It was found that the recommended excess should be at least 0.4 to ensure the reliability of at least $^{10-6}$ data; Rational memory used in data processing in Uzbek is 2^{16} bits.

3. Development of methods, algorithms, functional and software modules for monitoring and error correction based on the use of text in the optical detection system.

4. Develop methods and algorithms for tracking and correcting errors in text based on verifying that code combinations of letters and symbols belong to the subdomain of allowed values. Consideration of the efficiency area and limited capabilities of the developed algorithms with the criterion of probability of error detection. If the error probability value of the error $P_{\alpha_i} \approx 3{,}14 \cdot 10^{-3}$ / character is close to the practical conditions, signs of control and correction reliability have been found to increase to two magnitude levels.

5. The rules for describing the formal analysis of Uzbek word forms were studied, based on which a multi-stage model of morphological analysis was developed, the structure and algorithm of the software system for error management and correction with a dictionary of word forms that do not require in-depth knowledge of the language.

REFERENCES

1) Марчук Ю.Н. Компьютерная лингвистика:2007 [Computational Linguistics] учеб. Пособие –Москва, Восток-Запад, 2007. 280 с.
2) Пулатов А. Компьютер лингвистикаси. 2011. [Computer linguistics] Тошкent академнашр. 2011 й. 20 б.
3) Белоногов Г.Г., Зеленков Ю.Г. Алгоритм автоматизированного исправления орфографических ошибок в текстах. 1996. [Algorithm for automated correction of spelling errors in texts] М.: ВИНИТИ, 1986, 244 с.
4) Рахимов А. Компьютер лингвистикаси асослари. 2011 [Computer linguistics] Тошкент академнашр. Ўқув қўлланма. 2011 й. 280 б.
5) Abdurahmonova N. Mashina tarjimasining lingvistik asoslari.2012 [Linguistic bases of machine translation] –Toshkent: Akademnashr, 2012. 250 b.
6) Abdurahmonov X., Rafiev A., Shodmonqulova D. O’zbek tilining amaliy grammatikasi. 1992. [practical grammar of the Uzbek language.] Toshkent: o’qituvchi, 1992. 230 b.
7) Мальковский М.Г. Диалог с системой искусственного интеллекта. 2013 [Dialogue with artificial intelligence system] М.: МГУ, 2013, 216 с.
8) Пулатов А. Мухамедова С.Х. Компьютер тилшунослигида автоматик тахрир қилувчи дастурнинг лингвистик таъминотини яратиш асослари // [Fundamentals of creating a linguistic software for an automatic editing program in computer linguistics] Ўзбек тилшунослигининг долзарб масалалари. – Тошкент, ТДПУ илмий тўплами, 2003. 210 б.
9) Н.Махмудов, А.Мадвалиев, Н.Махкамов, М.Аминов., “Узбек тилида иш юритиш” [Office work in Uzbek language] изд-во Тошкент, стр 223., 1990 г.
10) И.И. Жуманов, Н.С.Мингбоев. Контроль достоверности информации стационарных случайных процессов. [Control of the reliability of information of stationary random processes] Узбекский журнал «Проблемы информатики и энергетики, №5-6; Ташкент, 1999.
11) Джураев М. К., Каршиев Д. М. НОВЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ПРОЦЕССЕ РЕФОРМИРОВАНИЯ СИСТЕМЫ ОБРАЗОВАНИЯ //Интернаука. – 2018. – №. 48-1. – С. 15-16.
12) Джураев М. К., Каршиев Ж. М. МЕТОД КОРРЕКЦИИ ТЕКСТОВ НА ОСНОВЕ ВЕРОЯТНОСТНОЙ МОДЕЛИ СОВЕРШЕНИЯ ОШИБОК //Педагогика ва психологияда инновациялар. – 2019. – №. 3.