Available online at www.scientiaresearchlibrary.com

Scientia Research Library

ISSN 2348-0424
USA CODEN: JETRB4

Journal of Engineering And Technology Research, 2022, 10 (1):1-7
(*http://www.scientiaresearchlibrary.com/arhcive.php*)

MECHANISMS FOR OPTIMIZATION OF DETECTION AND CORRECTION OF ERRORS IN COMPUTER TEXT PROCESSING SYSTEMS

Djumanov Olimjon Israilovich*, Tursunova Luiza Komilovna

Candidate of Technical Sciences, Associate Professor, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan
Graduate student, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan

---

ABSTRACT

Efficient algorithms have been developed with mechanisms for detecting and correcting errors in texts in natural languages based on models of morphological and n-gram structured grammar analyzes. Variants of error control with a dictionary and without the use of a dictionary of word forms with mechanisms for controlling the reliability of elements based on the distortion metric and the technique of minimizing the time from the beginning to the end of the work are implemented.

Keywords: dictionary, word forms, error detection and correction, optimization, search, recognition, control time.

INTRODUCTION

Formulation of the problem. As a measure of combating errors in texts, as a rule, studies are considered on the control and correction of spelling errors based on morphological and n-gram structured natural language grammar analyzes. [1,2].

However, these models require a large volume of vocabulary and specific knowledge of the processed language. In this regard, solving problems related to search, recognition, clustering, structuring, identifying distorted elements (letters, words) in a text, obtaining frequency characteristics in the form of small dictionaries for optimizing control and correcting spelling errors is an urgent research problem. Among them, the problem of identifying erroneous elements of the text is solved with the construction of a reference natural language dictionary, on the basis of which the text and search procedures are transmitted for comparison with the total volume of the dictionary and recognition of the desired word [3,4].

When solving the set tasks, it is also important to determine the rational volume of the required vocabulary, since the search for a word in an excessively large volume of the vocabulary of word forms negatively affects the performance indicators of the system, which lead to a decrease in the

search speed and significantly slows down the process of monitoring and correcting errors [5].

Algorithms for detecting and correcting errors in texts during their work use procedures for representing a chain of word forms, comparing them with a dictionary based on the mechanism of forming various hypotheses [6].

A hypothesis is accepted that a word form is used close to the original one, which is additional information. This information helps to identify valid elements of the text [7].

To reduce the cost of searching and processing information, the algorithm is based on the use of softer rules for making decisions about the reliability of information, which are determined on the basis of the condition of mandatory coincidence with the word form found in the dictionary, which boils down to the implementation of effective procedures for enumerating word forms. Two mechanisms of control and correction of errors are investigated: with the comparison of a word with a dictionary and a decrease in the number of calls to disk memory [8].

The calculation of these indicators directly affects the efficiency of the algorithm [9].

Four options for evaluating the efficiency of the error detection and correction algorithm are considered when [10]:

- the number of operations in the control procedure is counted from the start of the algorithm until the first version of the correction is received;

- the number of operations is counted until the correct option is obtained;

- the number of operations is counted until the last version of the fix is received;

- the number of operations is counted until the completion of the algorithm.

Minimizing the number of algorithm accesses to disk memory. To count the number of calls to disk memory, the structure of a specific dictionary of word forms of the Uzbek language is used [11].

Note that the dictionary is a large array stored on a disk with block access, also oriented towards the use of parallel computing technology when building a system based on an n-gram structured model [12].

The principle of block access to information proposed by us consists in organizing procedures using small sections of the dictionary, where the reading of a new block into memory is carried out by an elementary access to disk memory. At the next call to the analysis procedure, a new block is read only if it did not remain in memory from the previous call [13,14].

The dictionary is ordered lexicographically, so that words that are closely related in alphabetical order are in the same block of the dictionary.

We have summarized all the options for constructing an error correction algorithm designed to correct any one-letter, multiple elementary, and some combined isolated distortions. To assess the efficiency of algorithms for detecting and correcting errors, we introduced a specific distortion metric, which allows, along with other conditions, to take into account the distortion probability distributions of the tested chains.

Improving the reliability of information based on distortion metrics. The dictionary of word forms used is considered to be lexicographically ordered. The problem of error correction under consideration consists in finding all possible words for a given string of letters that is not contained

in the dictionary [15].

To simplify the presentation of research results and the implementation of algorithms, we present a variant where the proximity criterion (distortion metric) for one-letter substitutions is taken as a verification condition: two words are considered close if their lengths are the same and differ by exactly one letter.

We will assume that the letter string *W* is fed to the input of the information control system and denote the position of mismatch by *D*, and as *d* - the nearest letter.

In what follows, we denote by *w* the letters in the chain *W* at position *D*. The main property of position *D* is that no changes in the original chain *W*, leaving the first *D* letters of the chain *W* unchanged, can lead to an admissible chain.

Based on the adopted designations, we present the distortion metric for assessing the process of monitoring the reliability of textual information, according to which, to characterize the distorted information, the criterion used is represented as a bounded non-negative function $d(x,y)$, where $x, y \in \mathbf{X}$ is the set of the alphabet of the generated chains.

The introduced measure is symmetric: $d(x,y) = d(y,x)$ and the fulfillment of equality $d(x,y) = 0$ means match $x = y$, where *x* denotes the transmitted image symbol, and *y* denotes the received symbol.

Let us introduce an extension for the distortion metric on sequences of length *N* symbols $x^N = (x_1, x_2, ..., x_N)$ and $y^N = (y_1, y_2, ..., y_N)$ as follows:

$$d^N(x^N, y^N) = \frac{1}{N}\sum_{k=1}^{N} d(x_k, y_k). \quad (1)$$

The distorted part of information in a sequence of messages of length *N* is denoted by D_1, by *M* we denote the set of introduced specific characteristics of text elements.

The value D_1 characterizes the distortion of information, the maximum allowable when embedding control rules into it.

Let us introduce the value $p(\tilde{O}^N, \acute{O}^N)$ - the main statistical parameter, represented by the averaged distribution of errors, which is an important parameter of the efficiency of algorithms for detecting and correcting errors. The study was carried out both for a uniform distribution of distortions and for conditional distributions of the form $Q^N(y^N/x^N)$ [16].

The lower bounds of the reliability of the controlled texts are obtained for the conditional distribution function $Q^N(y^N/x^N)$, which is represented by distortions from set $\mathbf{X}^N$ to set $\mathbf{Y}^N$ in the form of

$$\sum_{x^N \in \mathbf{X}^N} \sum_{y^N \in \boldsymbol{X}^N} d^N(x^N, y^N) Q^N(y^N / x^N) p(x^N) \le D_2, \quad (2)$$

where D_2 is considered to be the lower limit of distortion. The essence of the introduction of the limitation of the value D_2 is that in the process of information control the user is trying to correct the detected error, i.e. invalid text element, adjusts the value D_2 to preserve the quality of the original word based on prior information.

For real systems, $D_2 \le D_1$ is usually performed.

For practical implementation, it is of interest to consider the case when the selected function f_N is limited by the average distortion between $\tilde{\mathbf{X}}^N$ and $\mathbf{Y}^N$:

$$\sum_{m,k^N,\tilde{x}^N,y^N} d^N(\tilde{x}^N, y^N)\, Q^N(y^N / f_N(\tilde{x}^N, m, k^N))\, p(\tilde{x}^N, k^N) \le D_2 . \quad (3)$$

Definition D_2 in accordance with (3) assumes that the controller knows the exact probabilistic characteristics of distortions of text elements.

Based on this, the assessment of the reliability of information control within the lower and upper bounds (D_1, D_2) is presented in the form of the probability of distortion of the checked word in a sequence of length *N:*

$$P_{e,N} = \frac{1}{|\mathbf{M}|} \sum_{m \in \mathbf{M}} P(\phi_N(Y^N, K^N) \ne m / M = m), \quad (4)$$

where the controlled chains of word forms are considered equiprobable and is chosen from the set ***M***.

The probability $P_{e,N}$ is the probability averaged over the set of all messages.

Minimizing the time from the beginning to the end of the algorithm. The main property of the function *d* is that no changes in the string *W* for the letter being checked, leaving the first *D–1* letters unchanged and leading to the appearance at position *D* of any letter greater in alphabetical order than *w*, but less than *d*, also does not lead to valid chain.

An additional requirement for the procedure for matching a word with a dictionary is the requirement to establish in the process of matching the position of the mismatch between *D* and the nearest letter greater than *d*.

Note that the dictionary search algorithms used in the system in any case use the marked properties of the information, but do not give them out as the results of the system operation.

To organize the work of the correction procedures, we used a tree-like dictionary organization model, which allows us to speed up the search for the information necessary for correction [17].

In a simplified version, the algorithm for correcting errors based on the information received is based on the use of the two words of the dictionary that are closest alphabetically to the given word. At the same time, information about the two nearest alphabetically to a given word of the dictionary is used for a priori screening out of hypotheses containing letters that are less alphabetically at the position of non-coincidence *D*.

At each iteration, the algorithm replaces the letter with a new letter *v* at some position *V* in the original word *W*, which is the current variable position.

Algorithm for adapting the position of the letter in the checked word. The algorithm includes the following steps.

Step 1. The position *V* is used as an adaptable parameter and the first position from the chain *W* is selected.

Step 2. Organization of the external cycle. The variable position *V* is fixed . The first letter of the alphabet is chosen as the first letter to replace *v*.

Step 3. Organization of the internal cycle. The letter in the variable position *V* is replaced by the selected letter *v*. The received hypothesis is presented to the comparison procedure with the dictionary.

Step 3.1. If the required word is found in the dictionary, it is issued as an error correction option.

Step 3.2. The letter in the word to replace *v* is presented, following the current value of *v* alphabetically. If there is no such letter, then the symbol of the end of the alphabet is taken.

If the searched letter from the word is in the dictionary, i.e. $D_2 = V$, then d_2 is taken as the next value of *v*.

Step 3.3. If the required word is not found in the dictionary, then the number *D* and the letter *d* returned by the analysis procedure are considered.

If $D>V$, then the next value of *v* is chosen as in step 3.2.

If $D = V$, then $v = d$ is chosen, which leads to a decrease in the number of hypotheses.

If $D <V$, the value of *v* becomes the end of the alphabet. The latter possibility is carried out only once, at the moment of the deepest recursion at step 4.

Step 4. A forward and backward passage of the current positions *V* and *v* is organized and compared with the letter at position *V* in the original word *W*.

Their matches with the original word are checked.

The procedures are then applied recursively to the rest of the word. By stopping the recursion, the condition of reaching *v* as a character of the end of the alphabet is accepted.

Step 4.1. The letter *v* selected for replacement is considered. If it is alphabetically less than the letter at this position in the original irregular chain, then a backward pass through the word positions is performed and the transition to step 3 is made.

Step 4.2. If, on a direct pass through the positions of the chain *W*, the letter for the replacement *v* turns out to be alphabetically larger than the letter in the original word *W* at position *V*, then the recursion is entered. The letter *v* is stored on the stack, the position number V is incremented, and you go to step 2.

Step 4.3. If the value of v is the end of the alphabet, then the recursion is exited: the number of the variable position *V* decreases, the previously stored letter *v* is taken from the stack, and the transition to step 3 is performed.

Step 4.4. The iteration continues at the given recursion level, interrupted in the forward pass. If the position number V cannot be decreased, then the algorithm ends.

CONCLUSION

Thus, the scientific and methodological foundations and effective mechanisms of search, recognition, clustering, approximation, information processing have been developed, which are based on the application of soft rules for making decisions about the reliability of information.

Mechanisms for detecting and correcting errors with a dictionary and without using a dictionary of

word forms have been implemented. A method for evaluating the effectiveness of algorithms for detecting and correcting errors has been developed and implemented, which contributes to the optimization of parameters such as the number of operations in the control procedure is counted from the beginning of the algorithm until the first correction option is obtained; the number of operations is counted until the correct option is obtained; the number of operations is counted until the latest fix is received; the number of operations is counted until the completion of the algorithm.

REFERENCE

[1] Akhatov, A. R., & Zhumanov, I. I. (**2007**). Algorithm for quality control of texts in electronic document management systems. *Journal "Bulletin of TUIT*, (2), 68-72.

[2] Jumanov, I. I., & Xolmonov, S. M. (February **2021**). Optimization of identification of non-stationary objects due to information properties and features of models. In IOP Conference Series: Materials Science and Engineering (Vol. 1047, No. 1, p. 012064). IOP Publishing.

[3] Ibragimovich, J. I., Isroilovich, D. O., & Maxmudovich, X. S. (November **2020**). Effective recognition of pollen grains based on parametric adaptation of the image identification model. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-5). IEEE.

[4] Kholmonov, S. M., & Absalomova, G. B. Methods and algorithms for improving the reliability of textual information of electronic documents. SCIENCE AND WORLD, 43.

[5] Kholmonov, S. M., & Absalomova, G. B. (**2020**). Improving the reliability of texts based on logical criteria and the knowledge base of electronic documents. In ENGINEERING SCIENCES: PROBLEMS AND SOLUTIONS (pp. 15-19).

[6] Zhumanov, I. I., & Sharipova, M. (**2013**). Optimization of Uzbek language spelling control based on stochastic search models. In MODERN MATERIALS, ENGINEERING AND TECHNOLOGY (pp. 129-133).

[7] Isroil, J., & Khusan, K. (November **2020**). Increasing the Reliability of Full Text Documents Based on the Use of Mechanisms for Extraction of Statistical and Semantic Links of Elements. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-5). IEEE.

[8] Jumanov, I. I., & Karshiev, K. B. (May **2020**). Mechanisms for optimization of detection and correction of text errors based on combining multilevel morphological analysis with n-gram models. *In Journal of Physics: Conference Series* (Vol. 1546, No. 1, p. 012082). IOP Publishing.

[9] Jumanov, I. I., Karshiev, K. B., & Tishlikov, S. A. (**2019**). Examination of the efficiency of algorithms for increasing hereliability of information on criteria of harness and the cost of processing electronic documents. *International Journal of Recent Technology and Engineering*, 8(2), 4133-4139.

[10] Israilovich, D. O. ., & o'g'li, T. F. S. (**2022**). Description of the Core of the Spelling Control Software Package in the Sphinx Framework Environment. Middle European Scientific Bulletin, 21, 133-138. Retrieved from https://cejsr.academicjournal.io/index.php/journal/article/view/1073

[11] Zhumanov, I. I., & Karshiev, H. B. **(2019).** Optimization of the reliability of information based on the database of electronic documents and features of the rules for controlling the knowledge base. Problems of Computational and Applied Mathematics, (3 (21)), 57-74.

[12] Jumanov, I. I., & Karshiyev, K. B. (**2019**). Effectiveness analysis of generalization of algorithms for increasing information reliability based on usage of information redundancy of electronic documents. Scientific Journal of Samarkand University, 2019(2), 9-14.

[13] Ibragimovich, J. I., & Azamat o'g'li, A. J. (**2022**). Optimization of Text Processing In Documents of Automated Office Work Systems. *EUROPEAN JOURNAL OF INNOVATION IN NONFORMAL EDUCATION*, 2(2), 374-378.

[14] Ibragimovich, J. I.., & Erkinovich, T. A.. (**2022**). Control of the Reliability of Textual Information in Documents Based on Neuro-Fuzzy Identification. Middle European Scientific Bulletin, 21, 144-149. Retrieved from

https://cejsr.academicjournal.io/index.php/journal/article/view/1075

[15] Ibragimovich, J. I.., & Abdusalyamovich, D. B.. (**2022**). Optimization of Neural Network Identification of a Non-Stationary Object Based On Spline Functions. International Journal of Innovative Analyzes and Emerging Technology, 2(2), 49–55. Retrieved from http://openaccessjournals.eu/index.php/ijiaet/article/view/1021

[16] Ibragimovich, J. I., & Baxromovna, M. M. (**2022**). Adaptive Processing of Technological Time Series for Forecasting Based on Neuro-Fuzzy Networks. *International Journal of Human Computing Studies*, 4(2), 30-35.