N. M. Bykov, Cand. Sc. (Eng), Assist. Prof.; D. E. Balhonskyy,
I. V. Kuzmin, Dr. Sc. (Eng.), Prof.

ANALYSIS OF STATISTIC CHARACTERISTICS OF MORPHEMES OF UKRAINIAN LANGUAGE

There had been developed an algorithm and soft ware for the determination of statistic and transitive probability of text morpheme. There had been given the theoretical analysis of the problem of building of efficient hierarchical strategy of text reorganization, and suggested the procedure for building and optimal tree of classification of text images.

Key words: *analysis of statistic characteristics, morphems, efficient strategy of recognition, hidden Macrovian nets, optimal classification procedure.*

Introduction. Practice in recognizing the handwritten symbols shows hat using graphic information for their description does not allow to receive satisfactory results from the point of view of speed and reliability, which causes the necessity of using linguistic information, which is contained in text document [1]. It is quite natural to use the context information, namely lexical and statistical information. Lexical information is very easy to use when elements for recognition are the morphemes – minimal parts of the word, which may be split (for example, prefix, suffix, root, ending). Then it is possible to use statistic and lexical information on morphemes for building the models of words of the text in the kind of hidden Marcovian nets (HMN) which would enable to apply the known algorithms for recognizing on these nets. More than that, the procedure of segmentation into morphemes will be executed rarely in comparison with the character - oriented recognition.

Task setting. The authors in work [2] developed software for building data base of Ukrainian morphemes to ensure the possibility of using morphological information in the task of recognizing the text document. As the result, there had been made the data base, which contains more than 60.000 morphemes of Ukrainian language. However, the usage of one data base without using the other statistic information will not allow to optimize the process of text recognition. This causes the necessity in solving the task of determination of statistic characteristics of morphemes in the kind of their transitional probabilities. Using data base of statistic characteristics allows to facilitate the process of recognition of hand written and other texts, written in the atypical styles due to module hierarchical architecture and apparatuses of hidden Markovian nets (HMN). After the procedure of text segmentation into morphemes and recognition of the next morpheme as the alternative to the following morpheme, such nets allow to choose the morphemes with the biggest transitional probabilities from the data base. That is, there is no necessity in comparing the graphic image of the morpheme (grapheme) to every possible standard. Decision making is made by choosing an alternative with the biggest total possibility. Such an approach to entering, processing and recognition of texts improves the running speed and reliability of the process. To realize the described ideas, the given paper solves the tasks of development of efficient strategy for text document recognition during its entering on computer and procedure of search for optimal classifier of text images, as well as developing algorithms for determination and analysis of statistic characteristics of morphemes of Ukrainian languages with an aim of their usage on the lexical level of recognition.

Theoretical analysis of the problem of building an efficient hierarchical strategy of text recognition. Any text document may be considered not only as a graphical image, but as a bearer of language information which is used for its transmission in this or that communicative system [3]. From this point of view, text graphic reflects different information levels, which are characteristics of communicative act: pragmatic, semantic, syntactic, lexical, morphologic, segmatic and affective [1]. There appears a question: which level of information and in what sequence shall be used in automated process of entering and recognition of text documents to receive the maximum possible

speed am minimum possible mistakes and cost. To solve this problem, the authors suggest the new technology of electronization the text documents, which, along with the recognition graphic images use partial text comprehension. The entering process is considered as the process of interaction between the entering device and language thesaurus of the computer system of text comprehension. During text image scanning, the entering device singles out the following attribute of the grapheme, belonging to this or that information level of language, which is used by the system for decreasing the entropy on the text unit and narrowing the range of candidates for decision making. In the work [3] the authors continued the formal task setting for entering process optimization and processing text documents, which considers it as the classification three of text images on different information levels.

Optimization of the process of text image recognition is done according to the information criteria of efficiency

$$Э_{\text{Р}} = \frac{\text{I}_{\text{Р}}}{C_{\text{Р}}}, \tag{1}$$

suggested in [3], where I_p – volume of information, received by the system of text recognition and comprehension, is determined with the consideration of entropy properties of text images ; $C_{\text{Р}}$ – cost of the system;

$$C_{\text{Р}} = C_{\text{Х}} + C_{\text{К}}, \tag{2}$$

where $C_{\text{Х}}$ – difficulty in calculation of attributive images description ; $C_{\text{К}}$ – difficulty in calculation of images classification.

Since the difficulty C_p of the recognizing system is the adapted sum of difficulties of each of the hierarchical levels of recognition, and informativity I_p is the non-descending function of probability of the correct recognition, the optimal strategy is the composition of recognizing algorithms, which maximizes the relation I_i / C_i on each of the levels. The succession of algorithms composition in the optimal strategy must correspond to the succession of location of the classification tree levels, which, in turn, corresponds to the information levels of the text documents.

The solution of the problem of choosing proper branching factor allows to narrow the search range in the optimized search procedure of optimum solution tree. The work [4] shows that the minimization of total classification error and time classification gives the boundaries of the branch factor B_r, which is chosen during the construction of the optimal decision tree:

$$2 \le B_r \le 5. \tag{3}$$

So, the set peculiarities of the decision tree allow to narrow the range of search when solving the task of determination of optimal classification tree of text images.

The solution of the task of building an efficient decision making strategy in the kind of classification tree may be fulfilled by the procedure of optimization of „ controlling search prior to return" [4]. In this procedure the criteria (2) controls the search of such decision tree structure among the possible ones, in which each searching stage chooses the configuration unit with the highest criteria value. For the set of tree unit Ω_i^h the searching procedure is executed in the following steps:

1. On the basis of the selected sign $x^h \in X$ there takes place one of the possible divisions $\pi^h \in \Pi$ of the unit Ω_i^h info subset units - scions $\{\Omega_j\}, j = \overline{1,m}.$ Sign x^h is chosen against the distinction matrix in a way that the branching factor is kept within the limits, determined in (3). There h – level (high) of classification tree, X – a priory signs alphabet .

2. The criteria values (1) for the received unit configuration has to be calculated.

3. Repeating steps 1 and 2, allows to build the other possible divisions with further calculation of the criteria value.

4.It is necessary to determine the configuration for which the criteria has the maximum value, determining at the same time the optimum set of signs $\overline{B}_r$ for the given tree unit and optimum step of the classification algorithm.

The table of pair- wise text grapheme distinction w_i and w_j is used as the 'distinction matrix" upon all the signs of their description from the a priory alphabet on the basis of the selected in the space sign of distance d_{ij}.

Analysis of the statistic characteristics of morphemes. Let us introduce the world grapheme on the morpheme level of the classification tree in the kind of succession ***O*** of the observation vectors:

$$O = \vec{o}_1, \vec{o}_2, ..., \vec{o}_L, \tag{4}$$

where $\vec{o}_l$ – vector of morpheme image.

In such a case the task of word recognition in the text may be considered as calculation of credibility maximum.

$$\arg \max_i \{P(w_i / O)\}, \tag{5}$$

where w_i is the *i*-th word from the dictionary.

According the Bayesian formula

$$P(w_i / O) = \frac{P(O / w_i)P(w_i)}{P(O)} \tag{6}$$

the most probable word grapheme in the text image is determined by probability. The direct evaluation of the joint conditional probability $P(\vec{o}_1, \vec{o}_2, ..., \vec{o}_L / w_i)$ from the body of the text shall not be appropriate which is caused by the big number of possible observed sequences. In the majority of cases the task of evaluation of fission density of conditional probability $P(O / w_i)$ is substituted by more simple problem of evaluation of parameters of Markovian model of the text generation M. This model is a machine with the final number of states; during the determination of state *i* ,the vector of grapheme image $\vec{o}_l$ with the probability of $b_i(\vec{o}_l)$ is generated. Apart from that the transition from the state *i* to the state *j* is described by the probability a_{ij}. The selection of the most probable grapheme of the word is executed by finding the most appropriate succession of state:

$$P(O / M) = \max_X \left\{ a_{x(0)x(1)} \prod_{l=1}^{L} b_x(\vec{o}_l) a_{x(l)x(l+1)} \right\}, \tag{7}$$

where $P_l(\vec{o}_l)$ – probability of morpheme image vector observation $\vec{o}_l$, $a_{x(l)x(l+1)}$ – probability of transition from grapheme $\vec{o}_l$ к $\vec{o}_{l+1}$, X – set of states which the model reproduces with the consideration of space localization of states in Markovian model of the text, the authors suggested the modification of this model which means the addition of the requirement (7) to the requirements:

$$\sum_{l=1}^{L} P_l(\vec{o}_{(l)}) = P(L), \tag{8}$$

where $P_l(\vec{o}_l)$ – average of distribution of the word grapheme length w_i.

Conclusions. Thus for the realization of the algorithm of recognition of the word grapheme in the text (7) on the morpheme level, it is necessary to determine the static and transitional

probabilities of morpheme in the text as well as statistic characteristics of the length of the morpheme and words (their images). The given work presents the developed algorithm of calculation of static and transitional characteristics of morphemes on the basis of application of the body test of the text and developed by authors in [5] morpheme data base of Ukrainian language. The results of algorithm work are fixed in the kind of the two matrixes, the first of which fixes the static probabilities of morphemes, and the second - transitional probability. Table 1 presents the second matrix.

In table, the accepted symbols: m_1, m_2, .., m_N – morphemes of Ukrainian language; P (m_i/m_j) – probability of transition between the i-th and the j-th morphemes.

Table 1

Matrix of transitional probabilities of morphemes.

Morphemes / Probabilities	Morpheme 1	Morpheme 2	...	...	...	...	Morpheme N
Morpheme 1	0	P (m_1/m_2)	...	...	...	...	P (m_1/m_N)
Morpheme 2	P (m_2/m_1)	0	...	...	...	...	P (m_2/m_N)
...	...	...	0	...	...	...	...
...	...	...	...	0	...	...	...
...	...	...	...	...	0	...	...
...	...	...	...	...	...	0	...
Morpheme N	P (m_N/m_1)	P (m_N/m_2)	...	...	...	...	0

The algorithm of creation of matrix of transitional probability of morphemes of Ukrainian language comprises the following steps:

1. Reading data base of Ukrainian language morphemes .
2. Reading word set from the test body of the text, on which the data base of probabilities will be built.
3. Initiation of cycle from the first to the last morphemes (i = 1; i <= N), where N – number of morphemes in data base.
4. Initiation of internal cycle from the first to the last morpheme (j = 1; j <= N), where N – the number of morphemes in data base.
5. Null the counter (k) of the found i-th and the j-th morphemes, following one another.
6. Initiation of the cycle from the first to the last word of the text (w = 1; w <= M), where M – number of words in the text.
7. Search for the i-th and j-th morphemes in the w-th word.
8. Return to step 5 for transition to the next word.
9. Upon the completion of cycle 6 (when all words are covered and the sum (counter k) of the found words in the whole set of the i-th and j-th morphemes is calculated, we determine the probability of transition between the i-th and the j-th morphemes: P (m_i/m_j) = k / N.
10. Write down the determined probability P (m_i/m_j) to the data base.
11. Upon the completion of cycle 4 (when all the j-morphemes are covered) return to the cycle 3.
12. Upon the completion of cycle 3 (when all the i- morphemes are covered) we formulate the report and quit the program.

Table 2

Example of determined probabilities

Morphemes / Probabilities	re	cog	ni	ti	o	n
роз	0	0,0457	0,001	0,0255	0,0548	0,00023
піз	0,0652	0	0,0522	0,0453	0,0985	0
нав	0,0001	0	0	0,0781	0,0001	0,0012
ан	0,0001	0	0,00112	0	0,0268	0,0556
н	0	0	0	0,123	0	0,0434
я	0	0,002	0	0	0,0897	0

REFERENCES

1. Пиотровский Р. Г. Текст машина, человек. — Ленинград: Наука", 1975. – 326 с.

2. Биков М. М. Використання інтелектуальних методів в розпізнаванні символів / М. М. Биков, Д. Є. Балховський, А. Раїмі // Інформаційні технології та комп'ютерна інженерія. – 2007. – № 2 (9). – С. 121 – 125.

3. Нова інформаційна технологія введення і оброблення текстових документів в автоматизованих інформаційно-пошукових системах // Автоматика-2008: доклади XV міжнародної конференції з автоматичного управління, 23 – 26 вересня 2008 р., – Одеса: ОНМА. – 992 с.

4. Биков М. М. Розробка ефективної стратегії прийняття рішень в комп'ютерних інтелектуальних системах / М. М. Биков // Вісник Хмельницького національного технічного університету. – 2005. – Ч.1. – Т. 2, № 2. – С. 22 – 30.

Bykov Mykola – Cand. Sc. (Eng), Assist. Prof., Professor with the department of computer control systems., тел.: (0432)-598430, e-mail: nmbdean@ksu.vstu.vinnica.ua.

Balhovskyy Dmytro – Post-graduate student with the department of computer control systems., тел.: (0432)-598222, e-mail: vinbudya@yandex.ru.

Kyzmin Ivan – Dr. Sc (Eng.), Professor with the department of computer control systems., тел.: (0432)-598222, e-mail: nmbdean@ksu.vstu.vinnica.ua.

Vinnytsia National Technical University.