УДК 004.415 (045)

**Badoryna L.**

# DEFINITION OF RELEVANCE OF TEXT EXPRESSIONS IN INFORMATION SYSTEMS

**Institute of computer technology**
**National aviation university**

*In the article we offer a model of calculation of a ratio parameter (relevance) of meaningful terms of reference definition in answers. Here is given an example of the calculation of a numerical synonymic parameter of meaningful terms*

## *Introduction*

Introduction of progressive forms of training and development of modern information technologies creates the necessity of automated assessment of the student's knowledge. Great value for automated educational systems have models of assessing answers not in the form of chosen variants, but in the form of a free text of any length with synonym concept estimation. A special urgency gets the problem of development of answers' analysis model on the task of the open type, demanding to enter from the keyboard the certain formulation of this or that term of a subject domain. There is an objective necessity of transition to computer testing of students' knowledge. Thus, on the foreground rises the problem of automatic assessment of students' knowledge. This problem is simple enough, if the student is offered to choose one or more right answers from a set of variants, but it becomes considerably difficult, if the procedure of testing provides a developed answer in any form, that is with his or her own words in natural language. In the latter case it is possible to appreciate the student's knowledge only by comparative text analysis of the answer with the set standard reference text and to assess their relevance. Thus, all wordforms, terms of the subject domain and grammatical structures of the statement should be considered and assessed with the use of all possible synonyms. [1]. Integration processes, introduction of telecommunication means, computerizing of human activity have represented a set of new problems and tasks in scientific area which is between computer technologies and linguistics. Development of modern information technologies in educational sphere creates the necessity of the automated control over students' knowledge. Great value for automated systems in educational sphere has the estimation of answer models not in the form of the selected variants, but in the form of a free text of any length with words-synonyms estimation. During the educational process using such means as computer information technologies, in particular, there is a number of theoretical and practical questions concerning the adequacy of these technologies, and also control of knowledge and habits. With the advent of multimedia and global networks (the Internet) it is obvious, that computers are capable to teach great volume of prime and instructive knowledge, representing them in attractive form which motivates the process of training. One of the most actual directions in modern information technologies is on the basis of algorithms the development of effective approaches to processing texts with the purpose of the linguistic analysis and structured text information and formation of the semantic analysis on the basis of texts.

Modern scientific researches of linguists [4], give to us a number of approaches which allow formalizing the representation of each linguistic unit in the form convenient for machining. At the automated control of knowledge of subject terminology there appears a problem of comparison of two definitions of one term: the definition, given by the teacher (reference definition), and the definition, given by the one who is being taught (answer). The estimation of correctness of the text answer is based on the method of absolute concurrence of the answer to one of

standards which are saved in the test system database. As the definition of the term is formed through the system of base concepts (terms), having its own definition, we offer to use quantity synonymic subject indicators to calculate a relevant parameter of the answer to the open type problem. Methods of text answer processing are complicating the opportunity of the element analysis of the text answer and the reference answer which is saved in the database, and impose restrictions on the format of the open type problem; therefore, it is necessary to construct a model of analytical calculation of a relevant parameter.

***Statement of the problem***

To estimate the relevance degree of standard reference definition and the answer of a trainee it is necessary:

To establish mutual monosymantic synonymic conformity with terms of standard reference definition and answer;

To calculate the value of a relevance parameter of standard reference definition and answer.

Statements are considered as a set of terms. Thus, the standard reference definition should be considered as a set of base terms, and the answer should be considered as a set of terms *t*, for each of them it is necessary to find a corresponding base term *e*. The search of conformity of a base term and an answer term proposes the definition of function $e = \varphi(t)$ and the calculation of the size of synonymic conformity $k = \theta(e,t)$. Thus, the pair <*e*,*k* > will allow characterizing a term *t* in relation to a term-standard *e* . It means conformity of answer terms with base terms.

Let *A* be a set of standard-term definition, *B* – a set of answer terms.

Then the description of standard definition and answer is as follows:

$$A = \{e_1, e_2, \ldots, e_i, 1 \le i \le N\},$$

$$B = \{t_1, t_2, \ldots, t_i, 1 \le i \le M\},$$

*N* – quantity of terms of standard definition;

*M* – quantity of answer-terms.

To calculate the conformity with terms of standard definition and answer it is necessary to characterize terms *t* according to terms-standards *e*. We are going to define synonymic conformity of answer-terms with standard definition. Comparing separate terms of standard definition and answer there can be following situations which should be solved.

I. One term of standard definition corresponds to only one base term of the answer.

It can be presented as a biactive display between sets *A and B :*

$$\alpha : B \to A, a_i = \alpha(b_j), a_i \in A, b_j \in B$$

In this case between terms of standard definition and answer there is a mutual monosymantic tie. All relations between terms of standard definition and answer should be brought to a similar kind.

II. One term of standard definition corresponds to some various answer-terms.

In this case there exist the intersected sets $\{a_i, b_i\} \cap \{a_i, b_b\} \neq \varnothing$ . Each of these sets is characterized by the function *q* designating a synonym parameter of the terms

$$a_i, b_j : k_m = \theta(a_i, b_j)$$

To achieve the aim it is necessary to remove from consideration one of the intersected sets by the following rule:

1. If $\theta(a_i, b_j) \tilde{>} \theta(a_i, b_b)$ it is possible to remove the set $\{a_i, b_b\}$ as the term $b_j$ is the closest synonym to the term $a_i$, and to use for further processing a parameter $k_h$, which describes numerical value of synonymic term $c_i$ and $b_j$ .

2. If $\theta\ a_i, b_j\ = \theta\ a_i, b_b\ =$ it is possible to remove any of sets $a_i, b_j$ *or* $a_i, b_b$ , as terms $b_j$ and $b_b$ are equally close synonyms to the term *i a* and to use for further processing the parameter *k*, describing numerical

value of synonymic term $c_i$ and the remained term of actual definition.

3. If $\theta(a_i,b_j) \prec \theta(a_i,b_b)$, it is possible to remove the set $\{a_i,b_j\}$ as the term $b_b$ is the closest synonym to term $a_i$, and to use parameter $k_a$, for the further processing describing numerical value of synonymic term $c_i$, $b_b$.

III. Several various terms of standard definition correspond to the same term of actual definition.

In this case there exist the intersected sets $\{a_a,b_j\} \cap \{a_i,b_j\} \neq \varnothing$. To achieve the stated aim it is necessary to remove from consideration one of the intersected sets according to the following rule:

1. If $\theta(a_a,b_j) \prec \theta(a_i,b_j)$, it is necessary to remove set $\{a_i,b_j\}$ and $k_h$

2. To use the parameter $k$ for further processing which describes the numerical value of synonymic term $c_a$ and $b_j$.

3. If $\theta(a_a,b_j) = \theta(a_i,b_j)$, it is possible to remove any of sets $\{a_a,b_j\}$ or $\{a_i,b_j\}$ and to use parameter $k$, for the further processing which describes the numerical value of synonymic term $i$ $c$ and the remained term of actual definition.

4. If $\theta(a_a,b_j) \prec \theta(a_i,b_j)$, it is possible to remove the set $\{a_a,b_j\}$ and to use parameter $k$, for further processing which describes the numerical value of synonymic term $c_i$ and $b_j$. In this case it is impossible to use proportionally to calculate both numerical parameters because of the peculiarities of the result in the system of assessment. There is a methodological aspect of the chosen decision: if semantics of the sentence is as follows, so it is necessary to use some terms, but the trainee used instead of them only one, and to his opinion, the generalizing term, hence, in according to his consciousness these terms are poorly distinguished, and it is necessary to make special methodical job.

IV. Some various terms of standard definition have some common synonyms. In this case to achieve the aim it is necessary:

1. To choose the set $\{a_i,b_j\}$, characterized by the highest numerical parameter $k_h = \theta\{a_i,b_j\}$.

2. To remove from the further consideration all other sets in which there are the chosen elements $a_i$ and $b_j$

3. Among other sets to continue choosing and removing the sets with the maximal parameter $k_a = \theta\{a_b,b_a\}$ according to the same rule until all intersected sets are not found.

In case if there are simultaneously several sets with identical maximal numerical parameter $k$, it is necessary to choose only one of them and again to make the analysis. As a result all intersected sets are removed. It means that mutual monosymantic conformity with significant terms of standard and actual definitions is established.

We are choosing the no intersected sets, which participate in the estimation of knowledge. The set of no intersected sets are considered, which have the biggest value of parameter k=1 or k=0,8.

***Calculation of a relevant parameter***

The model of calculation of a conformity parameter of meaningful terms in reference definition and the answer [1] is used for this purpose.

Let *WAB* – a relevant parameter of reference definition and the answer of the one being taught. Its value is estimated under the formula:

$$W_{AB} = f(\omega\eta) \qquad (1)$$

where $\omega$ – function of numerical values of a synonymy parameter of meaningful terms in reference and actual definitions;

$\eta$ – function of quantity of meaningful terms in reference definition and the answer.

The parameter $w$ is function of three variables: synonymy factor of terms $K_{AB}$ in reference and actual definitions, quantity of

meaningful terms in reference definition N and quantity of meaningful terms in actual definition M:

$$\omega = \theta(K_{A,B}, N, M).$$

The quantity of terms in reference definition is equal to A. The quantity of terms in the answer is equal to B to processing:

$$N = \|A\|, M = \|B\|.$$

At $N = const$ parameter $\omega$ has the following property:

1) At the increase of $K_{AB}$ the value $\omega$ increases.

2) At the increase of quantity of terms the answer of value *w* decreases.

3) If $\frac{\sum_{e \in B} k_t}{M} = 1$, thus $\omega = \omega_{\max}$.

The general parameter *KAB* is equal to the sum of the maximal factors of conformity terms in reference and actual definitions:

$$K_{A,B} = \sum_{e \in B} k_e$$

On the basis of the listed properties the received formula of calculation of factor $\omega$ :

$$\frac{\sum_{e \in B} k_t}{M} = \omega \qquad (2)$$

Function $\eta$ sets the dependence of a relevant parameter $W_{AB}$ from quantity of terms in reference definition and answer. It has the following properties:

If $M = N$ thus $\eta = \eta_{\max}$ .

Function $\eta$ is symmetric to its maximal value.

If $|A - B| \to \pm\infty$, thus $\eta \to 0$ .

On the basis of the listed properties the received formula of calculation of factor

$$\varphi(\Delta) = \frac{h}{\sqrt{\pi}} e^{-h^2\Delta^2}, h > 0. \qquad (3)$$

Parameter h and the range of option values Δ are done experimentally and can vary.

Following parameters are used:

$\Delta = 0{,}1 * (|M - N|)$; h=2,050.

Such values of parameters provide maximum value of the function in point (0; 1) and convergence to factor $\varphi$ = 0,01 at $M$ - $N$ =12 .

Having substituted formulas (2) and (3) in (1), we shall receive

$$W_{A,B} = \omega\eta = \frac{\sum_{e \in B} k_e}{M} \frac{h}{\sqrt{\pi}} e^{-h^2\Delta^2} \qquad . (4)$$

***Conclusion***

The developed model of processing the answers of the one who is taught, on the open type problem allows to connect reference definitions and answers in subject terms which enables to analyze the testeed's answer and to form the estimation of the latter's knowledge.

***Literature***

1. *Литвиненко А.Е., Шевченко А.В.* Проблемы моделирования и алгоритмизации в системах сравнительного анализа электронных текстов // I Міжнар. конф."Математичне та імітаційне моделювання систем МОДС'2006" (Київ, 2006): Тези доп. – К.: ІПММС НАН України, 2006. – С. 107 – 108.

2. *Широков В.А.* Ініформаційна теорія лексикографічних систем. – К.: Довіра, 1998. – 331 с.

3. *Цаленко М. Ш.* Моделирование семантики в базах данных. – М.: Наука, 1989. – 288 с.

4. *Бадьоріна Л.М.* Аналітична модель синонімічної релевантності текстових відповідей у комп'ютерних тестових системах // Вісн. НАУ. – 2006. – №4. – С. 60 – 63.