Then as distance $\rho_{вер}(P_l, P_j)$ between predicates P_l and P_j we shall take size

$$\rho_{вер}(P_l, P_j) = \frac{1}{s}\sum_{i=1}^{s} \rho^i(B_l^i, B_j^i).$$

For all properties of distance formulated in ([5]) are carried out for $\rho_{вер}(P_l, P_j)$.

Acknowledgements

This work was financially supported by the Russian Foundation for Basic Research, project no. 04-01-00858a.

Bibliography

[1] Lbov G.S., Startseva N.G. Decision Logical Functions and Statistical Robustness. Novosibirsk: Izd. Inst. Math., 1999.

[2] Vikent'ev A.A., Koreneva L.N. “Setting the metric and measures of informativity in predicate formulas corresponding to the statements of experts about hierarchical objects”, *Pattern Recognition and Image Analysis*, V. 10, N. 3, (2000), 303--308.

[3] Vikent’ev A.A., Koreneva L.N. “Model approach to probabilities expert statements”, “*Mathematical Methods for Pattern Recognition – 10”, Moscow,* (2001), 25-28.

[4] G.S.Lbov, M.K.Gerasimov. Determining the distance between logical statements in forecasting problems. In: Artificial Intelligence, 2’2004 [in Russian]. Institute of Artificial Intelligence, Ukraine.

[5] Викентьев А.А., Лбов Г.С., Коренева Л.Н. “Расстояние между вероятностными высказываниями экспертов”, *Искусственный интеллект, 2’2002, НАН Украины,* 58-64.

[6] Keisler G., Chang C. Model theory. M.: Mir, 1977.

Author’s Information

Alexander Vikent’ev – Institute of Mathematics, SB RAS, Acad. Koptyuga St., bl.4, Novosibirsk, Russia; e-mail: vikent@math.nsc.ru

ANALYSIS AND COORDINATION OF EXPERT STATEMENTS IN THE PROBLEMS OF INTELLECTUAL INFORMATION SEARCH[1]

Gennadiy Lbov, Nikolai Dolozov, Pavel Maslov

Abstract: *The paper is devoted to the matter of information presented in a natural language search. The method using the statements agreement process is added to the known existing system. It allows the formation of an ordered list of answers to the inquiry in the form of quotations from the documents.*

Keywords: *Search engine, natural language, coordination of statements, semantic graph*

ACM Classification Keywords: *I.2.7 Computing Methodologies – Text analysis*

Introduction

Efficiency of the search engine is determined by the use of various methods of relevant documents revealing and insignificant ones eliminating, as well as methods peculiar to the specific search engine or their certain kind (for example, specialized search engines). Existing search engines are based on the oversight of index databases of

[1] This work was financially supported by **RFBF-04-01-00858**

the processed documents. The purpose is revealing the objects satisfying some criteria. However, such systems do not analyze the sentences of the document for revealing their structure and interrelations.

In the paper an approach to the search engine construction based on the analysis of semantic structure of sentences and their interrelations in the document is offered. Such method allows to do the search considering the logic of sentences thus taking into account the sense of a document. Generally it provides a stricter criterion of significant documents selection, based on accordance to a certain logic structure reflecting the sense of inquiry.

The main issue solved by the offered algorithm consists in doing the logic analysis of sentences for the subsequent search, i.e. in formation of the ranged list of answers to the inquiry in the form of quotations from documents instead of the list of these documents. Intellectuality of this method lies in its simplification of sentences perception and analysis by a person.

This system was developed as a superstructure over an existing search engine ISS2 (Internal Search System) [1]. However, independent functioning of the offered system, for example, for doing the analysis in some interesting documents is also possible. The purpose is in providing search service on local and public network catalogues being storehouses of the information. For the effective search within several storehouses there is an option for aggregation of several search servers to a distributed system. The software contains the means of carrying out a safe remote management as well as all components status analysis done by a search engine.

Selection of Search System

To derive sentences structure the system uses working results a natural text translation system [2]. It describes the methods of translated documents processing for "natural" translation considering specific features of languages. In [5] various systems of parse such as «Dialing»: L. Gershenzon, T. Kobzareva, D. Pankratov, A. Sokirko, I. Nozhov (www.aot.ru); the program of scientific group FtiPL (Institute of linguistics) RGGU (T.Yu. Kobzareva, D.G. Lakhuti, I. Nozhov); LinkParser (www.link.cs.cmu.edu/link). The selection of basis for the developed method was stipulated among other things by a good description and demonstration of system abilities [2]. In this system the analysis is done through several steps, which simplified sequence is as follows: primary, morphological, parse and semantic. Each step uses the results achieved on the previous one. The purpose of primary analysis is in the analysis of the initial document which identifies its sentences, paragraphs, notes, stable statements, electronic addresses etc. As a result the table consisting of some fragments of the initial text and their descriptors is formed. At the following step words morphoanalysis and lemmatization is done, that is each word becomes respectfully attributed with its normal form, morphological part of speech and the set of grammems, defining its grammatical gender, number, case etc. In parse syntactic groups characterized by certain parameters (type of a group, position, parental group) are defined. On the step of the semantic analysis semantic relations describing certain binary links between dependent and operating members are formed. These binary relations are just used in the offered algorithm. Resulting semantic graph characterizes interrelated binary links in the initial text sentences which reflect their logic.

For the solution of the search issue the agreement of statements described in [3] is required on a certain step. So far the resulting sets of relations in the initial text are determined by multiple expert statements whereas in the inquiry text the are defined by a set of certainly true and agreed statements. The algorithm is offered for cases with one or several experts. At first the algorithm agrees the statements of one expert which leads to a number of formulas, and then a process of overall agreement of already agreed opinions of each expert is accomplished. The specific feature if this algorithm is that he identifies absolutely all regularities. Therefore the paper [4] describes an approach to reduction of dimension statements set given on sample with the purpose of the maximal reduction of its dimension at the minimal loss of information.

Co-ordination

Basing on the intermediate results of the system work [2], which are the semantic graphs of the sentences, the logic form is constructed for each sentence. This form is a model in the language of predicates calculus of two variables united in conjunctions. Each of such predicates is an elementary statement. The following problem is to accomplish the procedure of statements agreement in the models on the base of these received models of sentences in the text and inquiry. To do it the predicates of one type are isolated and their set (for each type of predicates) corresponding to the sentences the text is a set of agreed statements whereas their set corresponding to the inquiry is an agreed in advance statement. Considering that each predicate is a part of a sentence model, the crossing of the sets corresponding to agreed predicates of different types is taken. This crossing can be considered as the result of search in the document.

Hypotheses

For the further description of algorithm it is necessary to introduce the following assumptions:

1. Sentences having different predicate structures and different variables in them are considered as the facts of different types supplementing each other.
2. A sentence with the same predicates and with the same (i.e. synonymous) variables are considered supplementing each other, therefore one-type variables are designated by the same letter with an identical index.
3. In case of crossing variables from different predicates we obtain more complicated variant of sense addition.

Each semantic link in the graph defines some type of a two variables predicate. Let's designate with letters Xi, Yi, Zi etc. each predicate variable. As the predicate defines the relation between its variables, the sets of the one-type variables standing in a certain position in the predicate are designated by the same letter with different indexes. Variables in predicates crossing are respectively designated by the same letter with an identical index. Predicates are designated by the name of semantic links. Synonymous words standing in identical positions and in identical predicates are designated by the same variables.

Analysis and Co-ordination

For the sentences and inquiry agreement, inquiry predicates are considered separately. The predicates are picked out one by one from the inquiry and in the same time the predicates of respective types are picked out from the text sentences. Expert statements are agreed with the elementary inquiry predicate which is considered to be agreed in advance.

Decision of the Formulated Task Requires Some Modification of the Algorithm Offered in [3]

Let some statement with known characteristics requires to define its belonging to the certain image. The predicate sets corresponding one or another image are considered separately. The general formal writing of a sentence is done in the form of two-place predicates conjunction. The area of predicate is defined by nominal variables satisfying the list of admissible values. We shall designate T_{ji}^{k} the truthful areas of function and argument variables in the initial sentences inquiry, where i, j, k are the numbers of predicates, statements and the links between argument and function variables, respectively.

As variables are nominal the area of true statements is defined by variables satisfying the list of admissible values. Such list has to be based on a synonyms dictionary. Besides the lists of synonyms it is also necessary for such a dictionary to contain also factors of words affinity. For example, each word from a synonymic group corresponds to the list of synonyms with decreasing weights. To simplify the finding of truthful area it is possible to define the truthfulness of statement on the base of variables satisfying the list consisting of one admissible value. But this list can be expanded with synonyms. Aprioristic probabilities of statements are equal to 1/n (S), where n is a number of statements S.

In the offered system it is enough to accomplish the agreement at a level of one expert as for simplification the analysis is done only in one document, not between many documents. Since predicates are two-placed and variables in them are from different truthful areas, then for the agreement of one expert statement it is necessary to consider separately variables in predicates. Assuming that the statement obtained from the inquiry is true and agreed we define truthful areas from each predicates included in it. The further procedure is done for each separate predicate. T_{pi}^1 is a truthful area of the first variable in the predicate i the inquiry p. T_{pi}^2 is the same for the second variable. The order of choice of the first and the second (the function and the argument) variable can be interchanged for altering the character of agreement, but the choice of the second variable in a predicate as the main one is more logical. Lets designate T_{ji}^1, T_{ji}^2 truthful areas of variables in predicates of the initial text. Respectively, the statement satisfying:

1. $m(T_{ji}^2 \wedge T_{pi}^2) \geq \beta_{r1}$ and $m(T_{ji}^1 \wedge T_{pi}^1) \geq \beta_{r2}$ is true,
2. $m(T_{ji}^2 \wedge T_{pi}^2) \geq \beta_{r1}$ and $\ulcorner\ m(T_{ji}^1 \wedge T_{pi}^1) \geq \beta_{r2}$ is not likely
3. $\ulcorner\ m(T_{ji}^2 \wedge T_{pi}^2\ ) \geq \beta_{r1}$ and $\ulcorner\ m(T_{ji}^1 \wedge T_{pi}^1) \geq \beta_{r2}$ is denying
4. $\ulcorner\ m(T_{ji}^2 \wedge T_{pi}^2) \geq \beta_{r1}$ and $m(T_{ji}^1 \wedge T_{pi}^1) \geq \beta_{r2}$ is denying at a choice of the second variable as the main, and not likely in other case. β_{r2} is a parameter.

Thus we receive sets of statements: ω_1 - not likely, ω_2 - true, Ω - denying.

The following steps of the one expert statements agreement are similar to described in [3].

Ranging

Let's designate Nsi the number of all predicates in a sentence, Nsoi the number of agreed predicates of a sentence, Nr the number of predicates in an inquiry. Then for determination of the sentences relevance we have to calculate the ratio:

$$k = \frac{(N_{so_i})^2}{N_{s_i} \cdot N_r}$$

As a result we receive a set of agreed statements for the first type of predicates. The procedure of agreement is repeated separately for all other predicates and we obtain the sets of agreed statements of different type, each of which defines the sentence. Finding the crossing of all these sets we receive the set of sentences satisfying to the inquiry. The outcoming set forms the result in a usual language considering text paragraphs and document headings. Thus the trial algorithm of significant sentences allocation in the text is obtained; it reflects the first and the second assumption about the usual language.

Example (in Russian)

The simple text: **Рыбак собрался ловить рыбу. Рыбак взял удочку и ведро. Рыбак забросил крючок в реку и стал ждать. По реке проплывала лодка.**

And simple inquiries: **1. Рыбак ловит рыбу. 2. Рыбак взял наживку. 3. Мокрый рыбак.**

The sentence graphs constructed by the system [1] look as follows:

Sentences in the text:

The formula of the 1st sentence: $SUB(z_1, x_1) \cup OBJ(z_1, y_1)$

The formula of the 2nd sentence: $SUB(z_2, x_1) \cup OBJ(z_2, y_2) \cup OBJ(z_2, y_3)$

The formula of the 3rd sentence: $SUB(z_3, x_1) \cup OBJ(z_3, y_4) \cup TP(z_3, t_1) \cup SUB(z_4, x_1)$

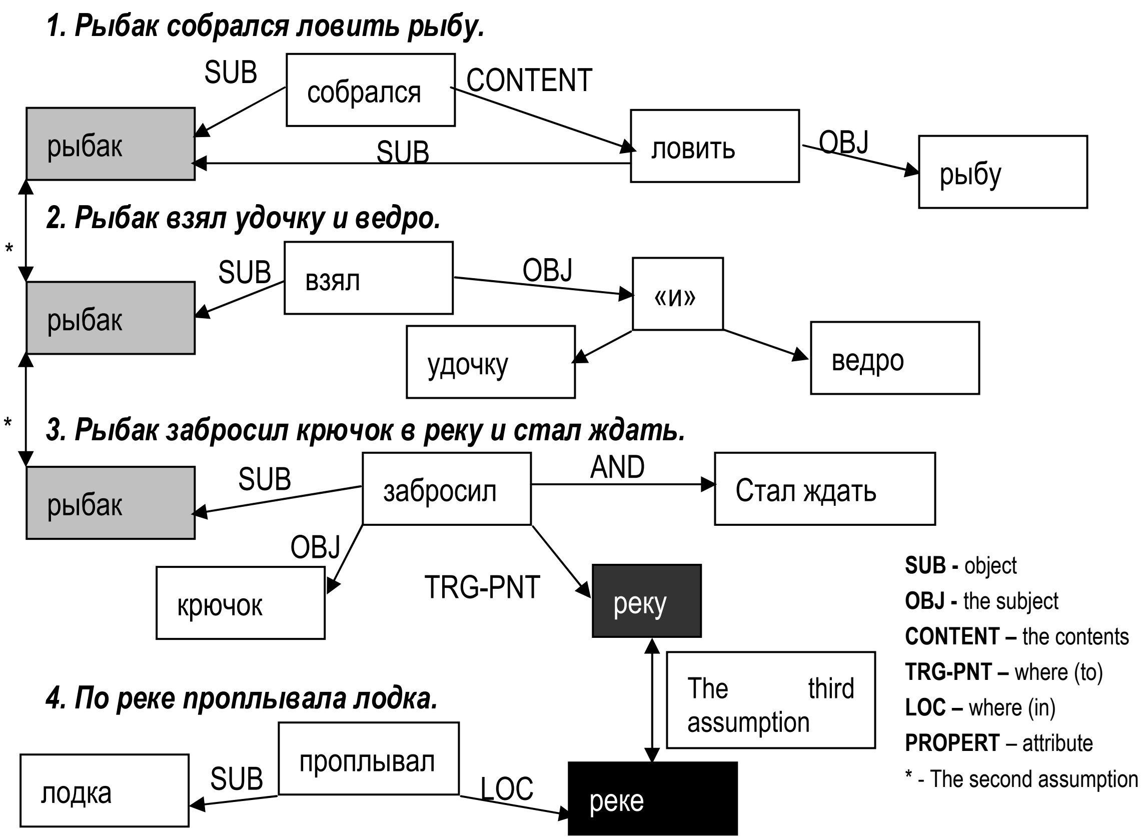
The formula of the 4th sentence: $SUB(z_5, x_2) \cup LOC(z_5, l_1)$

Sentences of the inquiry:

The formula of the 1st sentence: $SUB(z_1, x_1) \cup OBJ(z_1, y_1)$

The formula of the 2nd sentence: $SUB(z_2, x_1) \cup OBJ(z_2, y_5)$

The formula of the 3rd sentence: $PRT(x_1, p_1)$



SUB - object
OBJ - the subject
CONTENT – the contents
TRG-PNT – where (to)
LOC – where (in)
PROPERT – attribute
* - The second assumption

Conclusion

For the inquiry 1 the structure of inquiry and predicate variables are similar to one of the text sentences, therefore at least one sentence is in complete agreement with such inquiry. In the second inquiry there the structure is concurrent, variables in a predicate are distinct - the full agreement is not present, therefore the ranging will show only 25%, whereas a simple phrase «рыбак взял» will show 100%. The third inquiry contains the single predicate PRT designating the property of an object. Such predicate is not present in the text, therefore the algorithm agrees nothing. In other words, the sense of inquiry is not crossed with the sense of the text.

Bibliography

[1] P.P. Maslov. Designing Materials of the All-Russian scientific conference of young scientists in 7 parts. Novosibirsk: NGTU, 2006. Part 1. - 291 p. // pp. 250-251

[2] Automated text processing "DIALING" // www.aot.ru

[3] G.S. Lbov T.I. Luchsheva. The analysis and the coordination of expert's knowledge in problems of recognition // 2'2004, NAS of Ukraine, pp. 109-112.

[5] Nozhov I. The Parse // http://www.computerra.ru/offline/2002/446/18250/

Authors' Information

Gennadiy Lbov - SBRAS, The head of laboratory; P.O.Box: 630090, Novosibirsk, 4 Acad. Koptyug avenue, Russia; e-mail: lbov@math.nsc.ru

Nikolai Dolozov - NSTU, The senior lecturer, Cand.Tech.Sci.; P.O.Box: 6300092, Novosibirsk, 20 Marks avenue, Russia; e-mail: dnl@interface.nsk.su

Pavel Maslov - NSTU, The post-graduate student of of FAMI; P.O.Box: 6300092, Novosibirsk, 20 Marks avenue, Russia; e-mail: altermann@ngs.ru