INFORMATION SEARCH BASED ON ANALYSIS OF EXPERTS STATEMENTS[1]

Gennadiy Lbov, Nikolai Dolozov, Pavel Maslov

Abstract: The paper describes natural language processing. The proposed method, which uses statement coordination principle, is implemented to the described search system. This method allows to compile an ordered list of answers to the inquiry in the form of quotations from the document.

Keywords: natural language processing, coordination of statements, information search.

ACM Classification Keywords: H.3.3 Information Search and Retrieval

Introduction

The specified approach defines criterion of selection of significant sentences of documents, based on accordance to a certain logic structure reflecting the sense of inquiry, also allow revealing relevant documents in the order of inquiry level accordance to the documents.

Coordination of statements

Basing on outcomes of NLP (see below) 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. Let X_i, Y_i, Z_i etc. be each predicate variable.

The set (for each type of a predicate), corresponding sentences of the text is a variety of coordinated statements, and a set corresponding inquiry is beforehand coordinated statement. By quantity of the coordinated predicates the level of its accordance to inquiry is defined, being based that each predicate is a part of model of the sentence.

Let some statement with known characteristics requires to define its accordance to inquiry [1]. The general formal writing of a sentence is done in the form of two-place predicates conjunction. We shall designate T_{ji}^{k} as area of the validity 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. As such list the dictionary of synonyms is used.

As predicates two-place and their variables are defined on different areas of the validity for the coordination of statements, it is necessary to consider variables in predicates separately.

For each predicates contained in the statement the areas of validity are defined: 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. Let us designate T_{ji}^{1}, T_{ji}^{2} as truthful areas of variables in predicates of the input text. Respectively, the statement satisfying:

[1] The work was supported by the RFBR under Grant N07-01-00331a.

$$1.\quad \frac{\mu(T^2_{j1} \cap T^2_{\rho 1})}{\mu(T^2_{j1} \cup T^2_{\rho 1})} \geq \beta_{r2} \quad and \quad \frac{\mu(T^1_{j1} \cap T^1_{\rho 1})}{\mu(T^1_{j1} \cup T^1_{\rho 1})} \geq \beta_{r1} \quad \text{- true}$$

$$2.\quad \frac{\mu(T^2_{j1} \cap T^2_{\rho 1})}{\mu(T^2_{j1} \cup T^2_{\rho 1})} \geq \beta_{r2} \quad and \quad \frac{\mu(T^1_{j1} \cap T^1_{\rho 1})}{\mu(T^1_{j1} \cup T^1_{\rho 1})} < \beta_{r1} \quad \text{- not likely}$$

$$3.\quad \frac{\mu(T^2_{j1} \cap T^2_{\rho 1})}{\mu(T^2_{j1} \cup T^2_{\rho 1})} < \beta_{r2} \quad \text{- contradictious}$$

$$k = \frac{(N^i_{so})^2}{N^i_s \cdot N_r}$$

Where μ is a cardinal number of the argument; β_{rq} is a parameter is defined experimentally in [1], N^i_s is the number of all predicates in a sentence, N^i_{so} the number of the coordinated predicates of a sentence, N_r the number of predicates of inquiry.

To define the accordance of sentence to inquiry it is necessary to calculate ratio k.

Natural Language Processing

The specified approach of selection of significant sentences of documents has been realized programmatically in search system Internal Search System3 further ISS3 [2] (In this system some technologies of known system [3] are used), which operation scheme is showed in fig. 1, providing search service of documents on local and sharer network resources.

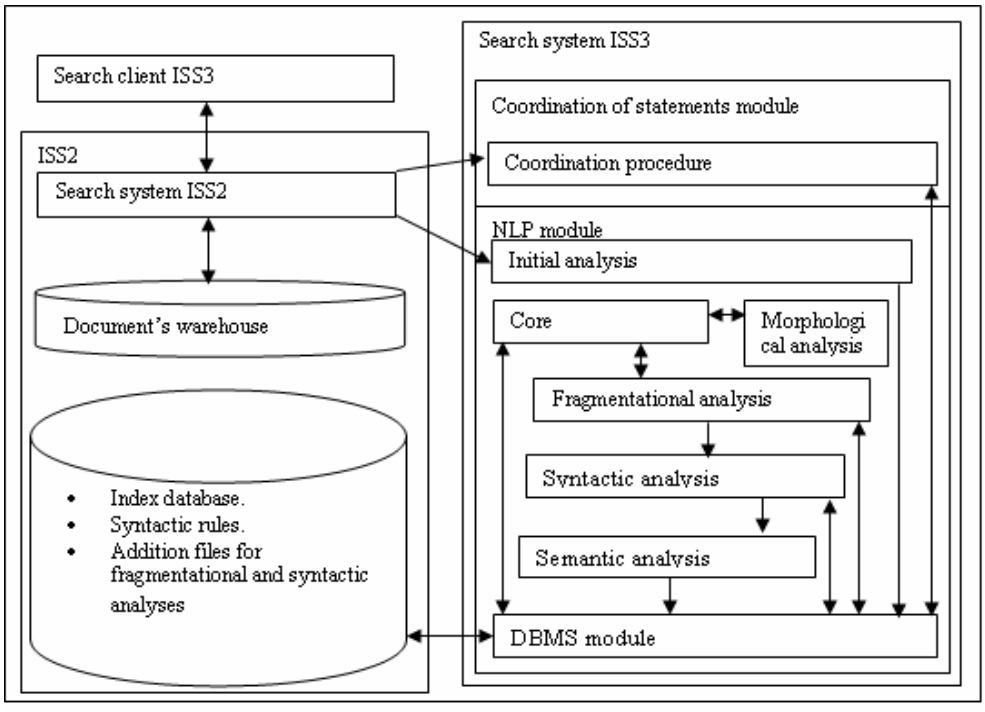

Figure 1. Operation scheme of ISS3.

To perform the subsequent procedure of the coordination of the statement, the developed system represents an input text as a sets of syntactic relations for each sentences of the text. It is achieved by the multilevel natural language processing realizing initial, morphological, fragmentational, and syntactic.

Stages of NL (brief description):

1) Initial text analysis process input text and then forms two separate tables. The first describes elements of the text and their arrangement in the text, and the second defines interrelation of fragments in the input text.

2) At a stage of the morphological analysis, for each lexeme a set of lemmas with attributes is created. Each lemma is represented as a normal form of a word, and attributes - a set of descriptors (a part of speech, number, case etc.).

3) Fragmentational analysis performs derivation of fragment from the input text. Fragments are the main and dependent clauses in structure of complex one, participial, adverbial participial and other isolated turns.

4) The purpose of syntactic parse is automatic construction of a functional tree of a phrase. Syntactic parse process outcomes of morphological and fragmentational analyses.

As a result NLP forms two separate tables for each input document:

1) The content of the document. The table of syntactic relation sets for each sentence of the input document, rows in which describe type and components of syntactic relations.

2) Structure of the document. The table containing the descriptions of the structure of the document (paragraphs, headers, etc.), derived at a stage of the fragmentational analysis and necessary to form the reply to inquiry.

Results

The results are showed below, illustrates processing of various search inquiries in the textual document (fig. 2). For convenience the information of inquiries is contained only in one document, however the system has no limitation for quantity of documents:

Тестовый файл.

Человек подошел к столу. Человек взял лист и ручку. Человек начал писать текст. На столе лежала черная кошка. Кошка заметила человека. Кошка подбежала к человеку и села на лист.

Рыбак собрался ловить рыбу. Рыбак взял удочку и ведро. Рыбак забросил крючок в реку и стал ждать. По реке проплывала лодка.

Figure 2. Test document.

In accordance with the inquiry "**человек взял ручку**" the title and the first paragraph are extracted:

Тестовый файл.

Человек подошел к столу. Человек взял лист и ручку. Человек начал писать текст. На столе лежала черная кошка. Кошка заметила человека. Кошка подбежала к человеку и села на лист.

The achieved level of conformity of the text to the inquiry, equal to 0.67 is defined by the presence of another syntactic relation "**взять листок**" in the fragment "**Человек взял лист и ручку**".

In accordance with the inquiry "**черная собака**" there is no results returned because the property "**черная**" is related with the element "**кошка**".

These simple tests have shown that the system extracts sentences in accordance with their syntactic structures and the inquiry structure, increasing the search precision in comparison with the offered system without natural language processing.

Conclusion

Approach for performing a search in the textual documents, based on the analysis of syntactic structure of sentences is offered. It allows to extract significant syntactic relation from the text in accordance with the syntactic structure of inquire. As a result of performance of algorithm, sets of the coordinated statements for all types of predicates are formed, each of which describes a certain fragment. To define conformity of the sentence to inquiry, the ratio specified above is calculated.

Bibliography

1. G.S. Lbov, T.I. Luchsheva. The Analysis and Coordination of Expert Knowledge in Problem of Recognition // 2'2004, NAS of Ukraine, pp. 109-112
2. P.P. Maslov. Proceedings of All-Russian scientific conference of young scientists in seven parts. Novosibirsk: NGTU, 2006. Part 1. – 291p. // pp. 250-251
3. The project DIALING // www.aot.ru

Authors' Information

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

Nikolay Dolozov – NSTU, The associate professor, candidate of science; P.O.Box: 630092, Novosibirsk, 20 Marks avenue, Russia; e-mail: dnl@interface.nsk.su

Pavel Maslov – NSTU, post-graduate student of Faculty of Applied Mathematic and Computer Science; P.O.Box: 630092, Novosibirsk, 20 Marks avenue, Russia; e-mail: mpp84@rambler.ru