INFORMATIONAL MODEL OF NATURAL LANGUAGE PROCESSING

Aleksandr Palagin, Viktor Gladun, Nikolay Petrenko, Vitalii Velychko,
Aleksey Sevruk, Andrey Mikhailyuk

Abstract: The formal model of natural language processing in knowledge-based information systems is considered. The components realizing functions of offered formal model are described.

Keywords: natural language processing.

ACM Classification Keywords: I.2.7 Natural Language Processing - Text analysis

Architecture of modern knowledge-based information systems (KIS) with natural language knowledge representation and processing explicitly includes ontological constituent, that in general can be interpreted as a conceptual knowledgebase. Such a knowledgebase is represented as a directed graph, which vertexes are represented as frames describing concepts, and arcs are the set of conceptual relations connecting among themselves concepts. Other important feature of the specified architecture is division and separate processing of the first and second order semantics [1], that generally means separating internal language and non language processing [2] and transition to the formal-logic representation of the source text.

The specified architecture features of modern KIS transform traditional natural language texts (NLT) processing model to the following formal model

$F = \langle T, W, SS^1, O, S^2, I \rangle$, where

T is a set of processed NLT;

W is a set of word forms contained in the T;

SS^1 is a set of first-order syntactic-semantic structures, describing T;

O is a set of ontological structures, converting the sets W and SS^1 into the S^2;

S^2 is a set of second-order semantic structures, describing the set of the T's scripts;

I is a set of information-code representations of the S^2.

Let us describe the objects of the formal model.

The T set represents the corpus of the NLT described with business, scientific and technical styles.

The $W \to SS^1$ chain in its classical sense represents grammatical analysis of NLT. In contrast to traditional linear and strong coded analyzing methods we use mixed one. The gist of its that in lexicographical database full W set is represented in the tables of two types: the tables of lexemes with corresponding morphological, syntactic, and semantic characteristics and the tables of inflections for all full-meaning varying parts of speech.

At the same time the algorithms of lexemes paradigm generation are simple; the lexical tables contains lexemes' stems and corresponding codes for selecting records from the tables of inflections. Non inflection changes are considered by corresponding algorithms.

The described grammatical analysis structure univocally corresponds to an effective mapping of functional operators into hardware realization, which is in particular based on PLIS (programmable logical integral schemas).

The set O of ontological structures in an ideal represents language-ontological pattern of the world, described in [1, 3].

The SS^1 set is iteratively formed and interpreted by the syntactic-semantic subsystem like “Konspekt” [4]. The main operation of syntactic-semantic analysis is recognizing of syntactical and semantic relations, linking text words. The recognition of links between notional words is carried out by the analysis of inflections and prepositions basing on lexical models without explicit traditional grammar rules using. For each sentence in source text parsing tree is built. The solving of semantic ambiguity is carried out in the way of calling to the set of ontological structures O. Basing on built semantic trees categorial net is built. The net represents semantic space S^2 of the text.

As a computer representation of such space it is convenient to use the growing semantic net for set of information code representations I, that is organized as pyramidal net, which receptors are corresponding to the names of objects, classes of objects, properties, states, actions, relations, semantic cases, modifiers [5].

The information conversion chain $T \to W \to SS^1$ and $O \to S^2 \to I$, per se, represents (accordingly) base procedures of analysis and understanding NLT, which interpretation tools are grammatical and semantic processors.

In applications for searching and processing large volume of text documents expedient is to use knowledge-based search system [6] providing initial and final stages of documents processing – searching in the Internet and saving documents in a database in the form of their synopses generated by “Konspekt” subsystem.

The described model of natural language texts processing in knowledge-based information system, containing “Konspekt” subsystem as a part, is a promising line of development ontological-based informational systems that make active use of ontology of natural language lexicon.

Bibliography

1. Palagin O.V., Petrenko M.G. Model’ kategorial’nogo rivnya movno-ontologichnoi kartyny svitu // Matematychni mashyny I systemy – 2006. – N 3. – pp.91-104.
2. Rubashkin V.Sh. Predstavlenie I analiz smysla v intellektual’nyh informatsionnyh sistemah. – M.: Nauka, 1989. – 191p.
3. Palagin A.V. Organizatsiya I funktsii “yazykovoy” kartiny mira v smyslovoy interpretataii EYa-soobshcheniy // Information Theories and Application. – 2000. – Vol. 7, N 4. pp.155-163.
4. Gladun V.P., Velichko V.Yu. Konspektirovanie estestvenno-yazykovyh tekstov. Proceedings of the XI-th International Conference “Knowledge-Dialogue-Solution”(KDS’2005).- Varna, Bulgaria.-2005.- 5. pp.344-347 vol.2.
5. Gladun V.P. Planirovanie resheniy. – Kiev: Naukova dumka, 1987. – 168p.
6. Sevruk O.O., Petrenko M.G. Znannya-orientovana poshukova systema na osnovi movno-ontologichnoi kartyny svitu // Tezy dopovidey XIII mizhnarodnoi konferencii “Avtomatyka-2006”. – Vinnytsya. – 2006. – 25-28 veresnya. – p.413.

Authors' Information

Aleksandr Palagin – V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: palagin_a@ukr.net

Victor Gladun – V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

Nikolay Petrenko, – V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: petrng@ukr.net

Vitalii Velychko – V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: vitaly@aduis.kiev.ua

Aleksey Sevruk, Andrey Mikhailyuk – V.M.Glushkov Institute of cybernetics of NAS of Ukraine, Prospekt akad. Glushkova 40, 03680 Kiev, Ukraine; e-mail: petrng@ukr.net