

# The methodology of semantic analysis for extracting physical effects

M A Fomenkova<sup>1</sup>, V A Kamaev<sup>1</sup>, D M Korobkin<sup>1</sup>, S A Fomenkov<sup>1</sup>

<sup>1</sup> Volgograd State Technical University, 28, Lenina Ave., Volgograd, 400005, Russia

E-mail: [dkorobkin80@mail.ru](mailto:dkorobkin80@mail.ru)

**Abstract.** The paper represents new methodology of semantic analysis for physical effects extracting. This methodology is based on the Tuzov ontology that formally describes the Russian language. In this paper, semantic patterns were described to extract structural physical information in the form of physical effects. A new algorithm of text analysis was described.

## 1. Introduction

Nowadays, the semantic analysis is one of the most topical and growing areas in computer science. In this work, the semantic analysis is considered as a tool for extracting structured physical information in the form of physical effects (PE). Today the task of identifying PE descriptions in the science texts to supplement the database of PE is the information basis for the new technical solutions development. It is an important and actual task.

Today, there is the system of identifying PE descriptions. It was realized using the semantic analyzer “Semantix” [1]. The system and the semantic analyzer that are described in this work were realized using the Tuzov ontology.

## 2. Program systems and tools for semantic analysis

Analysis of the effectiveness of software systems for semantic analysis is shown in Table 1, Table 2.

**Table 1.** The comparison table of efficiency of software systems

Program system	An algorithm of target entities identifying	The system's flexibility (the ability to customize settings, adding new entities)	License	Accuracy extracting (1 - 75% 2 - 65 - 75%, 3 - 65% less)	Completeness (1- more than 60%, 2 – less than 60%)
Extracting facts from the text files, RCO [2]	Patterns search	The parameters are hard-coded	Close	1	2
Attensity Text Analytics [3]	Neural network technology	The parameters are hard-coded	Close	2	1
NetOwl Extractor [4]	Neural network technology	The ability to add new entities	Close	1	1
IOFFE [1]	Patterns	The ability to add new	Open	3	2



search		patterns		
Table 2. Semantic analysis systems				
Software	Russian support	License	Base technology	Flexibility (focus on the subject area, the ability to customize, code availability)
Stanford nlp [5]	No	Open	Machine Learning Technologies	No
Malt parser [6]	Yes	Open	Machine Learning Technologies	No
Link grammar parser [7]	No	Open	Relations grammar	No
AGFL [8]	Yes	Open	AGFL-grammar	No
Tomita parser [9]	Yes	Open	CF-grammar and key words dictionaries	No

As is seen from the tables, most of the systems are flexible enough for adjustment to the subject area. It is necessary to create a new approach to work with the physical science text.

### 3. Semantic analyses for extracting structured physical knowledge in the form of physical effects of the Tuzov ontology

Ontology of the Russian language [10] is a formal description of the Russian language proposed by V.A. Tuzov (semantic roles are defined on the basis of semantic classes and morphological information). The base of computer semantics uses the functional model of the language:

- 1) the language is an algebraic system  $\{f_1, f_2, \dots, f_n, M\}$ , where  $f_i$  - basic functions of a language and  $M$  - the structure of a language, which is a set of basic concepts  $m_1, \dots, m_r$  and their hierarchy;
- 2) each language sentence can be represented as the superposition of basic functions  $f_i$ , and every word of the language is expressed by these functions. The exception - basic concepts  $m_j$ , belonging to  $M$ ;
- 3) grammar is inextricably linked with the language semantics, which basis is the semantic dictionary that describes more than one hundred thousand lexical units (words and phrases), and each word is described as a semantic formula consisting of basic functions.

**Table 3.** Some basic functions

Function	Description
Caus (x,y)	x is the reason y
Loc(x,y)	x is in y

The dictionary entry contains a header word and its interpretation in the semantic language:

EXPOSURE \$ 15142 (Caus\_o (AGENT: SOMETHING \$ 1 ~ Gen, Lab (OBJECT:! Acc, LOCATION: Prep)))

#### 3.1. Physical Effect

There is a formal description of a physical effect [11]. It consists of the following structure (A, B<sub>1</sub>, B<sub>2</sub>, C): A - an input stream of matter, energy or signals; B<sub>1</sub> - the initial state of a physical object B; B<sub>2</sub> - the final state of the physical object B; C - the flow of matter, energy or signals;

At the department, CAD, of Volgograd State Technical University, the physical effects fund was established. This fund has more than 1,200 PE descriptions.

#### 3.2. The model of structured physical information representation in natural language texts

The model of structured physical information representation in natural language texts was created to retrieve descriptions of physical effects [12]:

$$M_{PE} = \langle C, D, B, R_C, R_B \rangle, \quad (1)$$

where C - the set of predicates (relations), to describe the PEs in the text,  $c_i \in C$ ;

D - semantic roles and case arguments in predicates  $D_i \subset D$  - a list of roles / arguments of cases agreed with predicate  $c_i$ ;  $d_j \in D_i$ ;

B – a number of elements to describe PE (A, B, C),  $B_k \in B$ ,

where  $B_k \in \{\text{input (A), output (C), object (B)}\}$ ;  $\forall c_i \in C \exists d_j \in D_i [d_j \xrightarrow{\text{def}} B_k]$  - the operator that is associated with the role / case of the  $d_j$  argument at the  $c_i$  predicate;

$R_C$  – the relationship on  $C \times D$ , pair  $(c_i, d_j) \in R_C$  uniquely identifies the item of the PE description, consistent with the role of predicate / case  $d_j$

$R_B$  - the relationship at  $R_C \times B$ , pair  $((c_i, d_j), B_k) \in R_B$  defines a set of concepts corresponding to the element of PE description  $b_k$ ,  $b_k \in B_k$ .

### 3.3. The algorithm of semantic analysis for extracting the semantic roles

The algorithm of semantic analysis for extracting the semantic roles of Agent, Object, Place arguments is shown in Figure 1.

As a result of domain analysis, we can see the presence of semantic ambiguity. Semantic ambiguity arises because sometimes it is possible to match one semantic role of the Tuzov ontology (role "Agent", "Object", "Place") and several elements of the physical effect description ("Input", "Output", "Object" of physical effect).

To remove this kind of semantic ambiguity, one software module was developed. It is required for matching the existing descriptions of physical effects and semantic roles described in the templates on the basis of the Tuzov ontology.

The approach is to compare the semantic roles, elements of the physical effects descriptions and specific words, which may show the ambiguity. It was realized using the field (PE Description in the natural text) in the database of physical effects [13].

<Semantic role> - <Item of PE description>.

On the basis of statistics of this bunch, a template for the PE extracting is formed for each predicate.

The algorithm for constructing correspondences of semantic roles in the ontology and elements of the PE descriptions for the subject area of the predicates is shown in Figure 2. Thus, the overall physical effects extraction algorithm is shown in Figure 3.

## 4. Results

The main performance indicators are the accuracy and completeness of extraction.

Accuracy is characterized by the number of correctly retrieved elements from the total number of elements of PE description.

$$P = N_k / N_n \quad (2),$$

P - extraction accuracy of the PE,  $N_k$  number of correctly retrieved elements,  $N_n$  – the number of elements found in the text.

Completeness of the PE elements extraction indicates the amount of elements relative to the total number of PE elements in the text descriptions.

$$R = N_n / N \quad (3),$$

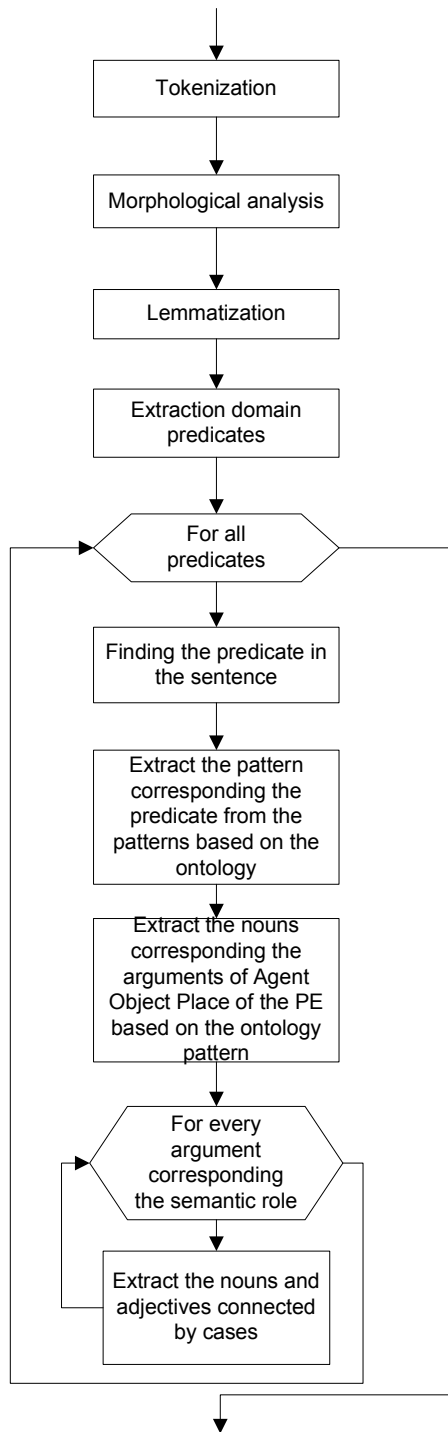
R - completeness of extraction,  $N_n$  – the number of PE elements found in the text. N – a total number of elements of the PE elements in the text.

F-measure is calculated as 4:

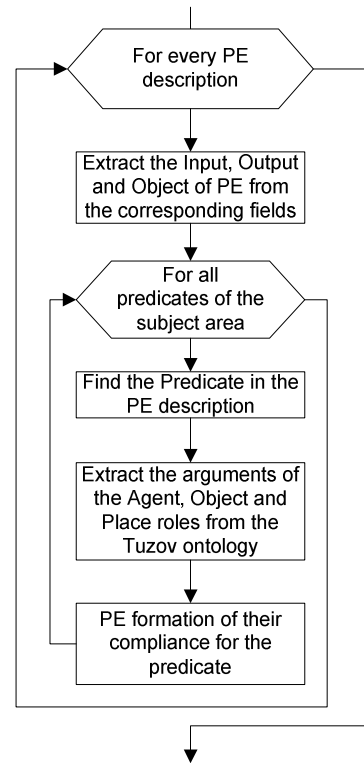
$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad (4),$$

$$\beta^2 = \frac{1 - \alpha}{\alpha} \quad (5),$$

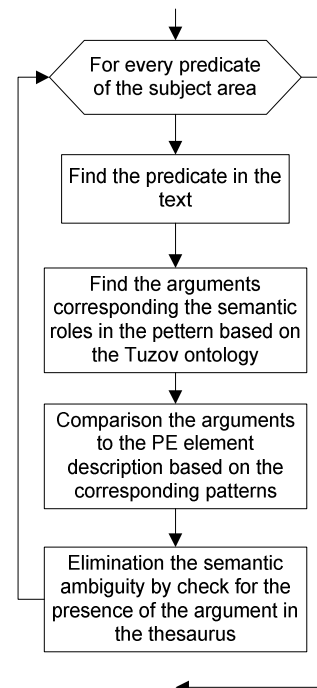
where  $\alpha = 0.3$



**Figure 1.** The algorithm of semantic analysis for extracting the semantic roles of Agent, Object, Place arguments.



**Figure 2.** The algorithm for constructing correspondences of semantic roles in the ontology and elements of the PE descriptions for the subject area of the predicates.



**Figure 3.** The extraction algorithm of overall physical effects.

Tests were carried out on the basis of the database of physical effects developed at the department of VSTU. 100 physical effects were selected. Tests were conducted on the base of physical effects descriptions in the database and compared with the fields “Input”, “Output” and “Object” of the physical effects. Tests were also conducted by the example of 31 patent documents (field “description”).

The results were compared with the results of program IOFFE [1] on the basis of the semantic analyzer “Semantix”.

The results of the effectiveness comparing are shown in Tables 4 and 5.

**Table 4.** Analysis of efficiency using the PE database

	Accuracy (%)	Completeness (%)	F-measure
Our system	68	59	61.5
IOFFE	57	53	54.14

**Table 5.** Analysis of the effectiveness using the patent array

	Accuracy (%)	Completeness (%)	F-measure
Our system	53	46	47.9
IOFFE	46	41	42.38

Efficiency Analysis showed that the developed system improves the efficiency to 4% for accuracy and completeness – to 7%.

Sample. Patent description: “The photoelectric conversion element may be a photodiode having a p-n junction or a pin junction, a phototransistor, or the like. When the incident light hits the semiconductor junction of the cell, this light leads to the appearance of the photoelectric effect, in which electric charges arise.” [14]

PE Input: “Light, any other electromagnetic radiation (energy - eV)”.

PE Output: “The electric charge (electron emission), (J)”.

PE Object: “Photoconductive material (photoconductor)”.

The results of the program show the results of the physical effect elements extraction:

PE Input: “Light”. PE Output: “electric charge”. PE Object: “phototransistor”.

## 5. Conclusion

The method described in this article allowed increasing efficiency of the PE elements extracting. The semantic analyzer based on the Tuzov ontology was created to increase the accuracy and completeness of the method. The approach was tested on the PE database and the patent array.

## 6. Acknowledgments

This research was financially supported by the Russian Fund of Basic Research (grants No. 15-07-09142 A, No. 16-07-00534 A) and the Russian Ministry of Education in scope of the base part (project 2586 of task N 2014/16).

## References

- [1] Korobkin D M, Fomenkov S A, Kamaev V A, Fomenkova M A 2016 Multi-agent model of ontology-based extraction of physical effects descriptions from natural language text *Information Technologies in Science, Management, Social Sphere and Medicine' (ITSMSSM 2016)* (Atlantis Press) pp. 498-501
- [2] Pleshko V V and Ermakov A Y 2009 Semantic interpretation in text analysis computer systems *Information Technologies* **155** (7)
- [3] Taylor J 2011 First Look – Attensity 5.5 *Decision, management, solutions*.

- [4] Krupka G R, Hausman K.I 1998 Description of the NetOwl TM Extractor System as used for MUC-7 (1998) *Proceedings of the MUC-7*.
- [5] Manning C D, Raghavan P, Schütze H 2008 *Introduction to information retrieval* (Cambridge University Press)
- [6] Nivre J, Hall J and Nilsson J 2006 MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)* (Genoa, Italy) pp 2216-2219
- [7] Tukeyev U A, Melby A K, Zhumanov Zh M 2011 Models and algorithms of translation of the Kazakh language sentences into English language with use of link grammar and the statistical approach *Proc. of IV Congress of the Turkic World Math. Society* (Baku) p 474
- [8] Azarova I 2002 The matching of AGFL subcategories to Russian lexical and grammatical groupings *Proceedings of the Second AGFL Workshop on Syntactic Description and Processing of Natural Language*.
- [9] Ogogrodnik P B, Serebryannaya L V 2014 Text analysis with Tomita parser *Proceeding of The International Conference, BSUIR, Minsk, 29th October 2014* pp 230-231
- [10] Tuzov V A 1998 *Computer Linguistics. St. Petersburg, St. Petersburg State University Press*
- [11] Korobkin D, Fomenkov S, Kolesnikov S., Lobeyko V, Golovanchikov A 2015 Modification of Physical Effect Model for the Synthesis of the Physical Operation Principles of Technical System *Creativity in Intelligent Technologies and Data Science. CIT&DS 2015: First Conference* ( Springer International Publishing) pp 368-378
- [12] Korobkin D M, Fomenkov S A, Kolesnikov S G 2014 Method of Ontology-Based Extraction of Physical Effect Description from Russian Text *Knowledge-Based Software Engineering : Proceedings of 11th Joint Conference, JCKBSE 2014* (Springer International Publishing) pp 321-330
- [13] Fomenkov S A, Kolesnikov S G, Korobkin D M, Kamaev V A, Orlova Y.A 2014 The Information Filling of the Database by Physical Effects *Journal of Engineering and Applied Sciences* **9 (10-12)** 422-426
- [14] Patent US 20100051095 A1 “Hybrid Photovoltaic Cell Using Amorphous Silicon Germanium Absorbers With Wide Bandgap Dopant Layers and an Up-Converter”