Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

New Trends
in
Classification and Data Mining

ITHEA
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0042-9

C\o Jusautor, Sofia, 2010

LSPL-PATTERNS AS A TOOL FOR INFORMATION EXTRACTION FROM NATURAL LANGUAGE TEXTS

Elena Bolshakova, Natalia Efremova, Alexey Noskov

***Abstract**: The paper describes main features of formal lexico-syntactic pattern language (LSPL) proposed for specification of linguistics information about NL expressions automatically extracted from Russian texts. A fully-implemented procedure for matching LSPL-patterns with text are presented, as well as developed programming tools for extraction of phrases specified by the patterns. Two applications of the language and the tools are discussed: terminological analysis of scientific texts and processing of NL sentences for question answering. LSPL-patterns developed for these applications are briefly characterized.*

***Keywords**: information extraction from NL texts, lexico-syntactic patterns, matching procedure, automatic terms recognition and extraction, analysis of NL phrases for question answering.*

***ACM Classification Keywords**: I.2.7 [Artificial Intelligence]: Natural language processing – Text analysis*

Introduction

Information extraction (IE) from natural language (NL) texts is one of the most important problems of modern computer linguistics and artificial intelligence [Grishman R., 2003]. Traditionally, IE aims at identification in texts of selected types of entities (names and titles), relations, or events [Hearst, 1998; Boudin F. et al., 2008]. As a rule, IE applications are based on shallow syntactic analysis of the text and exploits both heuristics and linguistics information about items to be automatically recognized in it.

Among programming tools commonly used to create IE applications we should point out well-known system for text engineering GATE [Bontcheva K. et al., 2002], and analogous systems, such as Ellogon [Petasis G. et al., 2002]. They are rather universal and propose special formal languages for annotating text segments and describing annotation transformations. As a consequence of their universality, the annotation languages require skilled users to develop application for a new problem domain and different natural language. The languages have no built-in devices for describing specific linguistics properties, in particular, grammatical agreement, which are typical for such flexional natural languages, as Russian.

So far, a more specific language called LSPL (Lexico-Syntactic Pattern Language) was proposed for formal specification of linguistics information about NL expressions to be automatically recognized within Russian texts [Bolshakova et al., 2007]. The key language structure is lexico-syntactic pattern that describes certain NL phrase - its words and other constituent elements, as well as their morphologic and syntactic properties. An example of pattern for the English phrase for topic actualization is *"let us consider" NP* with *NP* denoting a noun phrase. In general case, LSPL-pattern combines both lexical and syntactic information about the described phrase, and thereby LSPL language is convenient to formally specify a wide range of common scientific expressions used for automatic discourse analysis of scientific and technical texts [Bolshakova, 2008].

In contrast to the annotation languages, LSPL was created as a linguistically-oriented and purely declarative formal language that is easy to use for:

- formal specification of a wide variety of NL phrases (noun-noun and verb-noun combinations, adverbial and participle phrases, etc.) within information extraction systems based on surface syntactical analysis of texts;

- user's queries for text browsers performing search of NL phrases and expressions specified by their lexico-syntactic patterns.

Elaborated LSPL language includes devices for specifying within patterns both particular word forms, lexemes and arbitrary words of particular part of speech (POS), as well as their morphological attributes and conditions of grammatical agreement. The latter presents an important language feature proposed specially for description of Russian noun phrases.

For proposed LSPL language a procedure for matching a pattern with a given Russian text was developed, as well as corresponding programming tools for recognition and extraction of phrases specified by LSPL-patterns. Making use of the tools, we investigated two different applications: automatic extraction of terms in scientific texts and analysis of NL sentences for question answering. The former is relatively well studied for English and French texts [Jacquemin C., 2003]. Aiming at terminological analysis of Russian scientific and technical texts, we created a representative set of LSPL-patterns that describes linguistics properties of multi-word term occurrences in texts and then we experimentally studied these patterns. Another developed application of LSPL language (and its supporting tools) is analysis of NL queries in question-answering system. The language proved to be convenient for both applications.

The paper starts with an overview of LSPL language; basic principles of the matching procedure for LSPL-patterns are overviewed as well. Their applications for automatic terms extraction and NL query analysis in question-answering systems are then discussed and conclusions are drawn. Since LSPL language was primary proposed and used for formalizing Russian phrases, Illustrative examples are given mainly for Russian.

LSPL language and LSPL-patterns

Lexico-syntactic pattern formalizes structure and properties of some NL phrase (noun phrase or verb-noun phrase, etc.). The pattern has a name and a body, the latter is separated by symbol of equality. Pattern body includes elements describing constituents of the phrase to be formalized. The order of the elements corresponds to the order of constituents in the phrase. As a rule, the pattern also specifies conditions of grammatical agreement for its elements. For example, the pattern NP = A N <A=N> has the name AN and the body that consists of elements A, N and agreement conditions <A=N> . This pattern describes a simple noun phrase: adjective (A) and noun (N) that are fully grammatically agreed (i.e. in case, number, and gender).

Basic pattern elements are elements-strings and elements-words. _Element-string_ describes either a particular word form (e.g. Rus. *"задачей"* – word *problem* in instrumental case of singular), or particular symbols (for example, abbreviations or punctuation marks: *";"*). _Element-word_ describes a word, for which it may be specified:

- part of speech (POS: *N* – noun, *V* – verb, *A* – adjective, *Pr* – preposition , *Pn* – pronoun and so on);

- particular lexeme (i.e. all possible word forms of this word);

- particular values of morphologic attributes (they diminish the set of allowable word forms).

Morphologic attributes are written in angle brackets after the lexeme, with letter *t* denoting time, letter *p* denoting person, *c* – case, *n* – number, *g* – gender, etc.). For example, element-word

V<пониматься, t=pres, p=3, m=ind> describes Russian verb with lexeme *пониматься* taken in all forms of third person, present indicative (two word forms: *понимается* and *понимаются*). While describing an element-word, its morphologic attributes or its particular lexeme may be omitted, which makes it possible to allow within the corresponding phrase any word form of the given lexeme (e.g. N<файл>), or any word of the particular part of speech with needed values of morphologic attributes (e.g. A <;c=ins,n=sing> specifies an arbitrary adjective in instrumental case of singular).

Since LSPL-pattern often includes either several elements-words of different part of speech, or several different words of the same part of the speech, indices are used to distinguish the words. For example, the pattern NN = N1 N2<c=gen> includes two different nouns N1 and N2, the second is taken in gender case.

Agreement conditions describe relation of grammatical agreement for elements-words within the pattern. The conditions are written in angle brackets at the end of LSPL-pattern, similar to specification of values of morphologic attributes. They express the equality of values of morphologic attributes to be agreed. For instance, the pattern PnV = Pn V <Pn.n=V.n, Pn.g=V.g> specifies an arbitrary pair of prounoun and verb, which are agreed in number and gender (*Rus*.: *мы предположим*; *Eng*.: *we suppose*).

If some element of the phrase may occur in it successively several times, such a sequence is specified in corresponding pattern as *repetition of elements*, which is written in figure brackets. For example, repetition {N<c=gen>} describes sequence of nouns, each taken in genitive case. If the number of elements in the repetition is limited, it is specified in the pattern, for instance, {A}<1,3> N describes sequence including one, two or three adjective and a noun.

LSPL language also provides such useful device as *optional element* , which are written in square brackets, for example, the element ["не"] means, that particle *не* optionally enters the NL phrase under description. Another convenient device is *alternative variants* of the phrase – they should be written in the pattern through sign |. For instance, the pattern AP = A|Pa specifies Russian concept of adjective, i.e. adjective (A) or participle (Pa).

In order to describe patterns of complex phrases, one can use yet defined LSPL-patterns as auxiliary patterns within the main pattern. Let us consider the pattern NG = {A1} N1 [N2<c=gen>] <A1=N1> that includes the element-word N1 (principal word of the phrase), sequence of adjectives{A}, which are agreed with the principle word (<A1=N1>), and also optional noun in genitive case [N2<c=gen>]. Phrases with such structure are frequently used as terms in Russian texts (e.g., *восходящий процесс порождения, удаленный банковский терминал*). Based on the auxiliary pattern NG, the pattern S = NP V<t=past> specifies any phrase including a noun phrase NP and a verb in the past (e.g., *опорная точка уточнялась*).

Lexico-syntactic pattern may also have parameters, they are written in brackets, after all pattern elements and agreement conditions. The parameters fix some unvalued morphological attributes of pattern elements. For the LSPL-pattern AAN = A1 A2 N <A1=A2=N> (N), morphological parameters of the element-word N are specified as pattern parameters (the pattern describes noun phrase with elements-adjectives A1 and A2 fully agreed with noun N).

Pattern parameters are especially useful when the pattern is used as an element within another pattern. Suppose the pattern NG considered above has the parameter N1 (i.e. morphological attributes of the noun N1):

NG = {A} N1 <A=N1> [N2<c=gen>] (N1)

Then one can use the parameter for agreement. For instance, the pattern NG V <NG=V> describes phrase consisting of noun phrase NG and verb V grammatically agreed with it (e.g., Russian word combination *внутренний файл проверялся* is allowable, but *внутренний файл проверялись* is not, since the noun is not agreed with the verb).

Pattern parameters are also useful for specifying values of morphological attributes of the pattern used within the outer pattern; the specification is written in angle brackets, similar to specification of attributes of elements-words, for instance, in the pattern NG <c=gen> V the noun phrase NG is specified in gender case.

In overall, LSPL language is a flexible and powerful tool for describing lexical and grammatical properties of NL phrases to be recognized in texts.

Matching LSPL-Pattern with Text

For recognizing within a given NL text all phrases described by the particular LSPL-pattern, a matching procedure was developed. We call recognized phrases and corresponding text segments *variants of matching* of the pattern with the text. Each matching variant presents a text segment together with particular morphologic attributes of all its constituent words; the set of the particular values of morphologic attributes we call *syntactic interpretation* of the segment. When the segment consists of a single word, its syntactic interpretation is simply all morphological attributes of the word. In general case, for a given LSPL-pattern and text there exists several variants of matching, they correspond to different occurrences of the phrase described by the pattern.

Our matching procedure is based on special inner representation of the text – *graph of the text*. Nodes of the graph corresponds to space symbols, punctuation marks and all the other symbols that are not significant for matching; to be more precise, any segment of all such adjacent symbols constitute a node. Edges of the graph correspond to syntactic interpretations of text segments between the nodes.

While constructing the graph, first, segmentation of the text is done (words are delimited, as well as sequences of symbols that are not significant), and nodes of the graph are constructed and numbered from the beginning of the end of the text. Then morphologic analysis of all words is performed, and neighbor nodes are connected with edges represented morphologic interpretations of the words between them. If there exist several different morphologic interpretations of the same word, the corresponding nodes are connected with several edges. An example of graph representation for Russian sentence is presented on Figure 1 (the segmentation is also shown above the graph). One can notice that the number of morphologic interpretations for the same word form may be quite great. For example, word form *большой* has six interpretations, while the word *нечеткий* has only two.

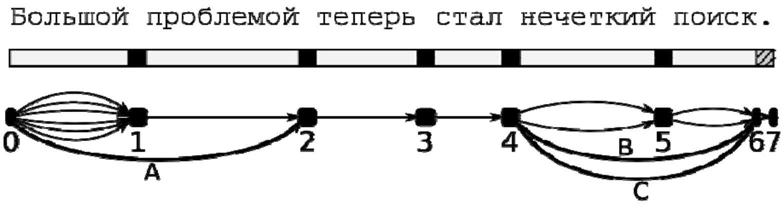

Figure 1. Graph of text with variants of matching the pattern NP = A N <A=N> (N)

Various ways in the graph of text corresponds to various possible combinations of morphologic interpretations of words. Therefore we consider the task of matching a pattern with the text as the task of searching a way (or a subway) within the graph that conforms to the pattern (i.e. pattern elements and agreement conditions).

Intermediate results of matching are also saved within the graph of text: any phrase recognized by matching with the pattern (main or auxiliary) is represented in the graph as a new edge connecting nodes pointing to beginning and end of the phrase (i.e. its text segment). This new edge presents matching variant for the pattern, and if the pattern has parameters, their values are additional attributes of corresponding syntactic interpretation.

In Figure 1 edge A represents the variant of matching of the pattern NP = A N <A=N> (N) with text segment *большой проблемой*, while edges B and C represent two different variants of matching of the pattern with segment нечеткий поиск (they differ in syntactic interpretation: B corresponds to nominative case and C to accusative). Thus, more than one variant of matching may be detected for the same text segment.

When LSPL-pattern includes repetition of elements, its matching also gives a new edge connecting all elements of repetition recognized in the text.

Therefore, the proposed graph of text is a convenient way to represent various syntactic interpretations and their combinations, as well as to uniformly process both elements-words and auxiliary patterns. It also allows optimizing of matching with the aid of indices constructed simultaneously with the graph. For this purpose, three types of indices are used: index of particular words, indices of parts of speech, and index of patterns yet matched.

Another optimization method also used by matching procedure is grouping of various syntactic interpretations, which diminishes the number of matching variants to be considered while searching way within the graph of text.

The described matching procedure is a core of programming tools developed to support LSPL language. These tools include console utilities for integration the core with various scripts, API for Java programming language, and graphic user interface. All the tools were first used to develop and to test automatic term extraction procedures for Russian scientific and technical texts.

LSPL-Patterns for Terminological Analysis of Scientific Texts

In order to formalize heterogeneous linguistics information needed to automatically extract terms and term definitions, an empirical study of terminology dictionaries and texts in several scientific fields (approx. 330 texts in computer science and physics) was performed. Based on the study, the formalization was done with the aid of LSPL language, resulting in a set of LSPL-patterns. The set comprises 6 groups of patterns that take into account various properties of term occurrences within Russian scientific texts. These groups, corresponding examples of patterns and examples of recognized term occurrences are presented in Table 1.

Table 1. LSPL-patterns for terminological analysis

N	Pattern Groups	Examples of Patterns	Examples of Terms and Term Occurrences
1	Morphosyntactic patterns of terms	*A1 N1 <A1=N1> (N1)*	активные долготы
		N1 A2 N2<c=gen> <A2=N2> (N1)	технология двойной накачки
2	Definitions of authors' terms	*Defin<c=acc>"называют"["также"] Term<c=ins> # Term<c=nom>*	Эту проблему называют также проблемой скрытого состояния
		"под" Term<c=ins> "понимается" Defin<c=nom> # Term<c=nom>	Под прерыванием понимается сигнал…
3	Contexts of introduction of terms' synonyms	*Term1 "("Term2")" <Term1.c=Term2.c>* *# Term1<c=nom>, Term2<c=nom>*	автокорреляционной функции (АКФ), зоны анализа (сегменты)
4	Dictionary terms	*N1<вектор> [N2<намагниченности,c=gen>| N2<состояния,c=gen>|"Умова"]*	вектор, вектор намагниченности, вектор состояния, вектор Умова
5	Lexico-syntactic variants of terms	*N1 N2<c=gen> # N1,* *N1 N4<c=gen> <Syn(N2,N4)>,* *N3 N2<c=gen> <Syn(N1,N3)>,* *A1 N1 <A1.st=N2.st>*	коллекция текстов – коллекция (N1), корпус текстов (N3 N2), текстовая коллекция (A1 N1)

6	Combinations of several terms	*N1 N2<c=gen> "," N3<c=gen> {"и"|"или"} N4<c=gen>* *# N1 N2<c=gen>,N1 N3<c=gen>, N1 N4<c=gen>*	шинам адреса, данных и управления – шина адреса, шина данных, шина управления
		N1 A2 N2<c=gen> <A2=N2> *# N1 N2, A2 N2*	разрядность внутреннего регистра – разрядность регистра, внутренний регистр

The first group describes morphosyntactic structure of one-, two- and tree-word terms frequently used in texts. Each pattern fixes part of speech of its element-words and morphological attributes of words (if necessary).

The second group formalizes typical one-sentence definitions of new terms introduced by authors of texts (so called *author's terms*); an example of such English definition is *Light quanta came is called photons*. Each LSPL-patterns of the group uses special auxiliary patterns Term and Defin. The former comprises all allowable morphosyntactic patterns of terms (i.e. patterns of the first group), the latter describes syntactic structure of phrases explicating meaning of new terms. Each pattern of the group also includes special element # Term <c=nom> , which specifies a constituent part of the recognized term definition to be extracted, as well as its lemmatization conditions (nominative case is specified for extracted term Term).

The third group includes LSPL-patterns of contexts typically used in Russian scientific texts to introduce synonymous terms (in particular, acronyms, such as *CPU* for term *central processing unit*).

As LSPL language proved to be convenient for describing entries of terminology dictionaries, the fourth group of patterns was constructed to specify particular terminological words and word combinations in two scientific fields – computer science and physics.

The last two groups of LSPL-patterns describe general derivation rules for text variants of terms. Term variation and methods of term variants recognition was well investigated for English and French text [Nenadic G. et. al, 2003], and we conducted analogous research for Russian texts. Besides variation of single term (cf. the group of lexico-syntactic variants in Table 1), we additionally considered typical combinations of several terminological word combinations and formalized their properties.

Each pattern of the fifth group fixes particular morphosyntactic structure of the term and specifies as the extracted element (i.e., after special sign #) morphosyntactic structure of its possible lexico-syntactic variants. In particular, if the structure of term is N1 N2<c=gen> #N1, the following lexico-syntactic variants are described:

i) insert (or deletion) of word (Rus. ввод данных – ввод, Eng. data input – input);

ii) substitution of a synonym (in the given problem domain) for constituent part of the term (*Rus. фрейм активации - запись активации; Eng. activation frame - activation record*);

iii) substitution of a word with the same root but another part of speech (*Rus.* шина адреса – адресная шина, Eng. address bus – bus of address).

The last group of LSPL-patterns describes (in a similar manner) derivation rules of text variants combining several terms. The rules take into account two different cases:

- combinations with coordinating conjunctions (Rus. шина адреса, шина данных, шина управления – шина адреса, данных и управления; Eng. address bus, data bus, control bus – address, data, and control bus);

- conjunctionless combinations (Rus. разрядность регистра, внутренний регистр – разрядность внутреннего регистра; Eng. capacity of register, internal register – capacity of internal register).

In both cases within described combinations one or more multy-word terms are discontinuous or truncated, and this is the real problem of their automatic recognition.

For each group of patterns, an automatic recognition procedure was developed and experimentally studied (in particular, a procedure for extracting new terms and their definitions). The recall of automatic recognition proved to be from 57% (for synonymous term recognition) to 85% (for dictionary terms), while precision varies from 32% (for synonymous term recognition) to 97 % (for authors' terms). In order to accomplish more accurate and full term detection and extraction, we then elaborated a strategy of consequent call of the procedures, which gives 7-14 % increase of F-measure (the combined measure of precision and recall).

LSPL-Patterns for Processing NL Sentences in Question-Answering System

LSPL language was also used to formally specify input NL phrases in a prototype question-answering system based on logical inference. The system is domain-independent, it gives answers to questions about existence of entities (animals, humans, games, cars, etc.) with particular properties (high, quick, black, difficult, etc.) or about properties of particular entities. Some properties of the entities are to be previously described by Russian sentences (these are initial statements). The system translates them to first-order logical formulas, which serves as axioms. In general case, axioms include universally and existentially quantified formulas-sentences (e.g. *all women like talking*, *some cats are black*), as well as formulas with implication (*if book is large, it is high-priced*).

The system uses axioms to infer answers to questions formulated in Russian (e.g. Are *all cats grey?*). For this purpose, the given question is translated to a logical formula and the resolution method is applied to prove it. Since questions are to be closed first-order formulas, the system gives either positive or negative answer to the given question.

LSPL-patterns developed for the question-answering system describes lexicon and syntax of input Russian sentences: either statements and questions. The patterns are divided into 5 groups presented in Table 2 together with corresponding examples (terms of computer games are mainly used in them).

Table 2. LSPL-patterns for question answering

N	Pattern Groups	Examples of Patterns	Examples of Phrases
1	Auxiliary patterns	*SubjDelim = "," ["а" "также"] | "и" | "или" |"а" "также"*	маги, колдуны, а также волшебники
		Exists = V<существовать> (V) | V<быть> (V)| V<бывать> (V)	эльфы бывают светлые
2	Patterns of entities	*EntityBase = {Adjective} N {SubjDelim Entity} <Adjective=N> (N)*	красный рыцарь
		Entity = EntityBase [Which Predicate {Delim Predicate} [","]] <EntityBase=Predicate> (EntityBase)	герой, который хорошо колдует
3	Patterns of properties	*Predicate = {Av} A (A)*	Очень сложный уровень
		Predicate = {Av} V (V)	Маг легко обучается

4	Patterns of statements	*Statement = Pn<некоторый> Entity Predicate {Delim Predicate} <Entity=Predicate>*	Некоторые феи хорошо поют
		Statement = [Pn<весь>] Entity ["-"] Predicate {Delim Predicate} <Entity=Predicate>	Все маги – бессмертные
		Statement = "если" Entity *["-"] Predicate {Delim Predicate} [","] "то" ["он"|"она"] Predicate {Delim Predicate} <Entity=Predicate>*	Если рыцарь быстро бегает, то он неуязвим
5	Patterns of questions	*Question = Predicate "ли" Entity {Av} <Av=Predicate=Entity>*	Бессмертны ли эльфы?
		Question = {Av}<1> "ли" Predicate Entity <Av=Predicate=Entity>	Долго ли живут орки?

The first group specifies general-purpose words (such as Rus. *быть*) and structure of auxiliary language constructs (in particular, enumeration phrases). The second and the third groups formalize respectively syntax of various phrases denoting entities (noun and participle phrases) and syntax of phrases denoting their properties (noun and verb phrases). The forth group comprises patterns of various statements to be translated to logical axioms: universal sentences, existential sentences, and sentences expressing implications. The last group includes patterns of user questions (similar to the previous group, universal and existential questions are allowable and specified). Most patterns of two last groups take into account various order of constituent words in the phrases (in Russian, the order is quite free), and as a consequence, these patterns are complicated.

We should note that the expressive power of LSPL language has made it possible very quick development of a procedure for translation NL sentences to logic formulas.

Conclusion

In the paper we have overviewed formal declarative language LSPL proposed to specify lexico-syntactic patterns of NL phrases to be recognized in Russian texts and extracted from them. We also described main features of matching procedure intended to recognize the phrases within a given text based on a particular set of LSPL-patterns and shallow syntactic analysis.

The programming tools (with the matching procedure as a core) developed to support the language were experimentally tested while investigating two quite different applications, the first is terminology analysis of scientific and technical texts, and the second is processing of NL phrases for question answering. The programming tools (including the user interface similar to a text browser driven by patterns) have demonstrated their working efficiency.

Our experience shows that LSPL language is well-suited for quite different NL processing tasks. Another potential applications of the language to be further investigated include text summarization, computer-aided editing of scientific and technical texts, and intra-document browsing and retrieval.

Bibliography

Bolshakova, E.I., Baeva N.V., Bordachenkova E.A., Vasilieva N.E., Morozov S.S. Lexicosyntactic Patterns for Automatic Processing of Scientific and Technical Texts. In: Proc. of 10th National Conference on Artificial Intelligence with International Participation 2006. Moscow, Fizmatlit, Vol 2, 2006, p. 506-514 (in Russian).

Bolshakova E. I. Common Scientific Lexicon for Automatic Discourse Analysis of Scientific and Technical Texts // International Journal on Information Theories and Applications. Vol. 15, 2008, No 2, p. 189-195.

Bontcheva K. et al. Developing Reusable and Robust Language Processing Components for Information Systems using GATE. In: Proceedings of the 13th Int. Workshop on Database and Expert Systems Applications, DEXA. Washington, 2002, p. 223-227.

Boudin F. et al. Mixing Statistical and Symbolic Approaches for Chemical Names Recognition. In: Computational Linguistics and Intelligent Text Processing. A. Gelbukh (Ed.). LNCS, No. 4919. Springer, 2008, p. 334-343.

Hearst, M.A. Automated Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, 1998, p.131-151.

Grishman R. Information extraction. In: Mitkov R. (ed.): Handbook of Computational Linguistics. Oxford University Press, 2003. p. 545-59.

Jacquemin C., Bourigault D. Term extraction and automatic indexing. In: Mitkov R. (ed.): Handbook of Computational Linguistics. Oxford University Press, 2003. p. 599-615.

Nenadic G., Ananiadou S., McNaught J. Enhancing Automatic Term Recognition through Variation. In: Proceedings of 20th Int. Conference on Computational Linguistics COLING'04, 2004, p. 604-610.

Petasis G., et al. Ellogon: A New Text Engineering Platform. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas, 2002, p. 72-78.

Authors' Information

Elena I. Bolshakova – *Moscow State Lomonossov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: bolsh@cs.msu.su*

Major Fields of Scientific Research: Artificial Intelligence, Natural Language Processing

Natalia E. Efremova – *Moscow State Lomonossov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: nvasil@list.ru*

Major Fields of Scientific Research: Automatic Extraction of Terms and Relations, Sentiment Analysis

Alexey A. Noskov – *Moscow State Lomonossov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: alexey.noskov@gmail.com*

Major Fields of Scientific Research: Artificial Intelligence, Software Engineering, Natural Language Processing