EPiC Series in Language and Linguistics
Volume 4, 2019, Pages 95–106
Proceedings of Third Workshop "Computational linguistics and language science"

Terminological Information Extraction from Russian Scientific Texts: Methods and Applications

Elena I. Bolshakova[1], Natalia E. Efremova[2], Kirill M. Ivanov[2]

[1] Lomonosov Moscow State University,
National Research University Higher School of Economics, Moscow, Russia
eibolshakova@gmail.com,
[2] Lomonosov Moscow State University, Moscow, Russia
nvasil@list.ru, ivanov.kir.m@yandex.ru

Abstract. Scientific texts contain a lot of special terms, which together with their definitions present an important part of scientific knowledge to be extracted for various applications, such as text summarization, construction of glossaries and ontologies and so on. The paper reports rule-based methods developed for extracting terminological information involving recognition of term definitions, as well as detection of term occurrences within scientific or technical texts. In contrast to corpus-based terminology extraction, the developed methods are oriented to processing a single text and are based on lexico-syntactic patterns and rules representing specific linguistic information about terms in scientific texts. The formal language LSPL for specification of the patterns and rules is briefly characterized, which is supported with programming tools and used for information extraction. Two applications of the methods are discussed: formation of glossary for a given text document and subject index construction. For these applications, both collections of LSPL patterns and extraction strategies are described, and results of their experimental evaluation are given.

Keywords: information extraction from texts, rule-based information extraction, processing of scientific texts, automatic terms recognition, extraction of term definitions, lexico-syntactic patterns

1 Introduction

Information extraction (IE) from natural language texts is one of the most rapidly developing areas of modern computer linguistics and text mining. As a rule, IE applications are based on shallow syntactic analysis of texts and exploit both statistics and linguistics information about items to be automatically recognized in them. In addition to traditional tasks of extracting named entities, their relations and facts, IE involves extracting terminological information from texts, including recognition of terms used in the texts, as well as their definitions. The latter is important for processing of scientific and technical texts containing a lot of special terms, which together with their

G. Wohlgenannt, R. von Waldenfels, S. Toldova, E. Rakhilina, D. Paperno, O. Lyashevskaya, N. Loukachevitch, S.O. Kuznetsov, O. Kultepina, D. Ilvovsky, B. Galitsky, E. Artemova and E. Bolshakova (eds.), CLLS 2018 (EPiC Series in Language and Linguistics, vol. 4), pp. 95–106

definitions present an important part of scientific knowledge required for various applications, such as text summarization, construction of glossaries and ontologies and so on.

Up to date, methods of automatic term recognition are well studied – see the most important results in [10, 16]. For term detection, certain combinations of linguistic and statistical criterion are usually applied, and machine learning methods are often used to reveal crucial features of terms to be extracted. The application of such technique is mainly oriented to compiling terminology dictionaries, constructing thesauri and ontologies, based on corpora of texts from specialized domains.

For extracting definitions of terms, some methods were proposed in recent works [1, 9, 11-13], presenting both rule-based and machine-learning approaches to information extraction. Some of the methods are intended to automatic or semi-automatic construction of problem-oriented glossaries. As a rule, glossary is a list of the most important terms from a source text together with their definitions.

One more application of terminological information extraction is constructing subject indexes for text documents, which is important for highly specialized texts. Subject, or back-of-the-book index usually contains significant terms from a particular document together with page numbers where they are used. Similar to computer-aided constructing of glossaries, automatic back-of-the-book indexing [5-7, 14-15] is a little-researched area with the central problem of selecting relevant terms from a given text document.

The main objective of our work is to study rule-based approach to terminological information extraction for two particular applied tasks: formation of glossary and construction of subject index, both tasks are being performed for a single text document. While processing single text, statistical measures proposed for corpus-based term recognition becomes less useful or do not work at all. For this reason, we mainly rely on various linguistic information about term usage in texts and formalize it with lexico-syntactic patterns of terms and term definitions. We assume that consideration of various linguistic features of terms and their occurrences in texts facilitates their detection in texts. Since the intensive use of terminological phrases with diverse structures is typical for scientific texts, in our study we consider terminological extraction from scientific and technical texts.

Our work continues the research [3] and elaborates methods for the particular applied tasks: construction of glossaries and subject indexes. In contrast to above-mention works, we handle Russian texts and have developed extraction methods with the aid of correspondent tools: LSPL language [2] intended to formally specify lexico-syntactic patterns of Russian phrases within IE systems, as well as programming tools[1] supporting LSPL. We have developed a representative set of LSPL patterns and extraction rules reflecting the features of terms usage in Russian scientific texts and including grammatical patterns of multiword terms, patterns of term definitions and contexts of introducing terminological synonyms.

The paper starts with an overview of some works close to our research. Key features of LSPL and its programming tools are also briefly characterized, and basic collections of lexico-syntactic patterns and extraction rules developed with their aid are described. Then, terminological information extraction is considered for tasks of glossary formation and subject index construction, including collections of applied pattern, rules, extraction procedures. The results of their experimental evaluation are reported and discussed.

2 Related Work

Most terms are multiword units, e.g., *небесное тело*, *высота бинарного дерева* (*celestial body*, *height of binary tree*). Based on assumption that terms are frequently encountered within texts in

[1] http://lspl.ru/

certain grammatical forms, statistical and linguistics features of terms are used for their detection in texts [10], including grammatical patterns of multiword terms and a wide range of statistical measures for ranging relevancy of extracted terms. The latter is needed because term extraction techniques do not guarantee extracted text units to be true terms (units may be non-term phrases, such as *схема работы*, *общий вопрос* − *work scheme*, *general question*), so the extracted units are only *term candidates* and, as a rule, need to be verified by human experts with the aid of a ranged list of term candidates.

For corpus-based term acquisition, the above-mentioned and primarily statistics-based extraction technique gives the acceptable results (the top of the ranged list contains most terms), but for reliable term extraction from a single text, special methods of filtering term candidates are to be developed.

The methods of extracting and filtering terms developed in works [5, 6] for constructing subject indexes exploit some grammar patterns of terms and various statistical features based on word occurrences, but despite the use of machine learning, they do not achieve sufficient quality of term extraction, giving precision and recall about 27-28%.

The papers [7, 15] describe subject indexing systems that rely on linguistic rules for term extraction, but their performance is not evaluated. The rules of InDoc system [15] specify various grammatical structures of multi-word terms and their text variants encountered in the text. Since these rules were elaborated for corpus-based terminology extraction, they most likely do not have high efficiency.

The recent paper [1] devoted to glossary construction for software requirements documents proposes term extraction method based on grammatical patterns of terms along with clustering of extracted terms with the use of certain syntactic and semantic similarity measures. In experiments with three particular software requirements documents, the method showed 21-51% of precision, whereas the recall was about 90%, resulting in 35-67% of F-measure (combined measure of precision and recall).

We should note that high recall and low precision is the common situation for most widely-used term extraction methods based primarily on statistics of word occurrences. Analogous extraction techniques are applied for the similar task of keywords extraction (but keywords may be non-terms, such as *литература для* − *literature for school children*), giving similar results: as it is reported in the review [8] of keyword extraction methods, the best achieved scores are about 35% of precision, 66% of recall, 45.7 % of F-measure.

As for term definition extraction, few recent works propose methods within both rule-based and machine learning approach.

The paper [11] describes a rule-based system for terminological definitions extraction from English medicine texts written for non-specialists. The proposed extraction method accounts for structures of terminological noun phrases, cue phrases used in definitions, and corresponding lexical and text markers (e.g., *is called*, *so-called* etc., and punctuation marks). The method showed quite good quality of definition extraction on 53 selected sentences: 86.95% of precision and 75.47% of recall, but this is mainly due to additional usage of a parser implementing deep syntactic analysis.

The work [12] addressed the problem of constructing lexico-syntactic patterns of term definitions, and a semi-supervised learning method was proposed and regarded as an alternative for constructing extraction rules by human experts since they cannot account for lexico-syntactic variety of terms definitions. The method relies on so-called Word-Class Lattices constructed on the basis of generalized representation of term definitions taken from the tagged dataset (term definitions from Wikipedia were used). An experimental evaluation of the extraction method based on the constructed Word-Class Lattices has showed precision and recall of 98% and 39% correspondingly, so the method does not improve recall typical for rule-based methods.

The paper [13] describes a comparison of two methods for term definition extraction: the first is based on bootstrapping of lexico-syntactic patterns, the second uses deep syntactic and semantic analysis of text. It is reported that both methods have showed sufficient quality for glossary construction, but scores are not given.

The machine learning approach for term definition extraction is presented in [9]. The CRF classifier was trained on the tagged dataset containing definitions and non-definition sentences. For learning, features of words (lexemes, stems, POS), features of sentences (position in the document, inclusion of acronym), and features of the document were used. The best evaluation results were 80% of F-measure obtained on ACL Anthology texts, and 85% obtained on Wikipedia benchmark corpus.

We should note that all above-considered works deal with processing of text documents in English or French, while in our work we consider Russian texts and exploit corresponding tools. For Russian, there are still no tagged datasets or corpora acceptable for learning term extraction models, so it is reasonable to explore rule-based methods. In order to extract terms and term definitions with high degree of reliability, we have elaborated a collection of lexico-syntactic patters and rules, and then we have evaluated it in two applied tasks. Our rule-based approach is somewhat close to that in [15], but differs in collections of rules, extraction strategies, and applied tools.

3 Rule-Based Terminological Information Extraction

3.1 Lexico-Syntactic Patterns and Extraction Rules

For our purposes, we use LSPL (Lexico-Syntactic Pattern Language) developed for the formal specification of phrases (primarily nominal word combinations) to be automatically extracted from Russian texts [2]. It is a declarative language flexible enough to specify both lexical and syntactic features of Russian phrases by their *lexico-syntactic patterns*. The language is supported by programming tools[2] for development of various information extraction applications, in particular, extraction of terminological information.

Lexico-syntactic pattern describes certain Russian phrase: its words and other constituent elements, as well as their morphologic and surface syntactic properties. Elements of patterns include:

- word forms, particular lexemes, arbitrary words of particular part of speech;
- particular morphological attributes of words (case, gender, number and so on);
- conditions of grammatical agreement of its elements;
- optional and repetitive elements;
- alternative variants of the phrase being described;
- auxiliary patterns intended to describe complex phrases, part-by-part.

The order of the pattern elements corresponds to the order of constituents in the phrase. For example, LSPL pattern *NP = {A} N1 [N2<c=gen>] <A=N1>* consists of the element-word *N1* (principal word of the phrase), repetition (sequence) of adjectives *{A}*, which are grammatically agreed with the principle word (*<A=N1>)*, and also optional noun in genitive case *[N2<c=gen>]*. The pattern describes such nominal phrases as *входной текстовый файл*, *стандартный ключ утилиты* (*input text file, standard key of utility*). Conditions of grammatical agreement are especially important for describing Russian noun phrases.

[2] http://lspl.ru/

In LSPL pattern, morphological attributes of elements (words) can be specified, for example, pattern *A<входной, c=nom, g=fem>* describes all forms of adjective *входной* (*input*) in nominal case (*c=nom*) and feminine gender (*g=fem*), but grammatical number is not indicated. At the same time, unchangeable words can be given as strings, as in the following pattern specifying alternative Russian conjunctions: *C = "или" | "и" |"но"*.

One can use yet defined LSPL-patterns as auxiliary patterns within the main pattern. 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 вычислялась − reference point was calculated*).

Lexico-syntactic pattern may have parameters corresponding to some free morphological attributes of any pattern element, which is 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)*:

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

The parameter makes it possible to assign grammatical agreement for noun phrase *NG* or to specify its morphological attributes. In particular, in the pattern element *NP<n=plur>* grammatical number of *NP* (of its principle noun) is specified as plural *(входные текстовые файлы)*; and in the pattern *NP V <NP=V>* noun phrase *NP* and verb *V* are agreed (so the phrase *текстовый файл загружался* is allowable, but *текстовый файл загружались* is not, since the noun is not agreed with the verb).

Overall, a collection of mutually related LSPL patterns describes a grammar (extended with conditions) for phrases to be recognized in texts.

LSPL rule consist of lexico-syntactic patterns in the left-hand side and also *extraction pattern* in the right-hand side of the rule (given after symbol =>). Extraction pattern describes text items to be extracted from the recognized phrase. For example, the rule

A1"и" A2 N <A1=A2=N> => A1 #N "+" A2 #N <A1~>N, A2~>N>

specifies extraction from coordinative phrase with conjunction *и* (such as *входной и выходной файл − input and output file*) of both their parts in normal form (#) and separated by sign *"+"* , with extracted adjectives (*A1*, *A2*) agreed (~>) with the noun *N* (*входной файл + выходной файл*). In general, LSPL extraction rule describes a transformation of the recognized phrase into extracted text (the transformation is specified by the extraction pattern). In the given example, the extraction pattern specifies normalization (or lemmatization, the sign #) of the first noun and grammatical agreement of adjective with the first noun (*A~>N1*).

Main programming tools supporting LSPL include:

- LSPL processor, which is the core component for extracting items from texts, according their lexico-syntactic patterns and rules; the processor is based on shallow syntactic analysis including text segmentation and morphological analysis;
- Command-line utility that calls the LSPL processor and outputs the results in XML format (it is intended to simplify development of IE applications);
- Environment for creating and debugging LSPL patterns and rules, with graphical user interface for visualization of text extraction results.

Comparing LSPL devices with those of the known open source systems for creating IE applications, which also have formal languages for extracted constructions, namely GATE [4] and Tomita-Parser[3], we should point out the following. Unlike the pattern language of more universal GATE intended for text engineering, LSPL is purely declarative language linguistically-oriented to Russian text, which greatly simplifies description of phrases and development of applications. In

[3] https://tech.yandex.ru/tomita/doc/

comparison with Tomita-Parser intended for extracting facts from Russian texts, LSPL is more easy to learn and it also provides user with environment supporting development of patterns.

3.2 Lexico-Syntactic Patterns for Russian Scientific Texts

The described LSPL language proved to be convenient to formalize grammatical structures of scientific terms, as well as contexts of their usage typical for Russian scientific text, which is necessary for rule-based terminological information extraction. Formalizing linguistics features of scientific terms and their contexts gave us a representative set of LSPL patterns and rules, which comprises three basic collections. Some LSPL patterns from each collection, together with examples of extracted terms are presented in Table 1.

The first collection of 12 patterns describes grammatical structures of one-, two- and three-word terms frequently used in Russian scientific texts. Each pattern fixes part of speech of its constituent words, morphological attributes of words (if necessary), and grammatical agreement of adjectives and nouns. In particular, the collection include patterns *A1 A2 N <A1=A2=N>* (e.g., *асинхронная стартстопная передача – asynchronous start-stop transfer*) and *A N1 N2<c=gen>* (*линейный блок данных – linear data block*). Here is an example of an LSPL extraction rule from this collection:

A N1 N2<c=gen> <A=N1> (N1) => A #N1 N2<c=gen> <A~>N1>

(the pattern in the right-hand side describes the result of term extraction: the first noun *N1* is normalized while the adjective *A* is agreed with *N1*).

Table 1. Collections of LSPL Patterns

Collections and Pattern Examples	Examples of Terms and Contexts of Terms
1. Grammatical patterns of terms	
A N <A=N> => A #N <A~>N>	*одноклеточные организмы (unicellular organism)*
A1 A2 N <A1=A2=N> => A1 A2 #N <A1~>N, A2~>N>	*двоичный симметричный канал (binary symmetrical channel)*
2. Term definitions	
V<получить, t=past, p=3, m=ind> "название" Term<c=gen> => #Term	*...получила название биогеоценоза... (....was named biogeocenosis...)*
"под термином" Term<c=nom> "будем понимать" = > #Term	*...под термином "изменение климата" будем понимать... (...under the term "climate change" we will understand ...)*
"ключевым" ["понятием"] "является понятие" Term<c=gen> =>#Term	*ключевым понятием является понятие класса (...the key concept is the notion of class)*
3. Contexts of term synonyms	
Term1 "("Term2")" <Term1.c=Term2.c> => #Term1, #Term2	*дезоксирибонуклеиновая кислота (ДНК) – (deoxyribonucleic acid (DNA))*

The second and the third collections of LSPL patterns specify typical contexts of term occurrences, primarily, contexts of definitions of new terms introduced by authors of texts (so-called *author's terms*).

The second collection contains 35 patterns covering most frequent Russian-language one-sentence definitions of terms in scientific texts, for example: *Генофондом называется совокупность генов особей, составляющих популяцию* (*The set of all genes in a population is called gene pool*). Each

pattern includes both particular lexical units (verbs *называть*, *определять* – *call*, *define*, and so on) and special auxiliary patterns *Term* denoting word combination with grammatical pattern specified in the first group of patterns. For example, the definition phrase *Под сильным взаимодействием понимается*... is described by the pattern:

"под" Term <c=ins> V<пониматься, t=pres, p=3, m=ind>

where *Term* should be in instrumental case (*c=ins*) and *V* describes Russian verb with lexeme *пониматься* taken in all forms of third person, present indicative (two word forms: *понимается* and *понимаются*)

The third collection includes 15 patterns of contexts typically used in Russian scientific texts to introduce *terminological synonyms* and *abbreviations* (including synonyms for author's terms), for example: *"…лямбда-выражение, или определяющее выражение функции..."* (... *lambda expression, or expression of function definition* ...). Synonymous term variants may be acronyms, such as *SD* for term *standard deviation*. Each pattern of the collection recognizes a pair of synonymous terms (with grammatical structure described by the pattern *Term*). In particular, for the above-considered example, the following rule extracts terminological synonyms in normalized form, relying on comma and lexical markers (words *или просто*, the latter word is optional):

Term1 "," "или" ["просто"] Term2 => # Term1 #Term2

We should note that for particular applied tasks the described lexico-syntactic patterns and extraction rules may be modified, in order to achieve needed functionality. Below we consider glossary formation and subject index construction.

4 Glossary Formation

Generally, glossary is a list of *glosses*, e.g. terms with their definitions, which is placed at the back of text document. Typical fragments of glossaries are presented in Figure 1. Besides term definitions, glossary may contain examples of their use, translations to another language, some relations between terms (e.g., synonymous).

An allele is a variant form of a given gene.
A genome is the genetic material of an organism.
The gene pool is the set of all genes in any population.

Аллель – одно из конкретных состояний гена.
Геном – совокупность наследственного материала, заключенного в клетке организма.
Генофонд – совокупность генов особей, составляющих популяцию.

Figure 1. Fragments of Glossaries

For extracting definition of terms, we rely on LSPL patterns and rules of the second basic collection that contains auxiliary patterns *Term* and *Defin*. The former pattern describes a grammatical pattern of the term being defined, while the latter specifies the structure of phrases explicating the meaning of the term. For example, the following rule

Defin <c=acc> "будем называть" Term<c=ins> => #Term "–" #Defin

describes such definition phrases as *Передачу, которая содержит зубчатые колеса с перемещающимися осями, будем называть планетарной передачей*, where *Term* should be in

instrumental case (*c=ins*) and it is extracted in normal form (*#Term*), whereas the *Defin* phrase is to be in accusative case(*c=acc*). The extraction pattern (after symbol =>) describes construction of needed gloss from the extracted *Term* and *Defin*. For the given example, the following gloss will be constructed: *Планетарная передача – передача, которая содержит зубчатые колеса с перемещающимися осями.*

To evaluate the quality of term definition recognition by the developed patterns, we united in groups all patterns with the similar lexical markers. For experiments, we use four Russian medium-sized (about 20 thous. words) educational scientific texts devoted to programming languages (PL) and heuristic search methods in artificial intelligence (HS). For each group of patterns that have worked on these texts (it turned out that only 22 patterns have worked), their precision was evaluated, the results are given in Table 2 (the number of patterns that have worked are indicated in round brackets). One can see that precision of the patterns significantly vary: some patterns of term definitions are ambiguous, but almost a half of them have excellent or good precision. Nevertheless, certain procedure is needed to select true term definitions among extracted sentences. Moreover, additional selection is necessary because not all definitions introduce significant terms to be included into glossary.

Table 2. Precision of term definition patterns

Lexical Markers	PL (Lisp)	PL (Refal)	PL (Prolog)	HS	Average
получить название (2)	1.00	-	1.00	1.00	1.00
ключевое понятие (2)	1.00	-	1.00	-	1.00
понимать (1)	1.00	-	-	-	1.00
будем говорить (1)	1.00	-	-	-	1.00
являться (2)	1.00	0.50	1.00	-	0.83
называться (5)	0.50	0.33	0.80	0.00	0.41
считаться (2)	0.67	0.00	-	-	0.33
представлять (1)	0.29	0.67	0.00	0.25	0.30
называть (3)	0.33	0.25	0.25	0.33	0.29
это (1)	0.40	0.40	0.13	0.13	0.26
т.е. (1)	0.11	0.00	0.29	0.10	0.12
есть (1)	0.00	0.33	0.09	0.00	0.11
Mean	0.57	0.26	0.49	0.24	0.56

We have also evaluated recall of term definition extraction from the processed texts, for all the collection of patterns, it proved to be quite high, 0.94. One reason of incompleteness is that several terms (more than one) may be defined in the same sentence. The problem can partly be overcome by adding new patterns.

The more difficult problem is that 49% of glosses (besides terms, they contain their definitions) could not be extracted because of linguistic variability of phrases explicating term meaning: some phrases have complex grammatical structures, the other do not have lexical markers at all, as in the following phrase defining term *косвенная рекурсия*: *Более сложным случаем является неявная, косвенная рекурсия, при которой рекурсивный вызов возникает во время вычисления функции* (*A more complex case is implicit, indirect recursion, when a recursive call occurs during computing the function*). Such phrases are not covered by our collection of patterns, and all they hardly can be described as single *Defin* pattern or even several patterns.

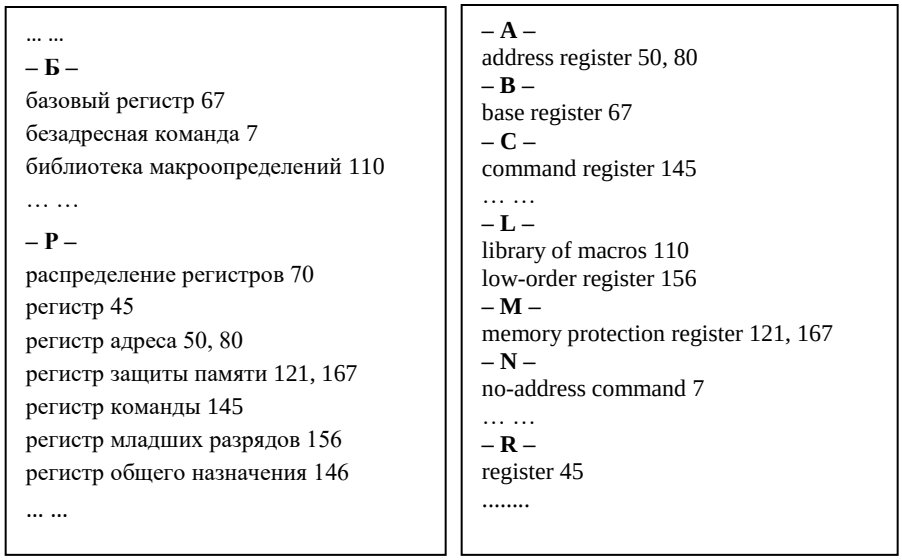
Another reason of incompleteness in gloss extraction is that some terms are not really defined in the extracted sentences, their definitions precede or follow these sentences. For example, the phrase *Эту форму называют функциональной блокировкой* (*This form is called a functional lock*) only

indicates that definition of the term *функциональная блокировка* is in the sentence before the extracted one.

Thus, to accomplish more accurate and complete gloss extraction, it is necessary, besides augmenting the collection of patterns and rules, to take into account term importance (through statistics of term occurrences in the text being processed) and also to extract information from neighbor sentences.

5 Subject Index Construction

To provide an easy access to the content of text document, subject index (back-of-the-book index) should contain significant terms from the document and their synonymous variants. Therefore, the central problem of subject index construction is extracting single-word and multi-word terms by applying linguistics and statistic criteria and filtering the more appropriate ones among extracted term. Typical fragments of subject indexes are presented in Figure 2.



... ...
– Б –
базовый регистр 67
безадресная команда 7
библиотека макроопределений 110
… …
– Р –
распределение регистров 70
регистр 45
регистр адреса 50, 80
регистр защиты памяти 121, 167
регистр команды 145
регистр младших разрядов 156
регистр общего назначения 146
... ...

– A –
address register 50, 80
– B –
base register 67
– C –
command register 145
… …
– L –
library of macros 110
low-order register 156
– M –
memory protection register 121, 167
– N –
no-address command 7
… …
– R –
register 45
.....

Figure 2. Fragments of Subject Indexes

In order to achieve sufficient quality of extraction we have somewhat modified extraction rules from the second collection of lexico-syntactic patterns and rules. Unlike glossary construction, for constructing index we use term definition patterns without *Defin* pattern, it was excluded from all patterns, since actual definitions of terms are not needed for a subject index. Moreover, we add a pattern with lexical marker *"так называемый"* (*so-called*) into the collection, here is the corresponding rule:

"так"{"называемый"|"называемая"|"называемое"|"называемые} Term => #Term

Such modified and augmented collection of term definition patterns makes possible to extract a more wide set of term candidates (in general, the usage of *Defin* limits the structure of phrases explicating terms, and hence it limits terms can be extracted).

Three basic developed collections of rules are independently applied for term extraction, giving correspondently three sets of extracted term candidates. The resulting sets are intersected, in

particular, there are terms extracted by patterns of grammatical structure and also by patterns of term definition (or patterns of terminological synonyms). Therefore, it is necessary to combine the sets, aiming to select the more significant ones for a subject index.

Based on some experiments with these pre-extracted sets of term candidates, we have elaborated a heuristic procedure selecting index terms. First, it discards from the sets those term candidates that a) are encountered in the pre-compiled list of stop word; b) consist of words from this stoplist; c) contain words from another pre-compiled stop list. Thereby many collocations of common scientific lexicon with the similar grammatical structures (e.g., *задача, простая задача*, *известная схема – problem*, *simple problem*, *known scheme*) are discarded.

Than the procedure iteratively forms a collection of index terms, taking elements from the rest of the sets and accounting for several factors of term importance:

- precision of the pattern of term definition, which was applied for term extraction: for the subset of very-high precision extraction rules, extracted terms are obligatory included into subject index;
- usage of term in heading/subheading of the text document;
- frequency of term occurrences: according to Zipf's law, the most significant terms are units with an average frequency and so they are selected to the index;
- lexical similarity of terms: a term candidate can be added into the index if it has common words (at least one) with any element yet included in it.

Table 3. Evaluation of term extraction for subject indexes

Text	Text size (in words)	Selected Index Terms		P	R	F
		#	Examples			
PL (Lisp)	21060	140	динамическое связывание (dynamic binding)	0.74	**0.84**	**0.79**
PL (Refal)	29301	208	функциональный терм (functional term)	0.56	0.82	0.67
PL (Prolog)	21376	77	хорновская формула (Horn formula)	**0.77**	0.72	0.75
HS (Heuristic search)	19471	98	редукция задач (reduction of tasks)	0.71	0.74	0.73
PS (Progr. Systems)	11699	67	функциональное тестирование (functional testing)	0.70	0.81	0.75
Mean	20581	118		**0.70**	**0.79**	**0.74**

To experimentally evaluate efficiency of the described extraction procedure, five educational scientific texts with subject indexes constructed by their authors were taken (one more text devoted to programming systems was added to those used in the experiments on term definition extraction). The results of evaluation measured in precision (P), recall (R), and F-measure (F) are shown in Table 3.

Our term extraction procedure demonstrates quite good performance, exceeding the known methods [1,5,6] of term extraction for subject indexing: its recall (in average 0.79) is sufficient, and precision (in average 0.70) is acceptable, as well as F-measure (0.74). For subject index construction, recall is more crucial, since for human editor it is easier to discard some terms from the than to add new ones.

6 Conclusion

In the paper we have described rule-based methods developed for terminological information extraction in two tasks of processing Russian scientific documents: formation of a problem-oriented glossary and constructing a subject index for a given document. The core of the methods is the described collections of lexico-syntactic patterns and rules representing linguistic information about terms in Russian scientific texts, they are used for recognition and extraction of terms and term definitions.

For subject indexing, the experimental evaluation of the proposed extraction procedure based on the patterns and simple word occurrence statistics has shown its good efficiency: in precision and F-measure, the procedure outperforms known methods of terms extraction for subject indexes. Nevertheless, new experiments for its verification and tuning are needed, on texts of various scientific domains and sizes.

For glossary construction, the described simple extraction strategy based only on patterns and rules is insufficient for detecting all term definitions and for extracting glosses from a given text document. Certain improvements are feasible by extending and refining LSPL extraction patterns and rules, but in order to reveal significant terms and their definitions, it is evidently necessary to develop a procedure taking into account both statistics of term occurrences and relations between sentences.

Another potential applications of scientific information extraction to be investigated within the rule-based approach include key phrase extraction and text summarization.

References

1. Arora, C., Sabetzadeh, M., Briand, L., Zimmer, F.: Improving Requirements Glossary Construction via Clustering: Approach and Industrial Case Studies. In: Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. ACM, New York, NY (2014).
2. Bolshakova, E., Efremova, N., Noskov, A.: LSPL-Patterns as a Tool for Information Extraction from Natural Language Texts. In: Markov, K., Ryazanov, V., Velychko, V., Aslanyan, L. (Eds.) New Trends in Classification and Data Mining, pp. 110-118. ITHEA, Sofia (2010).
3. Bolshakova, E., Efremova, N.: A Heuristics Strategy for Extracting Terms from Scientific Texts, Analysis of Images, Social Networks and Texts. Fourth Int. Conference AIST, CCIS, Vol. 542. Springer Berlin Heidelberg, pp. 285-295 (2015).
4. 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, pp. 223-227 (2002).
5. Csomai, A., Mihalcea, R.: Investigations in Unsupervised Back-of-the-Book Indexing. In: Proceedings of the Florida Artificial Intelligence Research Society Conference, pp. 211-216 (2007).
6. Csomai, A., Mihalcea, R.: Linguistically Motivated Features for Enhanced Back-of-the Book Indexing. In: Proceedings of Ann. Conf. of the Association for Computational Linguistics, ACL/HLT, Vol. 8, pp. 932–940 (2008).
7. Da Sylva, L.: Integrating Knowledge from Different Sources for Automatic Back-of-the-Book Indexing. In: Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI (2013).
8. Hasan, K.S., Ng, V.: Automatic keyphrase extraction*:* a survey of the state of the art. Proceedings of the 52th Annual Meeting of the ACL, pp. 1262–1273 (2014).
9. Jin Y., M.-Y. Kan, J.-P. Ng, and X. He. : Mining scientific terms and their definitions: A study of the ACL anthology. In EMNLP ’13, pp. 780–790 (2013).
10. Korkontzelos, I., Ananiadou, S.: Term Extraction. In: Oxford Handbook of Computational Linguistics (2nd Ed.). Oxford University Press, Oxford (2014) .

11. Muresan, A., Klavans, J.: A method for automatically building and evaluating dictionary resources. Proceedings of the Language Resources and Evaluation Conference (LREC) (2002).
12. Navigli, R., Velardi, P., & Roma, S. U. D.: Learning word-class lattices for definition and hypernym extraction (2010).
13. Reiplinger, M., Schafer, U., Wolska, M.: Extracting glossary sentences from scholarly articles: a comparative evaluation of pattern bootstrapping and deep analysis. Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, Stroudsburg, USA, pp. 55–65 (2012).
14. Wu, Z. et al.: Can Back-of-the-Book Indexes be Automatically Created? In: Proceedings of the 22nd ACM Int. Conference on Information & Knowledge Management. ACM, pp.1745-1750 (2013)
15. Zargayouna, H., El Mekki, T., Audibert, L., Nazarenko, A.: IndDoc: an Aid for the Back-of-the-Book Indexer. The Indexer, 25(2), pp. 122–125 (2006).
16. Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.: A Comparative Evaluation of Term Recognition Algorithms, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), pp. 2108-2111 (2008).