A Heuristic Strategy for Extracting Terms from Scientific Texts

Elena I. Bolshakova[1] and Natalia E. Efremova[1]

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

Abstract. The paper describes a strategy that applies heuristics to combine sets of terminological words and words combination pre-extracted from a scientific text by several term recognition procedures. Each procedure is based on a collection of lexico-syntactic patterns representing specific linguistic information about terms within scientific texts. Our strategy is aimed to improve the quality of automatic term extraction from a particular scientific text. The experiments have shown that the strategy gives 11-17% increase of F-measure compared with the commonly-used methods of term extraction.

Keywords: multiword terms · automatic term extraction · text variants of terms · term occurrences in scientific text · lexico-syntactic patterns

1 Introduction

Automatic extraction of terms from domain specific texts plays crucial role in various natural language processing (NLP) applications, such as compiling terminology dictionaries, constructing thesauri and ontologies, text abstracting, etc. The problem of automatic term recognition was studied over last two decades, and significant results were obtained – see, for example, [4-5], [11].

Most terms, including scientific ones are multiword units, e.g., *nonlinear plan*, *coefficient adjustment learning*. In order to recognize them within NL texts, shallow syntactic analysis along with statistical and linguistics criteria are used, based on assumption that terms are frequently encountered within texts in specific grammatical forms. The applied extraction techniques do not guarantee extracted text units to be true terms (e.g., non-term phrase *general plan* may be extracted), so they are only *term candidates* and usually need to be confirmed by human experts [5].

In most modern applications of automatic term extraction (in particular, for thesaurus construction), large text collections and corpora are processed, and exploiting statistical criteria of term recognition along with even poor linguistic information

(such as part of speech of words) gives acceptable quality of extraction measured by precision and recall. Statistical criteria include *tf.idf* measure (widely used in information retrieval [12]) and its numerous modifications, as well as their combinations (see, for example, [15]). Meanwhile, in such applied tasks as text abstracting and summarization, computer-aided writing and editing of specialized texts, recognition of terms is performed from a single text, and statistical measures becomes less significant because of less volume of the text under processing. Moreover, contrast corpora needed to compute *tf.idf* value are not always available. Thus, more comprehensive linguistic information is to be applied for reliable term extraction based on shallow syntactic analysis. Besides grammatical patterns of multiword terms, linguistic information may comprise local context data similar to that described in [16] for extraction of terms (concepts) from highly specialized texts.

In contrast to corpus-based term extraction, we consider the task of term recognition in a single text, the task is necessary for several applications, including computer-aided construction of glossaries and subject indexes for text documents [1], [6]. We should note that task of term recognition somewhat differs from keywords extraction (e.g., [13]), since terms denote concepts of problem domain, while keywords may be non-terms (such as *trend monitoring* or *banking application*). We assume that consideration of various linguistic features of terms and their occurrences within texts facilitates their detection in texts and improve the quality of term extraction. Since the intensive use of terminological phrases with diverse structure is typical for scientific papers, in our study we consider term extraction from scientific texts.

Heterogeneous information about terms in Russian scientific texts was formalized and represented as linguistic patterns of term structures, text variants of terms, and terminological contexts. For purposes of formalization, we used LSPL (Lexico-Syntactic Pattern Language) [3] developed for specifying linguistics features of Russian phrases. LSPL programming tools were exploited as well to implement several term extraction procedures. Each procedure is based on a collection of LSPL patterns representing specific linguistic information about terms in scientific texts.

The term extraction procedures were studied experimentally, and analysis of the results gave us a strategy for improving the quality of automatic term extraction from a particular scientific text. The strategy implies heuristics on how to combine sets of term candidates pre-extracted by the procedures. The experiments have shown that our strategy gives 11-17% increase of F-measure (the combined measure of precision and recall) compared with commonly-used methods for term extraction [7], [9], [17].

The objectives of this paper are:

- To clarify the categorization of extracted term candidates on the basis of linguistic information used for term recognition;
- To briefly describe a collection of LSPL patterns representing linguistic information of specific types;
- To overview the term extraction procedures developed for groups of patterns, as well as results of their experimental evaluation on Russian scientific texts;
- To describe our heuristics strategy for combining sets of term candidates extracted by the procedures.

2 Scientific Terms and their Lexico-Syntactic Patterns

In order to reveal heterogeneous linguistic information useful for automatic term extraction, we have performed an empirical study of scientific texts in Russian (approx. 330 texts), as well as terminological dictionaries in several scientific fields (including computer science and physics) [2], [8]. Based on the study, we formalized revealed linguistics features of multi-word terms and their occurrences in texts on the basis of LSPL language [3] intended to support development of various applications for information extraction from Russian texts.

LSPL is a declarative formal language flexible enough to specify both lexical and syntactic features of phrases to be extracted from Russian texts. Phrases are specified in the form of *lexico-syntactic patterns*; elements of patterns include particular word forms, lexemes, arbitrary words of particular part of speech (POS), morphological attributes of words, conditions of grammatical agreement. The agreement conditions are especially important for describing Russian noun phrases. To specify complex phrases, auxiliary patterns can be defined and used within LSPL pattern.

Formalizing linguistics features of scientific terms gave us a representative set of LSPL patterns, which comprises 6 groups. Every group of patterns corresponds to specific linguistic information used to recognize term occurrences in texts. Therefore, our categorization of extracted terms reflects types of linguistic information represented in the patterns.

Examples of LSPL patterns of each group, as well as examples of terms or term contexts extracted by patterns of the corresponding groups are presented in Table 1. Let us consider the groups and types of extracted terms in more detail.

The first group of patterns specifies grammatical structures of one-, two- and tree-word terms frequently used in scientific texts. This linguistics information is commonly-used by most term extraction methods. We should note that each LSPL pattern fixes not only POS of constituent term words (*N* is noun, *A* is adjective), but also their morphological attributes (if necessary). In both patterns given in Table 1 grammatical agreement of adjectives and nouns are specified (*A=N* and *A1=A2=N*). Symbol => denotes extraction of terms recognized by the patterns.

Regretfully, not only terms have the specified grammatical structures, in particular, collocations of common scientific lexicon may have the similar structure, e.g., *main problem*, *developed scheme*. Therefore, grammatical patterns of multiword terms are not enough reliable for term extraction, and hereinafter text units extracted by LSPL grammatical patterns we call ***term candidates***.

The second and the third groups of patterns specify typical contexts of term occurrences. Primarily, we consider phrases for definition of new terms, which are often encountered in scientific papers, for example: *"A set of entities created by expansion process we call reference markers"*. Such new terms that are explicitly introduced by authors and then used within their texts are called ***author's terms*** [2].

Each LSPL pattern of the second group specifies typical one-sentence definition of author's term, the pattern includes both particular lexical units (verbs *call*, *define*, and so on) and special auxiliary patterns *Term* and *Defin*. The former describes all allow-

able grammatical patterns of terms (i.e. patterns of the first group), while the latter specifies structure of phrases explicating meaning of the defined term.

Table 1. Groups of LSPL Patterns

Groups and Examples of Patterns	Examples of Terms and Contexts of Terms
1. Grammatical patterns of terms	
A N <A=N> => A N	*Rus. стерильный нейтрино –* *Eng. sterile neutrinos*
A1 A2 N <A1=A2=N> => A1 A2 N	*Rus. горячая темная материя –* *Eng. hot dark matter*
2. Definitions of authors' terms	
Defin <c=nom> V<называться, t=pres, p=3> Term<c=ins> <V.n=Defin.n> => Term	*Rus. … яркое кольцо называется кольцом Эйнштейна – Eng. … a bright ring is called Einstein ring*
Term "is" Defin => Term	*Rus. Вероятность есть степень возможности… – Eng. Probability is the measure of the likeliness…*
3. Contexts of introduction of term synonyms	
Term1 ("Term2") *<Term1.c=Term2.c>* *=> Term1, Term2*	*Rus. двумерный электронный газ (ДЭГ) –* *Eng. two-dimensional electron gas (2DEG)*
4. Dictionary terms	
N1<адрес> [N2<возврат,c=gen> *| N2<результат,c=gen>* *| N2<точка,c=gen>* *N3<вход,c=gen>]*	*Rus. адрес, адрес возврата, адрес результата, адрес точки входа – Eng. address, return address, result address, entry point address*
5. Combinations of several terms	
A1 "и" A2 N <A1=A2=N> => *A1 N <A1=N>, A2 N <A2=N>*	*Rus. гравитационная и инертная масса => гравитационная масса, инертная масса –* *Eng. gravitational and inertial mass => gravitational mass, inertial mass*
N1 A N2<c=gen> <A=N2> => *N1 N2<c=gen>, A N2 <A=N2>*	*Rus. разрядность внутреннего регистра => разрядность регистра, внутренний регистр –* *Eng. capacity of internal register => capacity of register, internal register*
6. Text variants of term	
A1 N <A1=N> => N, *A2 N <A2=N> <Syn(A1,A2)>*	*Rus. сильное взаимодействие => взаимодействие, ядерное взаимодействие –* *Eng. strong force => force, nuclear force*

In the first example of definition pattern given in Table 1, *Term* is to be in nominative case (*c=nom*), and the *Defin* phrase is to be in instrumental case (*c=ins*); for Russian verb *называться* morphological attributes of time and person (*t=pres*, *p=3*) and agreement in number with *Defin* phrase are specified as well. Each pattern of the

group includes the element => *Term* that denotes text item to be extracted from the recognized definition phrase.

The third group of LSPL patterns specify typical contexts encountered in scientific texts and intended to introduce ***term synonyms*** (including synonyms for author's terms), for example: *... the generalized momentum, also known as the canonical momentum...* . Term synonyms include, in particular, acronyms, such as *CPU* for term *central processing unit*.

Besides grammatical structures of terms and contexts of term definitions, linguistic information useful for term recognition are represented in terminological dictionaries (given that they are available for text processing). Such dictionaries seldom accumulate a complete set of terms in particular scientific field but yet fix many stable terminological words and word combinations. As LSPL language proved to be convenient for describing entries of terminology dictionaries, we have created a group of patterns that specify terms from several text dictionaries in the fields of physics and computer science. The patterns contain particular lexemes with particular morphological attributes (if needed), for example, genitive case (*c=gen*) for Russian words *возврат*, *результат*, *точка*, *вход* in the example pattern from Table 1 (the forth group of patterns) is specified. This pattern specifies four terms, symbol | separate alternatives, optional elements are written in square brackets. Term occurrences recognized in text by such dictionary patterns are called ***dictionary terms***.

The last two groups of LSPL patterns describe general rules for text variants of terms that often occur in Russian scientific texts. Variation of terms and methods to detect term variants is relatively well studied for English and French texts [10], [14]. In our empirical study we have revealed analogous term variants for Russian [8]. Besides variations of single term – cf. the group of ***text variants*** in Table 1 – we additionally consider typical ***combinations*** of several multi-word terms and formalized their features as LSPL patterns – cf. the fifth group of patterns.

Rules of combining several multi-word terms 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. поляризация волн, электромагнитные волны => поляризация электромагнитных волн – Eng. polarization waves, electromagnetic waves => polarization electromagnetic waves*).

In both cases within described combinations one or more multiword terms are broken (discontinuous) or truncated, and this is the real problem of their automatic recognition in texts. Any LSPL pattern of combination fixes its grammatical structure and also specifies (after sign =>) patterns of its constituent elements (i.e. grammatical structure of multi-word terms that compose this term combination).

Each pattern of the last group describes (in a similar manner) grammatical structure of the term being varied; grammatical structure of its possible text variants are specified after sign =>. In particular, if the structure of multiword term is *N1 N2<c=gen>*, its text variants comprises:

- insert (or deletion) of word from the term (e.g., *Rus. ввод данных => ввод – Eng. data input => input*);
- substitution of a synonym (in the given scientific domain) for constituent part of the term (e.g., *Rus. фрейм активации => запись активации – Eng. activation frame => activation record*);
- substitution of a word with the same root but another part of speech (e.g., *Rus. шина адреса => адресная шина – Eng. address bus => bus of address*).

We should note that to recognize synonymy variation of some constituent word of the term, certain dictionary of synonyms is to be incorporated into term extraction procedure. The element <*Syn(A1,A2)*> of the LSPL pattern given in Table 1 denotes necessary check-up of adjectives *A1* and *A2* in the incorporated synonymy dictionary.

3 Testing Term Extraction Procedures

For each described group of patterns, an automatic term extraction procedure was developed based on the LSPL programming tools.

Five developed procedures, namely procedures for extraction terms candidates, authors' terms, term synonyms, dictionary terms, and term combinations, are applied to a given text, and resulting sets of extracted terms are formed, along with recognized occurrences of the terms within the text. The procedures were tested with the aim to evaluate the quality of term extraction by traditional measures, i.e. recall and precision. For testing, a collection of Russian medium-sized texts (from 1597 to 4767 words) on computer science and physics were taken. Output results of the procedures were compared with sets of terms recognized and extracted by human experts.

The results of experimental evaluation of the procedures are shown in Table 2. Besides extraction of terms (in this case their presence in the given text was tested), we evaluate recognition of all occurrences of extracted terms within the text under processing. For example, let us consider a text fragment:

The geodetic effect represents the effect of the curvature of spacetime, predicted by general relativity, on a vector carried along with an orbiting body. The geodetic effect was first predicted by Willem de Sitter in 1916.

The term *geodetic effect* can be extracted from it, and also two its occurrences can be recognized (occurrences are required for such applied task as subject index construction). For term synonyms and term combinations, recognition of their occurrences in the texts was not evaluated, since both occurrences of synonyms and occurrences of terms extracted from combinations are recognized as term candidates.

Recall of automatic extraction proved to be from 58% (for term candidate recognition) to 94% (for dictionary terms), while precision varies from 25% (for term combination recognition) to 96% (for authors' terms). For recognition of term occurrences, the values are 60-89% and 49-78% for recall and precision respectively.

As for the procedure for extraction of term variants, information about its performance is absent in Table 2, because the procedure is intended to merely reveal variants among terms yet extracted by other procedures.

Table 2. Recall and Precision of Extraction Procedures

Procedure and Type of Terms	Extraction of Terms		Recognition of their Occurrences	
	Recall	Precision	Recall	Precision
Term candidates	58%	27%	60%	49%
Authors' terms	92%	96%	74%	78%
Synonyms of terms	64%	50%	–	–
Dictionary terms	94%	83%	89%	72%
Term combinations	82%	25%	–	–

Our analysis of detected cases of incompleteness and inaccuracy of term extraction from texts shows that restrictions of the applied linguistics criteria is evidently the main reason of the imperfect results. In particular, certain terms are not extracted because of their complex grammatical structure that is not represented in our LSPL patterns, however it is almost impossible to compile any exhaustive inventory of possible term structures. Some patterns of term definitions are ambiguous, and their addition to the set of extraction patterns increases recall of extraction but simultaneously decreases precision. Moreover, patterns of term combinations, as well as patterns of term candidates fix only their grammatical structure, so many non-term word combinations (e.g., *important problem of astronomy*) match the patterns. Dictionary terms are not recognized in the cases, when they are broken within term combinations, whereas some text units are falsely recognized as dictionary terms, though they are only fragments of other terms or non-term collocations (such as mathematical term *series* in common scientific collocation *series of experiments*).

In general, using additional linguistics information facilitates more reliable term extraction based on another specific information, in particular, accounting for patterns of term combinations increases the recall of recognition of dictionary terms, while accounting for dictionary terms increases the precision of recognition of authors' terms.

Since linguistics features of terms used for their extraction by the tested procedures are not mutually exclusive, the resulting sets of extracted terms are intersected, for example, extracted term candidates include terms recognized by other procedures. Therefore, in order to accomplish more accurate and complete term detection and extraction, it is reasonable to combine the output sets of our term extraction procedures. It should be pointed out that simple union of the sets is not appropriate because it evidently gives increase of recall with simultaneous decrease of precision.

4 Strategy for Combining Sets of Extracted Terms

Based on the results of the described experiments, we derived certain heuristics on how to combine the sets of terms recognized by the extraction procedures from the given text. In order to improve the overall quality of term extraction, we elaborated a strategy that iteratively forms a resulting set of terms by applying these heuristics while selecting elements of the pre-extracted sets of terms.

Let us denote the output sets produced by the extraction procedures as `TCAND` (term candidates), `AUTH` (authors' terms), `SYN` (term synonyms), `DICT` (dictionary

terms), COMB (conjunctionless term combinations), COMBCON (term combinations with conjunctions), TVAR (term variants).

The steps of our strategy are as follows.

S:=∅

Step 1. S1:=AUTH ∪ $DICT_{not\_part\_TCAND}$

Step 2. S2:=$DICT_{from\_COMB}$ ∪ $COMB_{with\_DICT}$

S:=S1 ∪ S2

CL: Step 3. S3:=SYN_{for_S}; S:=S ∪ S3

Step 4. S4:=$DICT_{from\_COMBCON}$ ∪ $TCAND_{from\_COMCON}$; S:=S ∪ S4

Step 5. S5:=$DICT_{from\_COMB}$ ∪ $TCAND_{from\_COMBCON}$ ∪ $COMB_{without\_TCAND}$

If S3 ∪ S4 ∪ S5 ≠ ∅ then S:=S ∪ S3 ∪ S4 ∪ S5; goto CL

Step 6. S6:=$TVAR_{for\_S}$

If S6 ≠ ∅ then S:=S ∪ S6; goto CL

Step 7. S7:=$TCAND_{freq>F}$

If S7 ≠ ∅ then S:=S ∪ S7; goto CL

Step 8. S8:=DICT

If S8 ≠ ∅ then S:=S ∪ S8; goto CL

The final set S of terms is formed incrementally. Initially S is empty, and whenever some element of abovementioned sets of pre-extracted terms are included into S, the element is removed from the source set. S1-S8 denotes terms, selected in the corresponding steps.

First of all, we put into S terms recognized with a high degree of precision. In step 1 all author's terms (AUTH) are included into S, dictionary terms are included as well unless they are fragments of term candidates ($DICT_{not\_part\_TCAND}$). Thereby, we do not approve as actual term any text unit recognized as dictionary, if all its occurrences in the text are embedded into occurrences of some term candidate (such as pair *dark matter – hot dark matter*). However, in Step 2 we add to S those remaining dictionary terms that are constituents of conjunctionless combinations ($DICT_{from\_COMB}$), since this condition increases the likelihood they are really terms. For similar reasons, any conjunctionless combination is added to S provided that it includes some dictionary term as constituent ($COMB_{with\_DICT}$).

In steps 3-5 synonyms and combinations of terms are considered. We include into S synonyms of all terms that belong to actual S (SYN_{for_S}). If any term combination with conjunction includes as constituent an element either from S, from DICT or from TCAND (as broken term), in Step 4 we add to S all constituents of this term combination (we denote them $DICT_{from\_COMBCON}$ and $TCAND_{from\_COMBCON}$). Similarly, in Step 5 we add to S all constituents of conjunctionless term combinations provided that they include as constituent (in broken form) any element either from S, from DICT or from TCAND ($DICT_{from\_COMB}$ and $TCAND_{from\_COMB}$). If there are no such conjunctionless combinations, we add to S whole conjunctionless combinations ($COMB_{without\_TCAND}$) instead of their constituent terms.

If the set S is extended in steps 3-5, these steps need to be repeated, otherwise the following steps are performed. Similarly, after each of the steps 6, 7, 8 in the case of extension of the set S, steps of our strategy are to be repeated from the step 3.

In step 6 term variants for all terms that belong to actual S ($TVAR_{for\_S}$) are added to S. Then we add to S all remaining term candidates ($TCAND_{freq>F}$) with frequency more than F, where F is computed as rounded weighted arithmetic average of term candidates frequencies. Thus, we exclude from consideration relatively rare term candidates.

In the final step 8 we include into S all dictionary terms that are not yet in S, and then if needed we repeat the Steps 3-7.

The described term extraction strategy produces a set of selected terms with higher degree of reliability. We have performed experiments to evaluate and compare recall and precision of the strategy and several commonly-used extraction methods, which we consider as baseline methods. For experiments, a collection of texts (approx. 33,000 words) in the same scientific fields (computer science and physics) was taken.

The baseline methods use information about frequencies of words and grammatical structures of terms. Method Mutual-Inf extracts two-word terms (possibly with prepositions) based solely on statistics of word occurrences and co-occurrences [17]. Method Mod-Mutual is a modification of the previous method, it additionally accounts for single word terms and POS of words [7]. Method C-Value recognizes multiword terms by using frequencies of words and information about embedded terms [9]. Method SP extracts multi-word terms according to their grammatical patterns.

The results of experimental evaluation of our strategy in comparison with the baseline method are shown in Table 3. The strategy shows significantly better performance in precision and compatible performance in recall compared with the method Mod-Mutual. In overall, for term extraction our strategy gives 17,6% increase of F-measure (the combined measure of precision and recall) compared with the best baseline method, and 11,7% increase of F-measure for recognition of term occurrences.

Table 3. Comparative Evaluation of the Proposed Strategy

	Extraction of Terms			Recognition of Term Occurrences		
Method	Recall	Precision	F-measure	Recall	Precision	F-measure
Mutual-Inf	27,3%	13,0%	17,6%	24,4%	20,4%	22,2%
Mod-Mutual	**54,1%**	37,4%	44,2%	**69,2%**	41,5%	51,9%
C-Value	35,5%	4,9%	8,6%	21,3%	5,9%	9,3%
SP	51,4%	22,6%	31,4%	37,3%	29,7%	33,1%
Strategy	53,6%	**73,1%**	**61,8%**	68,1%	**59,7%**	**63,6%**

5 Conclusion

Aiming to improve the overall quality of term extraction from a given scientific text, we propose a heuristic strategy based on various linguistics information including grammatical structures of multiword scientific terms, their text variants, and contexts

of their usage. The information about various features of term occurrences within Russian scientific texts has been formalized and represented as a set of LSPL lexico-syntactic patterns. Several term extraction procedures have been implemented with the aid of LSPL programming tools, each procedure uses a particular group of extraction patterns. This made it possible to experimentally evaluate efficiency of the implemented extraction procedures and then to reveal certain heuristics on how to combine and select output sets of terms extracted by the procedures in order to produce a set of terms with higher degree of reliability. Experimental evaluation of the proposed heuristic strategy shows significant increase of F-measure in comparison with the commonly-used methods of term extraction.

Nevertheless, our heuristic strategy needs additional verification on texts of various scientific domains and sizes. We believe that further improvement of the strategy is feasible by extending and refining LSPL extraction patterns. Necessary experiments can be performed without reprogramming our extraction procedures (only the input sets of patterns should be changed).

The described extraction procedures and strategy are undoubtedly useful for various NLP applications with complex processing of terminological units, especially for computer-aided abstracting of scientific texts and for computer support of scientific writing. In fact, writing support involves checkups of term consistency and accuracy within a particular scientific document, as well as construction of a problem-oriented glossary and a subject index for the document, the work can be done only by means of automatic term extraction procedures similar to those described in our paper.

Acknowledgements. We would like to thank the anonymous reviewers of our paper for their helpful and constructive comments.

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.: Recognition of Author's Scientific and Technical Terms. In: Gelbukh, A. (Ed.). Computational Linguistics and Intelligent Text Processing, LNCS 2004, pp. 281-290. Springer, Hedelberg (2001)
3. 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)
4. Bosma, W., Vossen, P.: Bootstrapping Language Neutral Term Extraction. In: Proceedings of the 7th Language Resources and Evaluation Conference, pp. 2277-2282. LREC, Valetta (2010)
5. Castellvi, M., Bagot, R., Palatresi, J. Automatic Term Detection: A Review of Current Systems. In: Bourigault, D., Jacquemin, C., L'Homme M.-C. (Eds.) Recent Advances in Computational Terminology, pp. 53-87. John Benjamins, Amsterdam (2001)

6. 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)
7. Dobrov, B., Loukachevich, N., Syromiatnikov, S.: Forming base of terminological word combinations from problem oriented texts. In: Proceedings of the 5th Russian scientific conference "Digital Libraries: Perspective Methods and Technologies, Electronic Collections", pp. 201-210 (2003) (in Russian)
8. Efremova, N.E.: Methods and Programming Tools for Extraction of Terminological Information from Scientific and Technical Texts: PhD Thesis, Lomonosov Moscow State University (2013) (in Russian)
9. Frantzi, K., Ananiadou, S., Mima, H.: Automatic Recognition of Multi-Word Terms: The C-value/NC-value method. In: Nikolau, C. et al. (Eds.) International Journal on Digital Libraries, vol. 3(2), pp. 115-130 (2000)
10. Jacquemin, C., Tsoukermann, E.: NLP for term variant extraction: synergy between morphology, lexicon, and syntax. In: Strzalkowski, T. (Ed.) Natural Language Information Retrieval, pp. 25-74, Kluwer Academic Publishers, Dordrecht (1999)
11. Korkontzelos, I., Ananiadou, S.: Term Extraction. In: Oxford Handbook of Computational Linguistics (2nd Ed.). Oxford University Press, Oxford (2014)
12. Manning, C. D., Raghavan P., Schütze H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
13. Matsuo, Y., Ishizuka, M.: Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. In: International Journal on Artificial Intelligence Tools, 13 (1), pp. 157-169 (2004)
14. Nenadic, G., Ananiadou, S., McNaught, J.: Enhancing Automatic Term Recognition through Recognition of Variation. In: Proceedings of 20th International Conference on Computational Linguistics COLING'04, pp. 604-610. Morristown, NJ (2004)
15. Nokel, M.A., Bolshakova, E.I., Loukachevich, N.V.: Combining Multiple Features for Single-Word Term Extraction. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Issue 11, vol. 1, pp. 490-501. RGGU, Moscow (2012)
16. Paice, C.D., Jones P.A.: The Identification of Important Concepts in Highly Structured Technical Papers. In: Korfhage, R., Rasmussen, E., Willett, P. (Eds.) Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 69-78. ACM, Pittsburgh, PA (1993)
17. Smadja, F., McKeown, K.: Automatically Extracting and Representing Collocations for Language Generation. In: Proceedings of the 28th Annual Meeting on Association for Computational Linguistics, pp. 252-259. ACL, Pittsburgh, PA (1990)