Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

TERM EXTRACTION FOR CONSTRUCTING SUBJECT INDEX OF EDUCATIONAL SCIENTIFIC TEXT

Bolshakova E. I. (eibolshakova@gmail.com)
Moscow State Lomonosov University, National Research University Higher School of Economics, Moscow, Russia

Ivanov K. M. (ivanov.kir.m@yandex.ru)
Moscow State Lomonosov University

Subject index, or back-of-the-book index, is a device intended to provide an easy access to relevant fragments of a text document. Subject indexes usually contain particular single-word and multi-word terms from the corresponding documents. Such indexes are especially useful for reading large documents with specialized terminology, as well as educational texts in difficult scientific and technical areas. The central problem of back-of-the-book indexing is recognition of terms to be included into the index. The paper describes a method developed for extracting and filtering terms from a given educational scientific text, with the purpose of reliable term selection in computer indexing systems. The method is primarily based on rules with lexico-syntactic patterns representing linguistic information about terms and typical contexts of their usage in Russian scientific and educational texts; simple occurrences statistics of terms is used as well. Experimental evaluation of the method has shown a considerable increase of precision and recall of term extraction compared with the widely-used standard techniques.

Keywords: rule-based term extraction, back-of-the-book index, subject indexing, educational scientific texts, lexico-syntactic patterns

ИЗВЛЕЧЕНИЕ ТЕРМИНОВ ДЛЯ ПОСТРОЕНИЯ ПРЕДМЕТНОГО УКАЗАТЕЛЯ УЧЕБНО-НАУЧНОГО ТЕКСТА

Большакова Е. И. (eibolshakova@gmail.com)
МГУ имени М. В. Ломоносова, НИУ ВШЭ, Москва, Россия

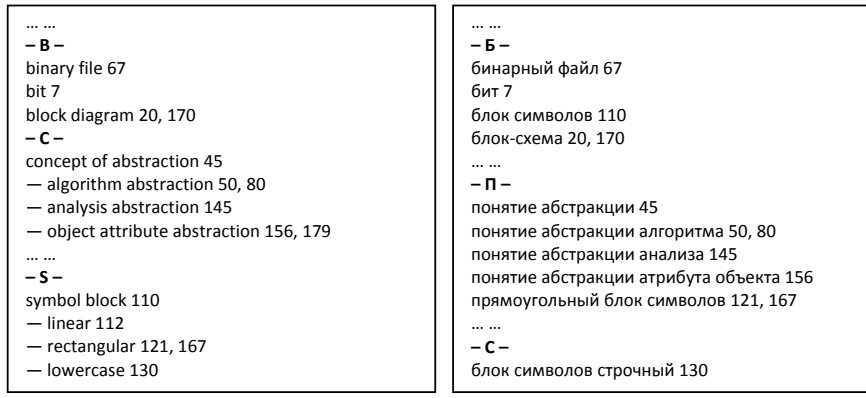
Иванов К. М. (ivanov.kir.m@yandex.ru)
МГУ имени М. В. Ломоносова, Москва, Россия

Предметный указатель к текстовому документу обычно содержит значимые однословные и многословные термины текста, вместе с номерами страниц, где они встречаются, что облегчает доступ к нужным фрагментам документа. Предметные указатели особо полезны для больших документов со специальной терминологией, а также для учебных текстов в сложных научных и технических областях. Центральной проблемой построения предметного указателя является выявление и отбор терминов для включения в указатель. В статье описывается метод, разработанный для извлечения из текста терминов и их последующей фильтрации в рамках программной системы поддержки построения предметных указателей. Метод основан на применении правил с лексико-синтаксическими шаблонами, отражающими лингвистическую информацию о терминах и контекстах их использования в учебно-научных текстах на русском языке. Экспериментальная оценка метода показала существенный прирост точности и полноты отбора терминов по сравнению с широко используемой стандартной технологией извлечения терминов из текстов.

Ключевые слова: извлечение терминов на основе правил, предметный указатель, учебно-научные тексты, лексико-синтаксические шаблоны

1. Introduction

Subject, or back-of-the-book, indexes are often constructed for large and medium-size text documents, such as books, manuals, tutorials etc., especially in highly specialized domains. As a rule, subject indexes contain significant terms from the corresponding documents, with associated page numbers. Such indexes are usually placed at the back of the text documents in order to facilitate navigating through them and locating needed information. Typical fragments of subject indexes are presented in Figure1 (hierarchical index in English and flat index in Russian).



... ...
– B –
binary file 67
bit 7
block diagram 20, 170
– C –
concept of abstraction 45
— algorithm abstraction 50, 80
— analysis abstraction 145
— object attribute abstraction 156, 179
... ...
– S –
symbol block 110
— linear 112
— rectangular 121, 167
— lowercase 130

... ...
– Б –
бинарный файл 67
бит 7
блок символов 110
блок-схема 20, 170
... ...
– П –
понятие абстракции 45
понятие абстракции алгоритма 50, 80
понятие абстракции анализа 145
понятие абстракции атрибута объекта 156
прямоугольный блок символов 121, 167
... ...
– С –
блок символов строчный 130

Fig. 1. Fragments of subject indexes

For educational texts written in scientific and technical domains (textbooks, manuals, tutorials, etc.) subject indexes are necessary devices. Indeed, scientific texts contain many specific terms with their definitions, and as a rule, students need to study the definitions and important contexts of term usage (explaining the corresponding scientific concepts), more than once, but without reading the full text. Regretfully, subject indexes are absent in many modern textbooks and manuals for students, especially in texts of rapidly developing scientific and technical domains. To now, automated back-of-the-book indexing is an under-investigated problem, and the high-laborious text indexing work remains mainly manual since modern word processing tools provide only technical assistance.

Among recent NLP works, relatively few papers are devoted to automating back-of-the-book indexing [5–7, 13, 14], and few subject indexing systems are known: In-Doc [15] and commercial TExtract[1].The central problem of index construction is extracting single-word and multi-word terms by applying linguistics and statistic criteria and filtering the more appropriate ones among extracted terms.

The works [5, 6] address the problem of extracting and filtering terms from a given text document; the proposed methods use some linguistic features of terms, statistical measures based on word occurrences, as well as machine learning, gaining precision and recall about 27–28%. The main difficulty of term detection for subject indexing relates with the fact that term extraction is performed from a single text, so various statistical measures developed and applied for corpus-based terminology extraction [9, 12] perform poorly or even are not applicable. The papers [13, 15] describes methods that are mainly based on linguistic rules for term extraction, but they are poorly described and do not provide enough information about their evaluation.

In our work, we consider the term extraction problem for subject index construction in relation to educational scientific documents, which makes it possible to use linguistic information about terminological features of such texts and thereby to achieve sufficient efficiency of the developed method. Our method of term extraction and

[1] http://www.texyz.com/textract/

filtering relies on rules and lexico-syntactic patterns accounting for grammatical structure of multiword terms, as well as typical contexts of their usage in the texts to be processed. Besides the linguistic rules, only simple term frequency statistics is used, without involving external text resources (so measure *tf.idf* widely used in information retrieval tasks [10] is not used).

Unlike above-mention works dealing with back-of-the-book indexing of English or French text documents [5–7, 13, 14], we consider Russian texts and exploit corresponding NLP tools. Our rule-based and corpus-free approach continues the work [4], it is close to those in [2, 15], but differs in the collection of extraction rules and in the strategy of filtering terms.

To evaluate our method, we use Russian educational scientific texts of medium size, mainly on computer science. The experiments have shown rather good performance, in average up to 70–79% of precision, recall and F-measure (the combined measure of precision and recall) for term extraction and filtering, which is considerably exceeds the results of statistics-based and machine learning methods [5, 6], as well as scores presented in [1] for several term extraction tools (approximately 20–47% of F-measure).

Since the results of term extraction based on modern NLP techniques are not strongly precise, the resulted list of terms needs to be validated and edited by a human expert in a problem domain, so any computer subject indexing system will inevitably be semi-automatic. Another reason for editing results by the expert is related with absence of standards on structure and content of indexes, and the work of human editors may be subjective and constructed indexes may vary in content and size. Our term extraction method is built into a research prototype of computer system[2] supporting back-of-the-book indexing and providing a user with graphical interface for setting parameters of the method and for editing results.

To clarify specialty of our approach, we begin with a brief explanation of term detection methods and their application for tasks close to subject index construction.

2. Related studies

Automated extraction of terms from texts is well investigated over last three decades. 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 [2, 9]. The elaborated extraction techniques do not guarantee extracted units to be true terms (in particular, a phrase of general lexicon like *main question* may be extracted), so resulted units are considered as *term candidates* and need to be filtered. The filtering task is usually performed by evaluating and ranking the extracted term candidates with certain statistical measures and machine learning (see [12, 16]).

It should be noted that developed methods and techniques are mainly intended for extracting terms from specialized text corpora, aiming to compile terminology dictionaries or to construct thesauri and ontologies in particular domains. For these

[2] https://github.com/ivanov-kir-m/SISTool

tasks, the methods have acceptable quality, but for processing single texts their effectiveness is not sufficient. Term recognition in single texts is often needed for keyword extraction [11], glossary construction [1], as well as for back-of-the-book indexing. For these tasks term extraction methods need to be modified and evaluated.

The methods developed in [5, 6] for back-of-the-book indexing rely on some grammar patterns of terms and various statistical term features based on word occurrences, but even applying machine learning they achieve about 27–28% of precision and recall. So whether the machine learning and pure statistical approach is a good choice for subject indexing seems questionable.

The works [2, 15] exploit linguistic rules for term extraction, which specify various grammatical structures of multi-word terms and their text variants encountered in the text. These rules were elaborated for corpus-based terminology extraction (from texts in French and English), and their performance for the back-of-the-book indexing task is not indicated.

The recent paper [1] describes term extraction method developed specifically for glossary construction for software requirements documents. The method uses grammatical patterns of terms along with clustering extracted terms based on certain syntactic and semantic similarity measures. In experiments with three particular software requirements documents, the method gives 35–67% of F-measure (with precision 21–51%, and recall about 90%) and slightly exceeds the best results of five term extraction tools taken for comparison. We should note that high recall and low precision is the common situation for most term extraction methods.

One can notice that the task of term extracting is quite similar to keywords recognition, but there is some difference, since terms denote concepts of a problem domain, while keywords represent main topics of the document (and may be non-terms, such as *economic trends*). However, the widely-used extraction techniques are applied for keyword extraction, and the best scores achieved on known datasets and reported in [8] are 35% of precision, 66% of recall, 45.7% of F-measure.

In contrast to the considered works, for reliable term extraction, we use a representative set of linguistic rules with lexico-syntactic patterns accounting for term features in Russian scientific texts. The formal rules are written in LSPL language [3], and the developed method has been implemented with the aid of LSPL programming tools[3].

3. Term Candidate Extraction

For extracting terms from a given text, the set of LSPL rules[4] were elaborated, based on lexico-syntactic patterns from [4]. The set encompasses three groups.

The first group of 12 rules specifies extraction of one- and multi-word terms by their typical grammatical structure (it is commonly-used by most term extraction methods [9]). The rules fix a part of speech of words (POS) and their grammatical characteristics (case, gender, etc.), for example, the pattern *N1 A N2<c=gen>*

[3] http://lspl.ru/

[4] https://github.com/ivanov-kir-m/SISTool/tree/master/Patterns

(*элементы двоичной арифметики*—elements of binary arithmetic), where *N1* is a noun, *A* is an adjective, *N2* is a noun in genitive case.

The second and the third groups of rules specify term extraction from typical contexts of term occurrences, primarily, contexts of term definitions. Such contexts are often encountered in educational scientific text, for example: *"An integrated lights-out we call remote management feature"*. Evidently, defined terms belong to significant terms to be included in the subject index.

The second group contains 53 rules for extracting terms from their definitions, covering most of the typical Russian-language phrases-definitions of terms. The rules include both particular lexical units (verbs *называть*, *определять*—*call*, *define*, and so on) and a special auxiliary pattern *Term* denoting phrase with grammatical pattern specified in the first group of rules. For example, the definition phrase *"Интегрированной средой будем называть..."* (*We call the integrated environment…*) is described by the rule:

Term <c=ins> «будем» «называть» => # Term

where *Term* should be in instrumental case (*c=ins*) and is extracted (=>) in normal form (*#Term*).

The third group consists of 25 rules specifying typical contexts for introducing terminological synonyms and abbreviations in Russian scientific texts, for example: *"… информационная система, или просто ИС"* (*... information system, or simply IS ...*). The rules recognize and extract pairs of term synonyms (they should have valid grammatical patterns), relying on commas and lexical markers (e.g., words *или просто*), in the following rule the word *просто* is optional:

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

As a result of all the extraction stage, three sets of term candidates are formed: M_{gram}, M_{def}, M_{syn}, respectively.

We have estimated the precision of term extraction for each group of rules. For this purpose two educational textbooks of medium-size on programming languages Lisp and Refal (112 and 95 pages respectively), together with their human-made subject indexes were used. For the first group of rules, experiments have shown high recall of term extraction but low precision (about 8–10%), which was expected. On the contrary, rules of the second group demonstrate high precision (90–95%) overall, due to lexical markers used in them. For similar reasons, the third group of rules shows a rather good precision: 63–67%.

Since rules and patterns of term definitions (from the second group) vary in precision, we have selected a subset of very-high precision rules, their extracted terms are labeled as *Trusted*. This label is used in our filtering procedure aiming at selection of the most important terms with the high degree of reliability.

4. Term Filtering

Based on the results of several experiments with the output sets of extracted terms M_{gram}, M_{def}, M_{syn}, we elaborated a heuristics filtering procedure that encompasses three stages.

At the first filtering stage, pre-compiled lists of stop words[5] are used. The first stop list contains words that cannot be terms (e.g., *метод, начало, отмена*—Eng.: *method*, *start*, *cancel*), while the second list contains words that cannot be part of terms, they are mainly adjectives (e.g., *данный, известный*—Eng.: *given*, *known*). From all the sets M_{gram}, M_{def}, M_{syn}, their elements are excluded that a) are encountered in the first stop list; b) contain words from the second list; c) consist of words from the first stop list. Thereby many collocations of the common scientific lexicon with the similar grammatical structure (e.g., *simple method*, *given scheme*) are discarded.

At the next filtering stage, the frequency of occurrences for all term candidates is calculated, and for frequencies of elements from M_{def}, the percentiles are calculated with the levels p_1=0.4 (rounding down) and p_2=0.95 (rounding up), respectively.

The third stage intended to account for several factors of term candidate importance: frequency of term occurrences, usage in headings/subheadings of document sections, as well as lexical similarity of terms (that is, they have common words, e.g., *tail recursion* and *high order recursion*). According to Zipf's law, the most significant terms are units with an average frequency, and the usage of percentiles makes it possible to eliminate unlikely term candidates (both rare and frequent).

The resulting set R of subject index terms is incrementally formed according to following steps (initially R is empty).

Term candidates from the set M_{def} labeled as *Trusted*, whose frequency is in the range $[p_1,p_2]$, are added to the set R.

Term candidates from the set M_{gram}, whose frequency is in the range $[p_1,p_2]$ are added to R, provided they are encountered in some heading or subheading of the processed document (if any).

Term candidates from the set M_{gram}, whose frequency is in the range $[p_1,p_2]$, are added to R, provided they have common words (at least one) with any *Trusted* term, whose frequency is out of the range $[p_1,p_2]$.

Remaining term candidates from the set M_{def} (unconsidered in step 1) having common words (at least one) with any element from current R are added to R.

Term candidates from the set M_{def} or M_{syn}, which are synonymous to a term from R, are added to R.

All pairs of synonyms from the set M_{syn}, whose overall frequency is in the range $[p_3,p_4]$ for percentiles with levels p_3=0.35 and p_4=0.95, calculated for overall frequencies of synonymous pairs, are added to R.

Term candidates from the set M_{gram} with frequency in the range $[p_1,p_2]$ are added to R, provided they have common words (at least one) with an element from current R.

The order of the steps was determined experimentally, as well as the levels of percentiles (p_1,p_2,p_3,p_4), but the levels may be regarded as parameters be changed by a user of the subject indexing system.

[5] https://github.com/ivanov-kir-m/SISTool/tree/master/Dictionaries

5. Experiments and discussion

The encountered problem for performing experiments is the lack of human-built indexes in many Russian educational texts (textbooks, tutorials, etc.) available in electronic form (whereas many printed books have them). So we have performed experiments with 5 medium-sized (about 20 thous. words) tutorials taken from the educational resourse[6]: they are devoted to programming languages (PL), programming systems (PS), heuristic search methods (HS) in artificial intelligence. All these textbooks contain back-of-the-end indexes constructed by their authors, we regarded them as etalon sets of terms and evaluated the quality of our term filtering procedure by recall, precision, and F-measure. For comparison, we also have processed and evaluated the manual devoted to academic writing (AW), since it can hardly be attributed to scientific or technical text. The results of the evaluation are shown in Table 1.

While measuring precision and recall we had to account cases when formally different term candidates denote the same concept, for example: *условная конструкция — условие*; *conditional construction—condition*)—we considered them as term variants.

Our filtering procedure significantly reduces the set of extracted term candidates, leaving in average about 8% of the terms. For 5 scientific texts, recall proved to be from 0.72 to 0.84, while precision varies from 0.56 to 0.77. The recall is sufficient for constructing subject indexes, and precision is acceptable, as well as F-measure. The low recall obtained for the manual on academic writing (the last row of the Table 1) is partially explained by lack of explicit definitions of certain important but relatively rare used terms (e.g., *аннотация—abstract*).

Table 1. Recall and Precision of term extraction and filtering

Text	Size (in words)	Number of terms		Precision (P)	Recall (R)	F-measure
		Extracted	Selected			
PL1	21,060	1,591	140 (8.80%)	0.74	**0.84**	**0.79**
PL2	14,322	1,012	169 (16.70%)	0.56	0.82	0.67
PL3	21,376	1,612	77 (4.78%)	**0.77**	0.72	0.75
HS	19,471	1,806	98 (5.43%)	0.71	0.74	0.73
PS	25,526	3,372	208 (6.17%)	0.70	0.81	0.75
Mean	**20,351**	**1,879**	**138 (7.34%)**	**0.70**	**0.79**	**0.74**
AW	11,699	1,884	67 (3.56%)	0.72	0.55	0.62

Our analysis of detected cases of incompleteness and inaccuracy of term extraction shows that the main reason relates to restrictions of the applied linguistic rules and lexico-syntactic patterns. In particular, certain terms are not extracted because of their complex or unusual grammatical structure (e.g., term *поиск вглубь* with pattern *N +Adverb*), which is not represented in the current collection of patterns. We also found in the texts complex phrases (with ellipsis) that define at the same time several terms, and corresponding phrase patterns are absent now. Another reason

[6] http://al.cmc.msu.ru/node/4

for low recall is incorrect tokenization of terms with hyphens and non-letter symbols (such as *И/ИЛИ-граф—AND/OR graph*), which leads to loss of the terms.

The analysis also shows that some extracted terms absent in the etalon subject indexes (such as term *logic programming* from the manual on Prolog) are terms relevant for indexing, and they may be omitted by human indexer because of his/her subjectivity or intent to get a more short index. So, in subject indexing task recall is more crucial than precision (provided that the number of extracted terms is not too large), since for human editor it is easier to discard some terms than to add new ones to subject index being constructed.

Overall, our method of term extraction and filtering considerably increases precision and recall in comparison with the known statistics-based methods [5, 6] and it also outperforms F-measure of the method [1]. At the same time, there are perspectives to improve the quality of term extraction, in particular, by accounting for more complex patterns and refinement of text tokenization.

Taking in mind that precision may depend on the size of processed text (the larger is text, the more terms are extracted), we have performed another experiment. Two texts (PL1 and HS) were divided into their section (chapters), which were processed and evaluated separately—the results are given in Table 2. The rows *Total index* contain scores for total indexes obtained after merging term sets extracted separately. One can notice that for the first text (PL1) the separate processing and merging give worse F-measure (0.64 instead of 0.79), but for the second one (HS), F-measure is slightly better (0.75 instead of 0.73), and in both cases recall increases. Therefore, the strategy of separate indexing of text sections and merging of resulted indexes seems reasonable (when sections are conceptually relatively independent) and may be chosen by a user of the indexing system.

Table 2. Evaluation of merging terms extracted from text sections

Text	Section	Size (words)	Number of terms		Precision (P)	Recall (R)	F-measure
			Extracted	Selected			
PL1	1	9,886	803	50 (6.23%)	0.83	0.71	0.76
	2	5,573	329	14 (4.25%)	0.58	0.40	0.47
	3	4,880	593	314 (52.95%)	0.42	0.87	0.56
	4	6,907	426	75 (17.60%)	0.67	0.83	0.74
	Total index				**0.50**	**0.89**	**0.64**
HS	1	4,150	523	37 (7.10%)	0.77	0.69	0.73
	2	10,853	1,062	100 (9.41%)	0.58	0.70	0.63
	3	4,468	536	23 (4.30%)	0.75	0.80	0.77
	Total index				**0.72**	**0.79**	**0.75**

6. Conclusions and Future Work

We have proposed and described the term extraction method for constructing back-of-the-book index of a given educational scientific document in Russian. The

method was experimentally evaluated, it demonstrates quite good performance (in average, 70–79% of F-measure) exceeding the widely-used standard methods, mainly due to the rules and lexico-syntactic patterns representing specific term usage in educational scientific texts. Thus, perspectives of rule-based methods for subject index construction of single documents seem encouraging.

The described method is implemented (with the aid of C# programming language) in a research prototype system supporting index construction. The user of the system can set parameters of the method, as well as indicate text fragment to be processed, and then verify and edit the results.

In order to accomplish more accurate and complete term extraction for subject indexing task, we evidently need to perform more experiments with texts. Future research directions are following:

- To elaborate additional lexico-syntactic patterns, in particular, patterns of non-standard phrases of term definitions;
- To improve the filtering procedure by experimenting with its parameters and the order of its steps;
- To develop methods for detecting and clustering synonymous variants of terms.

References

1. *Arora, C., Sabetzadeh, M., Briand, L., Zimmer, F.* (2016), Automated Extraction and Clustering of Requirements Glossary Terms, IEEE Transactions on Software Engineering, Vol.43, Issue 10, pp. 918–945.
2. *Aubin, S., Hamon, T.* (2006), Improving Term Extraction with Terminological Resources, Advances in Natural Language Processing, Springer Berlin Heidelberg, pp. 380–387.
3. *Bolshakova, E., Efremova, N., Noskov, A.* (2010), LSPL-Patterns as a Tool for Information Extraction from Natural Language Texts, New Trends in Classification and Data Mining, ITHEA, Sofia, pp. 110–118.
4. *Bolshakova E. I., Efremova N. E.* (2015), 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.
5. *Csomai, A., Mihalcea, R.* (2007), Investigations in Unsupervised Back-of-the-Book Indexing, Proc. of the Florida Artificial Intelligence Research Society Conference, pp. 211–216.
6. *Csomai, A., Mihalcea, R.* (2008), Linguistically Motivated Features for Enhanced Back-of-the Book Indexing, Proceedings Annual Conf. of the ACL, ACL/HLT, Vol. 8, pp. 932–940.
7. *Da Sylva, L.* (2013), Integrating Knowledge from Different Sources for Automatic Back-of-the-Book Indexing, Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI.
8. *Hasan, K. S., Ng, V.* (2014), Automatic keyphrase extraction: a survey of the state of the art, Proceedings of the 52th Annual Meeting of the ACL, pp. 1262–1273.

9. *Korkontzelos, I., Ananiadou, S.* (2014), Term Extraction. In: Oxford Handbook of Computational Linguistics (2nd Ed.), Oxford University Press, Oxford.
10. *Manning, C. D., Raghavan P., Schütze H.* (2008), Introduction to Information Retrieval, Cambridge University Press, New York, pp. 405–416.
11. *Matsuo, Y., Ishizuka, M.* (2004), Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information, International Journal on Artificial Intelligence Tools, 13 (1), pp. 157–169.
12. *Pecina, P., Schlesinger, P.* (2006), Combining Association Measures for Collocation Extraction, Proc. of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 651–658.
13. *Reinholt, K., Lukon, S., Juola, P.* (2010), A Machine-Aided Back-of-the-Book Indexer, Proceedings of DHCS 2010, Chicago, Illinois.
14. *Wu, Z. et al.* (2013), Can Back-of-the-Book Indexes be Automatically Created? Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, pp. 1745–1750.
15. *Zargayouna, H., El Mekki, T., Audibert, L., Nazarenko, A.* (2006), IndDoc: an Aid for the Back-of-the-Book Indexer, The Indexer, 25(2), pp. 122–125.
16. *Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.* (2008), A Comparative Evaluation of Term Recognition Algorithms, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pp.2108–2111.