DOI 10.1007/s10559-014-9677-6
Cybernetics and Systems Analysis, Vol. 50, No. 6, November, 2014

AN ITERATIVE APPROACH TO THE TERMINOLOGY EXTRACTION FROM UKRAINIAN-LANGUAGE SCIENTIFIC TEXT CORPORA

A. M. Glybovets[a†] and I. V. Reshetnov[a]

UDC 681.3:658.56

Abstract. *This article describes a combined method for the acquisition of valuable terms and relations from raw texts with the help of an iterative algorithm for automated terminology extraction from Ukrainian-language scientific texts. Special attention is paid to the analysis of lexicographical features of characteristic text fragments of documents. Specific features of Ukrainian-language documents are taken into account. The emphasis is on solving the applied problem of terminology acquisition from input texts in the widely used pdf format with obtaining output term relations in the RDF format.*

Keywords: *statistical method, lexicographic method, thesaurus, term, "general-particular" relation, hyponymy.*

INTRODUCTION

The creation and updating of specialized dictionaries lag behind the progress in investigations by virtue of the following objective reasons: the complexity of fields of study and variability of concepts with time [1]. At the same time, researchers have a very great need for mutual understanding at the conceptual level, which requires an available and unified terminological base and a qualitative retrieval system to search for terms used in scientific documents.

One of efficient methods for improving the relevance of search results of these systems lies in using a thesaurus [2]. Among methods for the construction of thesauri, an automated method is best suited for the field of scientific research by virtue of high rates of renewal of information and the ensuing high cost of participation of experts in this work. Within the framework of a number of investigations carried out at the Cathedra of Informatics of the National University of Kyiv-Mohyla Academy (NaUKMA) with a view to creating a retrieval system for scientific documents, the development of its component for the automated construction of thesauri will improve its quality.

The main objectives of this publication are the description and implementation of a method for extracting terminology from incoming scientific texts that underlain the above-mentioned system. This work analyzes existing approaches to the construction of thesauri and describes the developed method of automated determination of important Ukrainian-language terms and terminological relations between them. This method is implemented in the form a Web service. The analysis of efficiency of the method is based on real data of the scientific Ukrainian-language periodical press. The developed component became a natural component part of the retrieval system of Ukrainian-language scientific documents.

In developing the method, the limitedness of documentary data corpora produced in the Ukrainian language was taken into account, which has required the allowance for the possibility of iterative addition of scientific documents to terminological bases with the subsequent updating of the thesaurus content. The attention is focused on the description of solving the applied problem of construction of a terminology with the description of relations in the format RDF for incoming texts in the widely used format pdf.

[a]National University of Kyiv-Mohyla Academy, Kyiv, Ukraine, [†]*andriy@glybovets.com.ua.* Translated from Kibernetika i Sistemnyi Analiz, No. 6, pp. 53–62, November–December, 2014. Original article submitted July 3, 2014.

 1060-0396/14/5006-0866

1. A REVIEW OF EXISTING APPROACHES

1.1. The role of a thesaurus in information retrieval. A thesaurus is a controlled vocabulary containing semantic relations between terms and improving the process of searching for related terms [3].

Usually, information needs of a user of a retrieval system does not correspond to terms that occur in documents, or the user misunderstands the terminology of the knowledge domain in which he searches for information. Under such conditions, one of methods of improvement of search results is the use of thesauri of terms of object domains [4]. Thesauri are tables of terms and relations between terms with the specification of types of relations (NT, BT, USE, and RT) [3]. Information systems can use thesauri at the stage of indexation of documents to provide a more correct classification of documents into categories or during a search to extend a user retrieval request by adding related terms.

The main problem of compiling thesauri is that, for the majority of commercial databases that spread scientific information, such thesauri are created by experts in knowledge domains and also specialists in compiling thesauri. In newest knowledge domains such as bioinformatics or computer engineering in which terminology is in the making and a large number of new publications are issued, terminological dictionaries become outdated very quickly and should be updated more often, i.e., experts should be involved again. Contrary to this approach, there are methods of automated construction of thesauri that use all latest publications on some subject as their corpus and construct interrelations between terms on their basis. The updating of terminological relations with the help of such a system is much simpler and cheaper. In [2], basic methods of automated construction of thesauri are considered that have different efficiency and time complexity estimates and are based on different principles, namely, statistical and lexicographic. It is just the description of the development of a new method using a combination of ideas underlain these approaches which is the subject of this work.

In [5], the following definition of a thesaurus is given. A thesaurus is a lexico-semantic model of a conceptual reality or its representative that is expressed in the form of a system of terms and their interrelations, proposes access with the help of many aspects, and is used as a processing and search system within the module of an information retrieval system. Note that the author focuses attention on the fundamental integrity of the theoretical model of a thesaurus and practical application of program modules with the corresponding functionality.

1.2. RDF as a thesauri representation format. The format RDF is one of most widespread means of data and metadata representation for semantic web technologies. Simply speaking, this format is based on the idea of representation of information in the form of triplets "subject–predicate–object." This model that seems to be general and simple at first sight can successfully satisfy thesaurus needs for the description of thesaurus content. One more important distinctive feature of the format is its wide international support at the level of implementation of applied systems. As has been noted earlier, the role of a thesaurus is determined not only by the accuracy and scope of presented terminological relations but also by the practical applicability of a program module, ease of access, and suitability for electronic data processing. It is precisely the possibility of publication of data processed with the help of the thesaurus program component directly on the Internet in the generally accepted format that has allowed the authors to propose the format RDF and a supporting system of Web services with a program interface as the final format for access to the thesaurus.

In the authors' opinion, among concrete RDF specifications, the format JSON-LD presented in the ISO-25964 [6] standard corresponds to the best advantage to the stated problem of publication of thesaurus resources in the form of a Web service. The basic conceptions of the format are as follows [7]: international resource identifiers (IRI); a context that is used mainly for specifying IRI abbreviations; identifiers of nodes and typed values.

It suffices to use the proposed basic elements of the format in order that thesaurus data minimally satisfy the standard.

1.3. Automated methods for constructing thesauri. Methods of automated compiling thesauri can be divided into the following two basic classes: statistical methods that intensively use characteristics of term frequencies and positions in documents as the base for different models of finding out relations between terms, and lexicographic methods that use information from the field of processing human speech for implementing syntactic, morphological, and other types of analysis of texts to determine semantic relations from information obtained only from a text. Lexicographic methods usually use language corpora that are collected by experts and contain rules of general usage of words, word forms, and synonymic series. Implementations of many methods are promoted by software packages for carrying out the initial analysis of free texts. In turn, utilities indexing and ranking terms are similar tools for statistical methods.

Many statistical methods for searching for dependencies between terms are based on the creation of an index of terms providing the best description of contents of documents, which usually requires the ranking of terms according to their importance degrees. The most applied weighing technique developed for algorithms of retrieval systems is the use of term frequency (TF), inverse document frequency (IDF), and also their combinations.

In information retrieval, the method of joint use of terms is one of approaches to the formation of multiword terms [8]. In this method, the basic elements for computations are the frequency of occurrence of a term in definite contextual frameworks of different size such as the whole document, chapters of the document, paragraphs, and other elements. In this case, the smaller the distance between words in the context of the selected frame, the larger the measure of their joint use is assigned. Some authors have doubt about the quality of terminological relations found with the help of this method. For example, the inefficiency of a thesaurus formed by this method as applied to search problems is mentioned in [9]. The author of [9] proposes his own approach, namely, he introduces the concept of a conceptual space as a network of terms and weighted associations between them that can reflect concepts and relations between them in the corresponding information space represented in the form of a corpus of documents in a database. A model of associative search included in this method is close to mental methods for the representation of information needs of users of a retrieval system in the form of a network of terms and relations between them that are usually fuzzy.

Lexicographic methods of searching for relations between terms are based on the principle of the direct specification of relations between words with the help of language means, and the nature of a relation can be determined from the syntactic and lexical structure of sentences. From the viewpoint of lexicographic methods of searching for relations in a terminology, of interest is the method of compiling terminological word combinations as one of the most productive methods in word formation. Thus, the use of lexicographic methods for filling thesauri by terminological relations can be considered as most suitable as to its principles. In [9, 10], one of the most widespread problems of all statistical methods is mentioned, namely, the problematics of indexing phrasal terms or, in our case, terminological word combinations. In particular, the authors of the mentioned works alluded to the need for the construction of qualitative solutions based on linguistic distinctive features of texts including the use of the technique of part-of-speech tagging as one of main tasks of improving statistical methods in information retrieval.

In the next lexicographic method, the concept of hyponymy plays the key role. Hyponymy is a relation of inclusion between a genus and species in a lexico-semantic system. Generic words are called hyperonyms and specific ones are called hyponyms. It is obvious that the hyponymy phenomenon directly alludes to a relation of the type “general–particular” between terms and is an integral component part of thesauri. Developing this idea, the M. Hearst [11] created an automated lexicographic method for extracting hyponyms from texts.

Two main problems solved with the help of this approach are the elimination of the need for previously formed knowledge bases on an object domain and the possibility of applying the method to various text corpora. In [11], a set of lexico-syntactic patterns is specified that directly refer to the sought-for lexical dependencies that are easily recognized by software tools or even immediately in a text. The hypothesis of the method sustains the presence of a large amount of useful information about an object domain in the text itself that can be found out by both man and an algorithm without resorting to excessively concrete details of definite phenomena and things and without requiring any deep lexicographic or semantic analysis from a system.

This technique of searching for taxonomic relations was proposed by H. Alshawi [12]. He used a hierarchy of patterns for interpreting definitions consisted mainly of indicators of parts of speech and symbols-masks. The authors consider that the main drawback of this approach is the problem of selection of a set of patterns that would indicate the direction of a relation with the same accuracy in texts of different styles.

Summing up main achievements of the method, it makes sense to mention a comparative low cost of its application to the automated acquisition of semantic relations in documents. The method is positioned as an alternative to statistical methods and, in comparison with them, has the advantage of accuracy in processing rare relations between terms that occur sporadically in a text and cannot be successfully processed by statistical methods. The patterns and strategies of truncation of modifiers of nouns that are presented in the investigation do not pretend to be complete and leave definite freedom for future additions.

2. AN ITERATIVE COMBINED METHOD FOR CONSTRUCTING TERMINOLOGIES

This section describes the main stages of the proposed iterative method for constructing terminologies with the help of combining lexicographic and statistical methods.

2.1. Block diagram of the algorithm. The process of construction of a terminology from a text corpus can be divided into the following two basic steps: (1) extraction of a set of words that occur in texts of documents and correspond to terms in the knowledge domain of these documents; (2) definition of relations used in the thesaurus on the set of these terms.

The problem of meaning-based extraction of terms from the set of all words of a document is similar to the usual operation of indexing texts by retrieval systems, and it is this approach that is used in our method for obtaining an ordered list of unique words of a corpus with applying the term weighting technique TFIDF. Such a sequence begins with words that best characterize the contents of documents and, hence, are candidates for terms.

To restrict this list of words, an operator can be introduced that would make it possible to determine the boundary list element that is followed by ordinary words that are not terms.

For our method, this operator can have the following variations. “Stop list” is the operator eliminating the number of words that is specified by its parameter from the tail of the sequence. This approach uses one of popular methods of eliminating stop words in retrieval systems, but it remains sensitive to the size of a text corpus. The proportional operator is similar to the operator “stop list” but its parameter restricts the tail of the sequence to a definite percent of words that is based on the statistical distribution of the number of terms in corpora of scientific texts.

As a result of computational experiments, the following decision was made: apply the proportional approach to restrict the input list of terms proceeding from its advantages in processing unprepared text corpora.

Terms were estimated according to the metric of the document frequency of a reference corpus. It is obvious that the use of methods of restricting lists of words will be efficient only under the condition of applying a reliable weighing scheme, which, in turn, will depend in our case on the method of calculating the document frequency component of terms that is sensitive to the structure and size of corpora.

To provide reliable weighing, this work proposes to solve the problem of small text corpora by using and filling a reference system of document frequencies of terms. The reference system is based on the construction and indexation of a large and varied educational corpus of scientific texts with the subsequent storage of the obtained document frequencies as reference ones. The complete corpus of articles of the journal “Scientific Papers of NaUKMA” is proposed as the document base for this corpus.

After obtaining a first-priority list of terms for compiling a thesaurus, the character and direction of relations between terms should be determined.

We introduce the concept of a characteristic text fragment that is a direct occurrence of a term in a document in a definite context. From the variety of methods for the consideration of the context of using words, for example, parts of surrounding word combinations and turns of speech, sentences, and windows with fixed numbers of words, we have selected sentences as the basis for our investigations proceeding from available tools that would allow us to apply the methodology of tagging by parts of speech as the basis for lexicographic methods.

The next step consists of finding characteristic text fragments for all terms from the list. This search can be linearly performed but, allowing for the possibility of scaling the developed method, it is proposed to use one of retrieval systems with open source code that would return all documents from our corpus that contain a definite term and, thereby, to restrict the space of linear search. Then, linear search for characteristic fragments is performed among the found documents.

At the next stage of performing the method, all found characteristic fragments are analyzed using different methodologies to determine types of relations. The application of a simple method of joint use of terms in one characteristic fragment allows one to establish a relation between terms (RT) if they enter into characteristic text fragments together with the initial term, and the use of a set of definite lexicographic patterns makes it possible to find relations of the BT, NT, and RT types.

Next, we extend the thesaurus by adding terminological word combinations. The use of lexicographic patterns is based on the method of finding terminological word combinations that, in turn, allows one to select not only single-word terms but also multi-word terms whenever patterns match text fragments. Of course, there are a lot more terms of the second type. Thus, a by-product of applying lexicographic patterns is the extension of the top-priority list of terms by adding terminological word combinations. This could not be reached at the first stage with the help of indexing within the framework of used tools.

To use the lexicographic patterns defined by us, we introduce the following formal notation. A Lexicographic Pattern (LP) is an ordered list of matching operators. A matching operator is a command that requires the execution of the operation of searching for matches of the type Noun Phrase (NP), Exact Word (EW), or a symbol from a synonymic series.

NP is a matching operator that searches for a noun phrase by applying the list of parts of speech matching rules that are specified for each of such operators. As a result, it returns all noun phrases found in a phrase in the order of specified parts of speech matching rules and also positions of the noun phrases found in the phrase. For matching operators of this type, their role (the indices 1 and 0) can be specified as their parameter.

The role of an *NP* operator is specified by the indices 1 or 0 that allude to the leading or secondary role of this operator in a pattern (which is written as NP_1 or NP_0).

EW is a matching operator that searches for occurrences of a concrete symbol or a word in a phrase from a list of possible alternatives and returns positions of occurrences of such words.

W is a window operator that specifies minimal and maximal window frames and plays the role of a matching mask for any sequences of words in a sentence.

IT is an iteration operator that denotes a repetitive sequence of operators in a pattern.

A convergence rule (MR) is a given sequence of tags of parts of speech to which should correspond a subsequence of words in a sentence.

Tags of parts of speech (N, A, and P) are parameters of a configuration of matching rules for extracting terminological word combinations denoting a noun (N), an adjective (A), and a preposition (P), respectively.

The satisfaction of a pattern lies in finding a set of subsequences of words satisfying matching operators; each position of such a subsequence is consistent with both the order of occurrence in a phrase and the position of an operator defined in the pattern. All possible matches for particular operators must be combined into the resulting sets.

For example, to fix a lexicographic pattern responsible for direct definitions with using a dash in our formal notation, we should write

$$LP = (NP_0\,(MR < A, N >), EW(\text{“—”}, \text{”-”}), NP_1\,(MR < N, N >)).$$

The following phrase satisfies this pattern: “A sociological investigation – a system of procedures for obtaining scientific knowledge on social phenomena and processes.” In this case, to the first matching operator corresponds the terminological word combination “sociological investigation,” the word matching operator has two alternatives, namely, a dash symbol and also a hyphen for processing the cases of replacement of this symbol in the initial text, and to the last operator corresponds the word combination “a system of procedures.”

Thus, matching operators of the type *EW* in a pattern play the role of fixed points of the pattern, whereas *NP* operators play the role of variables that are filled in extracting word combinations from phrases during the satisfaction of the pattern.

We will interpret relations on the basis of a text that matches a pattern. In compiling such a pattern, the value of the leading or secondary role in the pattern is additionally assigned to the corresponding parameter of *NP* operators that determines the relations between the matches obtained for these *NP* as follows: the relation BT is established between the matches obtained for NP_0 and NP_1; the relation *NT* is established between the matches obtained for NP_1 and NP_0, and the relation *RT* is established between the matches obtained for operators with identical roles.

This interpretation is based on the fact that, in the majority of patterns, at the corresponding places of terminological word combinations, in parts of sentences are either homogeneous attributes, or adjectival nouns, or generalizing words, or, for example, in the case when direct definitions in a text are matched with a pattern, a term and its generic membership. Thus, when a text fragment matches a pattern, the direction of a relation is correctly determined.

To implement the Hearst algorithm for searching hyponyms, we should first teach the system to recognize phrasal word combinations. The proposed approach consists of fixing nouns in a sentence with subsequent selection of surrounding words according to rules.

Taking into account the similarity of the scientific style of formulating definitions in many languages, the idea of localization of the patterns developed by M. A. Hearst for the Ukrainian language with adding new patterns seems to be successful.

To restrict the framework of the investigation and to reach a definite result for specific and, at the same time, most used methods for creating a terminology, only terms-nouns and noun phrases are used.

The following categories are selected from patterns responsible for relations between terms in a sentence:

- direct definitions and definitions using punctuation symbols and patchwords inherent in the Ukrainian language;
- Hearst patterns;
- a pattern for the denotation of part–whole relations.

All the presented patterns are extended by adding synonymous and similarly used words in pattern formulas. In matching a pattern with sentences, all words are reduced to their normal form, which makes it possible to reduce the necessary number of pattern variations. Thoroughly developed patterns are presented in Table 1.

TABLE 1. List of the Developed Lexicographic Patterns in Formal Notation

Name	Formal Notation of Pattern Rules
MR1-9	$MR<NPNN>,\ MR<ANNN>,\ MR<ANAN>,\ MR<ANN>,$ $MR<NAN>,\ MR<NN>,\ MR<AN>,\ MR<N>$
LP1	$NP1, EW<'-'\ |'-'>, EW<$ ‘це’ |‘є’|‘означає‘ |‘вважається‘, $NP0$
LP2	$EW<$’такий‘$>, NP1, EW<$‘як’$>, \{ITNP0, EW<','>\}, EW<$’’і |’або’	’й’	’та’$>, NP0$
LP3	$NP0, ITEW<',', NP0, EW<$’і’|’або’|’й’|’та’ $>, EW<$ ‘інший’$>, NP1$
LP4	$NP1, EW<',', EW<$ ’включаючи’|’а саме’|’зокрема’|’особливо’$>,$ $ITNP0, EW<',', EW<$ ’і’|’або’|’й’|’та’$>, NP0$
LP5	$NP0, W<0,3>, EW<$ ’бути частиною’|’входити в’$>, W<0,3>, NP1$
LP6	$NP1, W<0,3>, EW<$ ‘складатися з’|‘підрозділятися на’ $>, W<0,3>,$ $ITNP0, EW<',', EW<$ ’і’|’або’|’й’|’та’$>, NP0$

In applying rules of a pattern, their order is taken into account and, therefore, nouns are first found as matched elements of a word combination with a large number of words and, hence, more rarely used elements.

2.2. Mathematical model and algorithmic formalization of the method. We introduce the following notations: D is a set of text documents, LP is a set of lexicographic patterns, T is a set of thesaurus terms, T_F is a list of important single-word terms of a corpus, which is sorted out using the TFIDF metric and bounded by a function $limit(T)$, T_E is a set of multi-word terminological word combinations, and R is the set of relations of the thesaurus, $R_i \in \{(T_1, T_2, Rl)\}$, where $Rl \in \{RT, BT, NT\}$ and $T_1, T_2 \in T_t$ is the a set of characteristic text fragments for a term t, S_C is a set of sentences of a characteristic text fragment C, Lem_S is a set of lemmatized words of a sentence S, and M_{lp} is a sequence of terminological word combinations that match a lexicographic pattern.

We also introduce the following functions:

$lm(T): \{t \mid t \in T_S\} \to \{t' \mid t' \in T_F\}$ is a function of restricting a sorted list of terms;

$extract(d): D \to \{t \mid t \in T\}$ is a function of extracting terms from a document;

$sort(T, d): \{T\} \times D \to (t')_1^{|T|}, \forall t_i, t_j, i \geq j \Leftrightarrow tf(t_i, d) \cdot idf(t_i) \geq f(t_j, d) \cdot idf(t_j)$ is a function that constructs a sequence of sorted terms of a document in the order of decreasing the TFIDF metric;

$tf(t, d): T \times D \to R$ is a function computing term frequencies in a document;

$idf(t): T \to R$ is a function that associates with each term its inverted document frequency from a reference corpus;

$findCF(t): T \to \{c \in C_t\}$ is a function searching for characteristic fragments of a term;

$split(c): C_t \to \{s \mid s \in S_C\}$ is a function partitioning a characteristic text fragment into sentences;

$lem(s): S_C \to (lem \mid lem \in Lem_S)$ is a function of extracting a sequence of lemmas from a sentence;

$match(lp, s): LP \times S_C \to \{m \mid m \in M_{lp}\}$ is a function of pattern satisfaction that returns a set of sequences of matched terminological word combinations in the sequence order of pattern positions;

$inrs(M_{lp}): \{m \mid m \in M_{lp}\} \to R$ is a function of establishing relations over a set of sequences of pattern matches.

The developed method for the construction of a thesaurus can be presented by the algorithm shown in Fig. 1. The following formalization of matching rules for a lexicographic pattern is also used:

$LP = \{(pe)_1^t \mid pe \in PE\}$ is a set of lexicographic patterns specified as a set of pattern elements.

$PE = \{NP_0, NP_1, EW, W, IT\}$ are pattern elements;

$NP_0 = \{((mr)_1^m, 0) \mid mr \in MR\}; NP_1 = \{((mr)_1^m, 1) \mid mr \in MR\}$ are sets of commands of searching for terminological word combinations with the indication of the leading (1) or secondary (0) roles of a word combination in a pattern;

$MR = \{(tag)_1^k \mid tag \in \{'N', 'A', 'P'\}\}$ is a set of matching rules specified by sequences of tags of parts of speech;

$EW = \{(ew)_1^n \mid ew \in Lem\}$ is a set of commands of searching for an exact word match on sequences of alternatives of lemmas;

$W = \{(\min, \max) \mid \min, \max \in N\}$ is a set of commands of searching for windows that is specified by pairs of minimal and maximal window lengths in a sentence;

$IT = \{(it)_1^t) \mid it \in PE\}$ is a set of commands of searching for iterations that is specified over subsequences of pattern elements;

1. $\forall d \in D$:
   $T_i := extract(d)$
   $(t)_{i=1}^{|T_i|} := sort(T_i, d)$
   $T_F := T_F \bigcup lm((t)_{i=1}^{|T_i|})$
2. $T_F := lm(sort(T_F))$
3. $\forall t \in T_F$:
   $C_t := findCF(t)$
   $\forall c \in C_t$:
   $S_C := split(c)$
   $\forall s \in S_C$:
   a. $\forall lem \in lem(s)$, если $lem \in T_F \Rightarrow R := R \bigcup (t, lem, RT) \bigcup (lem, t, RT)$
   b. $\forall lp \in LP$:
      1. $M_{lp} := match(lp, s)$
      2. $R_{lp} := inrs(M_{lp}); R := R \bigcup R_{lp}$
      3. $T_E := T_E \bigcup terms(M_{lp})$
4. $T := T_F \bigcup T_E$
5. $Thesaurus := (T, R)$

Fig. 1. Algorithm for constructing a terminology.

$P_M = \{((l)_1^v, p) \mid l \in Lem,\ p \in N\}$ is a set of phrase matches specified by pairs of lemma sequences and positions of the first lemma;

$M_{lp} = (p)_1^v \mid \exists\, lp = (pe)_1^l, \forall s \in S, \forall p_i \in apply(pe_j, s),\ pe_j \in \{NP_0,\ NP_1\}$ is a sequence of phrase matches for NP_0 and NP_1 operators of a pattern;

$apply(pe, s): PE \times S \to \{p \mid p \in P_M\}$ is a matching function for a pattern element and a phrase that associates a set of phrase matches with the pattern element;

$match(lp, s): LP \times S_C \to \{m' \mid m' \in M_{lp}\}$ when $\forall lp = (pe)_1^l$, $\exists (m)_1^l \mid m \in apply(pe_i, s)$ such that $\forall m_i, m_j$, $i < j \Leftrightarrow m_i = ((l)_1^{vi}, p_i)$, $m_j = ((l)_1^{vj}, p_j)$, $p_i \le p_j$, $\{m'_{\overline{1,n}}\} \subseteq \{m_{\overline{1,l}}\}$

$$inrs(M_{lp}): \{m \mid m \in M_{lp}\} \to \left\{ r \left| \left\langle \begin{array}{l} r = (T_1, T_2, BT) \mid \exists lp = (pe)_1^l, \exists s_1, \exists pm_1 \in apply(pe_i, s_1),\ pm_1 = (T_1, p_1), \\ \exists s_2, \exists pm_2 \in apply(pe_j, s_2),\ pm_2 = (T_2, p_2),\ pe_i \in NP_1,\ pe_j \in NP_0, \\ r = (T_1, T_2, NT) \mid pe_i \in NP_0,\ pe_j \in NP_1 \\ r = (T_1, T_2, RT) \mid (pe_i, pe_j \in NP_0) \cup (pe_i, pe_j \in NP_1) \end{array} \right\rangle \right. \right\}$$

CONCLUSIONS

This article describes the solution of the problem of iterative construction of the terminology used in Ukrainian-language scientific text corpora. Based on the proposed method and developed algorithm, a program module is created in the form of a Web service with possibilities of constructing thesauri in the RDF format from initial texts in the pdf format. The format JSON-LD of thesauri is chosen with allowance for the possibility of publication of the obtained terminological relations in the standardized form for accessing network resources and from the viewpoint of understanding a thesaurus as a ready-made software module of a retrieval system for scientific materials. Among types of relations between terms for search procedures, the preference is given to relations “general–particular” that were determined with the help of lexicographic analysis of sentences of texts with a view to finding out hyponymic relations between terms.

The developed module for constructing thesauri is based on a method described in this work and destined for searching for important terms and relations in a text. The first stage of this method that provides the search for important terms in document corpora is realized using the proposed method of weighing, sorting, and filtering terms of documents with the help of the metric of document frequency of a reference corpus. As such a corpus, the archive of the Ukrainian-language journal "Scientific Papers of NaUKMA" was used that underlies the constructed reference index of document frequencies of terms.

The second stage of the developed method realizes the application of lexicographic patterns for searching for hyponymic relations in initial texts. To provide successful implementation, an open source software support was used for solving utilitarian problems of lemmatization of terms and tagging words of sentences according to parts speeches, and also the lexicographic patterns proposed by M. Hearst [11] were adapted to the Ukrainian-language rules of word usage. The authors of the present publication have developed an extensible software package with the functionality of controlling the application of lexicographic patterns.

A test of the existing implementation of the proposed method on the basis of thematic corpora of scientific texts has shown the efficiency of the first stage of the algorithm and also an adequate accuracy of the second stage with the developed patterns. All limitations of the lexicographic method of searching for hyponymy make impossible the attainment of the completeness of a search for relations in a text because of the single-valuedness of contexts of terminological relations used in patterns and a low statistical frequency of their occurrence in the text. The problem can be eliminated by increasing the number of patterns and extending synonymic series determining word patterns, which requires the involvement of experts in lexicography, and also by improving the method of tagging parts of speech with the help of stochastic methods of eliminating ambiguity in determining parts of speech of some words.

The obtained program module has shown its usability as applied to test data corpora and can be used as a component part of a retrieval system for scientific materials.

REFERENCES

1. S. I. Landau, Dictionaries: The Art and Craft of Lexicography [Ukrainian translation], K.I.S., Kyiv (2012).
2. M. Lassi, Automatic Thesaurus Construction, University College of Boras, Sweden (2002), http://www.academia.edu/506142/Automatic_thesaurus_construction.
3. Types of Relations in a Thesaurus, Web. 5/10/2014, http://publish.uwo.ca/~craven/677/thesaur/main06.htm.
4. H. Chen, T. D. Ng, J. Martinez, and B. Schatz, "A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system," J. Amer. Soc. for Inform. Sci. (1997), http://arizona.openrepository. com/arizona/bitstream/10150/105991/1/chen21.pdf.
5. U. Miller, "Thesaurus construction: Problems and their roots," Inform. Proc. & Management, **33**, No. 4, 481–493 (1997).
6. "ISO 25964 — the International Standard for Thesauri and Interoperability with Other Vocabularies," ISO 25964 Thesaurus Schemas, Web. 08 April 2014, http://www.niso.org/schemas/iso25964/.
7. JSON-LD 1.0, Web. 08 June 2014, http://www.w3.org/TR/json-ld/.
8. H. Chen, T. Yim, D. Fye, and B. Schatz, "Automatic thesaurus generation for an electronic community system," J. Amer. Soc. for Inform. Sci., **46**, No. 3, 175–193 (1995).
9. H. Chen, K. Lynch, K. Basu, and T. D. Ng, "Generating, integrating, and activating thesauri for concept-based document retrieval," IEEE Expert., **8**, No. 2, 25–34 (1993).
10. G. Grefenstette, Automatic Thesaurus Generation from Raw Text Using Knowledge-Poor Techniques, Rank Xerox Research Centre (1993), http://www.academia.edu/4186829/AUTOMATIC_THESAURUS_GENERATION_FROM_RAW_TEXT_USING_KNOWLEDGE-POOR_TECHNIQUES.
11. M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in: Proc. 14th Conf. on Comput. Ling. (COLING '92), **2**, (1992), pp. 539–545.
12. H. Alshawi, "Processing dictionary definitions with phrasal pattern hierarchies," Comput. Ling., **13**, Nos. 3–4, 195–202 (1987).