UDC 004.93:159.95
DOI: 10.15587/2312-8372.2018.149962

Bisikalo O.,
Yahimovich A.,
Yahimovich Y.

DEVELOPMENT OF THE METHOD FOR FILTERING VERBAL NOISE WHILE SEARCH KEYWORDS FOR THE ENGLISH TEXT

Об'єктом дослідження є процес обробки вербальної інформації для визначення ключових слів в тексті. Найважливішим етапом пошуку ключових термінів є розрахунок їх ваг в аналізованому документі, що дозволяє оцінити їх значущість відносно один одного в даному контексті. Для вирішення цього завдання існує багато підходів, які умовно діляться на дві групи: вимагають навчання і не потребують навчання. Під навчанням мається на увазі необхідність попередньої обробки вихідного корпусу текстів з метою вилучення інформації про частоту зустрічальності термінів у всьому корпусі. Альтернативним підходом є використання лінгвістичних онтологій, які є більш-менш наближеними моделями існуючого набору слів заданої мови. На базі обох підходів були створені системи для автоматичної екстракції ключових термінів. Тим не менш у напрямку пошуку ключових слів не припиняються дослідження з метою підвищення точності і повноти результатів, а також з метою використання методів вилучення інформації з тексту для вирішення нових завдань.

Охарактеризовано існуючі підходи до визначення ключових слів. Краща якість обробки тексту досягається лінгвістичними методами або ж при їх комбінації зі статистичними. Систему автоматичного визначення ключових фраз з тексту природною мовою слід розробляти з використанням морфологічного словника і синтаксичних правил.

У ході дослідження використовується підхід до визначення ключових слів, який базується на знаходженні синтаксичних зв'язків між словоформами у реченнях англомовного тексту за допомогою інструментальних можливостей сучасних лінгвістичних пакетів. У межах загального підходу зменшення вербального шуму у методі, що пропонується, досягнуто за допомогою формалізованих операцій: заміна займенників на відповідні до них іменники; вилучення шумових зв'язків; вилучення шумових слів; вилучення стоп слів. Описані операції можна використовувати як додаткові модулі, що покращують результати знаходження ключових слів як для розробленого методу визначення ключових слів англомовного тексту, так і для інших алгоритмів знаходження ключових слів.

Ключові слова: *фільтрація вербального шуму, ключові слова англомовного тексту, лінгвістичний пакет, DKPro Core, синтаксичний аналіз.*

1. Introduction

At present, the volume and dynamics of information to be processed in lexicography and terminology, as well as in information retrieval tasks, make the task of automatically determining keywords especially important. Very actively in modern information technologies (IT) they use keywords to create and develop terminological resources for efficient processing of documents, in particular, indexing, summarization, clustering and classification [1].

There are a large number of automatic keyword extraction systems available that are designed and developed for processing natural languages. These systems are based on certain methods for determining keywords, which are divided into linguistic and statistical. Linguistic methods are based on the meanings of words, in particular, they use ontologies and semantic data about a word. These methods are resource-intensive at the early stages: development of ontologies, for example, is a very labor-intensive process [1]. On the other hand, statistical methods are accompanied by significant amounts of «verbal noise», which significantly affects the quality of the definition of keywords. Therefore, hybrid methods are the most promising for research, for which the speed of statistical text processing is enhanced by the capabilities of modern linguistic packages.

The relevance and practical value of the research direction is that the found keywords can be used to improve the accuracy of the analysis of site content and raise the position of the site in search results.

Keyword – a word in the text that can, in conjunction with other keywords, represent the text. The set of keywords is close to the annotation, plan and outline, which also represent a document with less detail, but, unlike keywords, associated with syntactic structures.

Verbal noise or noise words – a term from the theory of information search by keywords. These are words that do not carry a semantic load, so their use and role for the search is irrelevant [2].

In the process of processing, an exception is made to the words from the text under study, which, by definition, cannot be meaningful to what constitutes «noise». In contrast to the key, these words are called neutral or stop (stop words). These are words related to the official parts of speech, as well as pronouns [3].

2. The object of research and its technological audit

The object of research is the process of processing verbal information to identify keywords in the text.

ISSN 2226-3780

The subject of research is methods for finding keywords in the text, as well as approaches to reducing verbal noise in the process of searching for keywords.

Keywords have a number of essential features:

– high degree of repeatability of these words in the text, the frequency of their use;

– ability of a sign (words as verbal signs of a certain concept) to condense, collapse information, expressed in whole text, to combine «its main content». This feature is particularly pronounced in the keywords in the title position.

Having a properly selected set of keywords will allow to:

a) quickly find the article to the user when searching the database;

b) to see the article when viewing other similar articles;

c) rather, understand the thematic and terminological area of both one article and the journal as a whole.

All this serves one purpose: to attract the attention of readers to the article, which is the main task of any media [4].

However, the choice of keywords is a very difficult operation and requires a balanced approach. It is necessary to choose the keywords that most accurately reflect the specifics of the topic in question. It is necessary to avoid random and common phrases, it is not recommended to repeat the same keywords several times. So, the process of searching for keywords is analytical [5].

3. The aim and objectives of research

The aim of research is improvement of the accuracy of determining keywords from English text based on the development of a method for reducing the influence of verbal noise.

To achieve this aim it is necessary to solve the following objectives:

1. To consider approaches to reduce verbal noise when finding keywords.

2. To calculate the numerical indicators of the connections between words and analyze the results obtained as the basis of the method.

3. To formalize the operations for each stage of the method and determine the quantitative characteristics of the relevance of the results obtained in comparison with analogues.

4. Research of existing solutions of the problem

Among the main directions of solving the problem of searching keywords in the text, identified in the resources of the world scientific periodicals, can be highlighted [6, 7]. To separate single keywords using methods based on Zipf's law. Such methods depend on the setting of the range of frequencies in which words significant for the text are found. Since words that occur very often, basically turn out to be verbal noise, and words that occur rarely, in most cases, do not have a decisive semantic meaning. Therefore, in each case, it is necessary to use a number of heuristics to determine the width of the range, as well as techniques that reduce the influence of this width. One of the ways, as indicated in [8], is exceptions, with candidates for keywords, words that can't be meaningful to the volume components of the noise. But in this paper, noise reduction based on syntactic information is not considered.

The work [9] is devoted to improving the results of calculating the weights of terms based on the TF-IDF algorithm. However, a common feature of such systems is that they require the availability of information obtained from the entire collection of documents. In other words, if the method based on TF-IDF is used to create a document view, then the arrival of a new document in the collection requires recalculation of the term weights in all documents. So, any applications based on the weights of the terms in the document will also be affected. This greatly hinders the use of methods for extracting key terms that require learning in systems where dynamic data flows must be processed in real time [10].

To solve this problem, the TF-ICF algorithm is proposed in [11]. As a development of this idea in [12, 13] it is proposed to use Wikipedia as a learning thesaurus. For calculations, the information contained in the annotated encyclopedia articles with manually selected key terms is used. However, the order of passage of terms in the document and their syntactic role are not taken into account.

An alternative solution to the problem, outlined in [14], involves the use of linguistic ontologies that are more or less approximate models of the existing set of words of a given language. However, these methods are resource-intensive at the early stages: the development of ontologies is a very laborious process.

A method that serves to automatically form a thematic body frame with a WEB is shown in [15]. However, the selection is governed by the timing relationship threshold.

The authors of work [16] emphasize the importance of using nominal groups selected with a parser as candidates for keywords. Although this statement may be considered by other syntactic units used in the definition of keywords.

Seotool is a free online service that will help check the relevant written text of the key words (automatically generating the keys for the specified text). This will help to get a higher ranking in the search engines Yandex and Google, since the page will be the keywords that correspond to the content of the page on which they are placed. Also, this service will help in generating the semantic core of the site (when enabled, remove the HTML code). However, in the generation of keywords and phrases only the first thousand words of the entered text are used.

There is a possibility of a percentage comparison of words with a template. The words of the analyzed text (content) will be compared as a percentage with the list of words of the entire template (text) by morphological analysis. If the percentage equality with any of the words in the template is taken into account, the word is taken into account, otherwise it is not taken into account. The maximum number of words of a pattern should not exceed 250 words [17].

Rise-Top will help to make «sketches» of keywords for the site based on the use of the specified text for the analysis. As a selection of keywords, the words with the highest density in the order of decreasing their density are applied to the entire text [18]. But in the generation of keywords, only the first 1000 words of the processed text are also used.

Advego is the largest provider of content and related services for Internet sites in RuNet. For optimizers and site owners are offered unique articles, reviews, publications. Promotion in search engines and promotion in social networks are provided. The resource also has the ability to define keywords [19].

Thus, the results of the analysis allow to conclude that the question of developing a method for filtering

verbal noise in the process of searching for keywords is promising and requires further study.

5. Methods of research

To improve the accuracy of determining keywords, statistical text processing methods are involved, the speed of which is enhanced by the capabilities of modern linguistic packages.

One such package is the DKPro Core, a set of software components for natural language processing, based on the Apache UIMA framework.

The DKPro Core package is more than a set of analysis components that interact with each other. It was built to improve the productivity of researchers working with automatic language analysis. The approach of DKPro Core is that researchers should be able to focus on their real scientific issues, and not on the development of appropriate technologies [20].

The quantitative characteristics of the relevance of the results, based on the analysis of the literature, are completeness (in Jacquard and absolute) and accuracy (in Euclidean and Manhattan distances). The interpretation of the selected criteria to the conditions of the task of defining keywords is carried out.

Jacquard completeness, in this case, is determined for two sets of keywords – given by the author (reference) and programmatically defined, equal to the ratio of the number of elements of the intersection of these sets with the number of elements of their union. That is, it is the quotient of the division, where the numerator contains the number of keywords correctly found by the program, and the denominator is the difference between the sum of the elements in the two sets and the number of keywords correctly found.

Absolute completeness is found as a ratio of the number of keywords correctly found by the program with the number of keywords.

The Euclidean distance is determined by the formula:

$$d_e = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

where n – the number of keywords; x_i – the position of the i-th keyword defined by the author; y_i – the position of the i-th keyword defined programmatically.

Manhattan distance is determined by the formula:

$$d_m = \sum_{i=1}^{n} |x_i - y_i|.$$

The use of a pair of formal criteria for completeness and accuracy, will allow a more objective assessment of the relevance of the obtained search results for keywords.

6. Research results

According to [21], this approach to defining keywords is proposed, there are three main stages:

1) creating a multi-level markup of the text;

2) the use of syntactic markup, taking into account the complex dependencies between pairs of lemmas;

3) reduction of verbal noise.

The essence of the approach, in contrast to the known analogues, is determination of the number of links for individual words and the subsequent selection of the first n words with the largest number of links, where n is the number of necessary keywords.

Creation of multilevel markup of text and syntactic markup, taking into account the complex dependencies between pairs of lemmas, is achieved by means of DKPro Core [20].

Verbal noise filtering is proposed to provide with the following operations:

– replacing pronouns with their corresponding nouns;
– removal of noise connections;
– removal of noise words;
– withdrawal of stop words.

Replacing pronouns with their corresponding nouns (replace pronouns) allows reducing the number of pronouns, as well as increasing the number of nouns that can be keywords. For the method of reducing verbal noise when defining key words in English text, it is proposed that the replacement of pronouns is performed by means of DKPro Core [20].

Let's consider the removal of phrases with types of links that do not carry significant semantic load. As a result of the research, it is revealed that such connections are DET, EXPL, FIXED, PUNCT, REF, ROOT.

DET – the connection of the determinant that exists between the nominally main word and its determinant. Most often, a word that has a tag part of a DET speech will have the same DET identifier connection and vice versa. A well-known exception is that in some of the data sets, the possessive determinant (for example, such as «my») at some point receives the tag of a part of the DET speech, but the NMOD link, which is parallel to other possessive constructions. But this is not completely the same for different languages, in some languages it is much clearer than in English, it is expressed how the possessive determinants relate to adjectives, therefore the relation NMOD is not subject to doubt [22].

Examples of DET links are shown in Fig. 1.

EXPL is a relationship that fixes plugin or pleonastic values. Such nominal values appear in the argument position of the predicate, but do not fulfill any of the predicate's semantic roles. The main sentence predicate (verb or predicative adjective or noun) is the main word. In English, this applies to some ways of using it and there: existential there, as well as it when used in exhibition constructions [23].

Some languages have no such English-like expressions, this applies to most pro-drop languages (a speech in which certain classes of pronouns can be omitted when they are pragmatic or grammatically inertial). Also, this phenomenon is often referred to as zero or zero anaphora [24]. In languages with similar utterances, they can be located where the main argument usually appears: the subject and the direct (and even indirect) application [25].

Examples of EXPL links are shown in Fig. 2.

FIXED is used for certain constant grammatical expressions that behave as functional words or short adverbs.

The verbose expressions are annotated in an equal structure, where all subsequent words in the expression are attached to the first one using a permanent label. The assumption is that these expressions do not have an internal syntactic structure (except from a historical point of view) and the structural annotation is in principle arbitrary. However, in practice, it is very important to use the consistent instruction of all constant verbose expressions in all languages [26].

Examples of FIXED links are shown in Fig. 3.

ISSN 2226-3780

PUNCT is used to denote any part of the punctuation in a sentence or part of the text if the punctuation is stored in typed dependencies.

PUNCT ratio tokens are always attached to the content of words and can never have dependencies. Since PUNCT is not a normal dependency relationship, the usual criteria for defining a headword are not applied. But the following principles are used:

1. A punctuation mark separating coordinated units is added to the following link.

2. A punctuation mark preceding or following an independent unit is attached to this unit.

3. Within the relevant division, the punctuation mark is attached to the highest possible node that maintains perspective.

4. Paired punctuation marks (for example, quotations and brackets, sometimes also hyphens, commas, etc.) should be attached to one word, if this does not violate the perspective [27].

Examples of PUNCT links are shown in Fig. 4.

REF is referent of the main word to the noun phrase, which is a relative word that introduces a relative position by modifying the noun phrase. For example, for the sentence: «I saw the book which you bought», the REF link will be between the words book and which [25].

ROOT is root grammatical relation, indicates the root of a sentence. The fake ROOT node is used as the main node. The ROOT node has an index of 0, since the indexation of real words in a sentence begins with 1. There should be only one root node in each tree. If the main predicate is absent, but there are many single dependencies, then one of them rises to the position of the main (root) one, and other singles join it [28].

An example of a ROOT link is shown in Fig. 5.

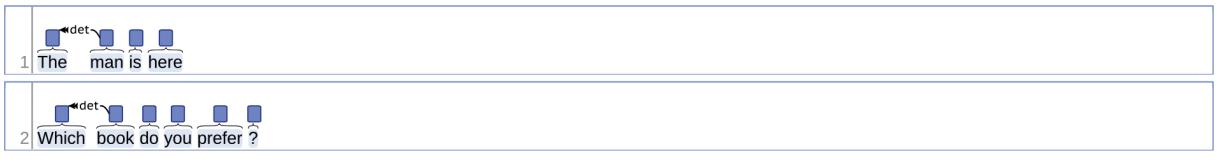

Fig. 1. Examples of DET noise links

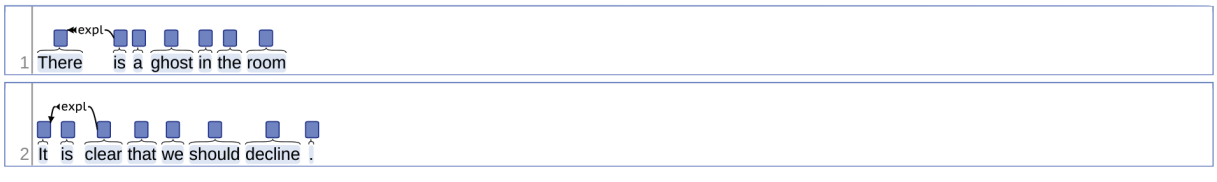

Fig. 2. Examples of EXPL noise links

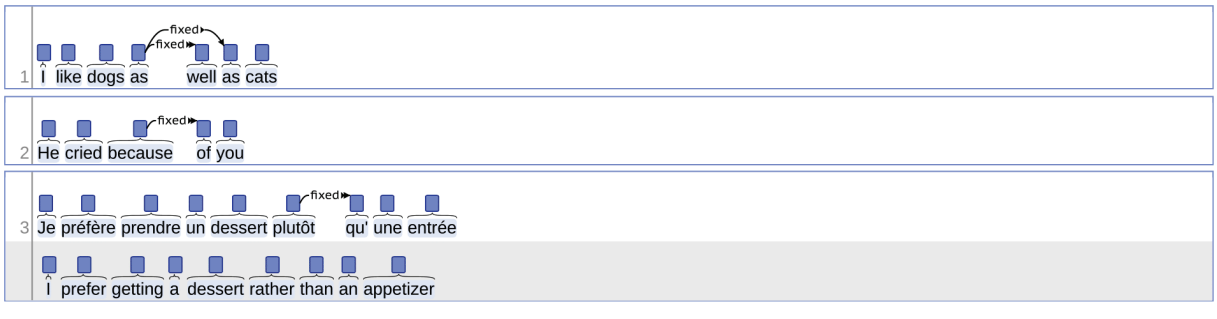

Fig. 3. Examples of FIXED noise links

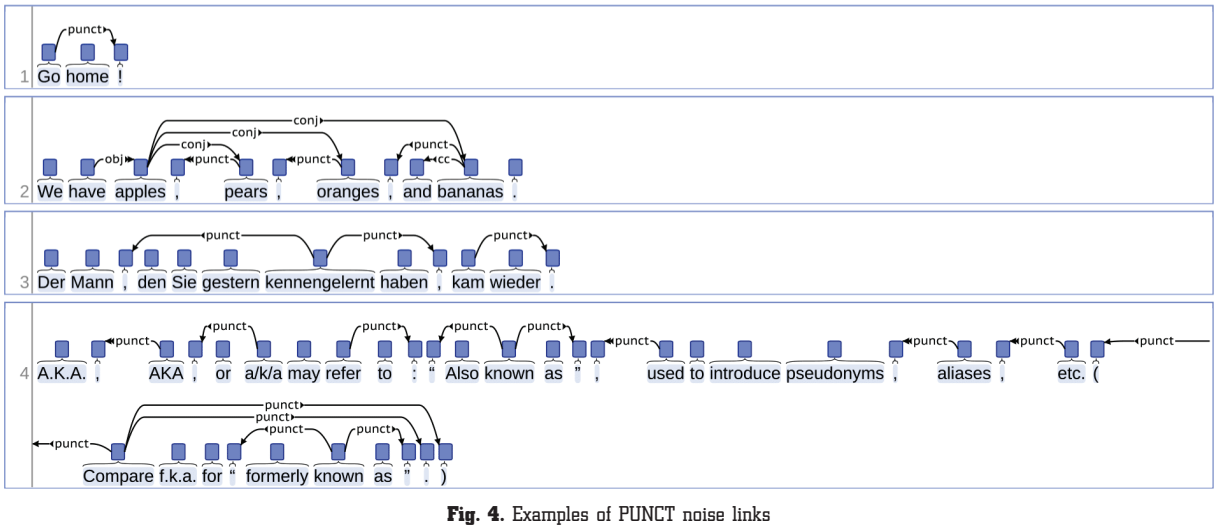

Fig. 4. Examples of PUNCT noise links

Fig. 5. Examples of ROOT noise links

Let's consider deleting noise words related to non-informative parts of speech, have tags: CC, CD, DT, EX, IN, LS, MD, PDT, POS, PRP, PRP$, RP, SYM, TO, UH, WDT, WP, WP$, WRB, -LRB-, -RRB-.

CC – coordinating combinations: and, but, nor, or, yet, plus, minus, less, times (multiplication), over (division), also for (because), so (i. e., so that), &, 'n, both, either, et, neither, therefore, v., versus, vs., whether.

CD – number, count, quantity: one, two, 2, mid-1890, nine-thirty, forty-two, one-tenth, ten, million, 0.5, forty-seven, 1987, twenty, '79, zero, 78-degrees, eighty-four, IX, '60s, .025, fifteen, 271, 124, dozen, quintillion, DM2,000.

DT – determinant: a, an, every, no, the, another, any, some, all, both, del, each, either, half, la, many, much, nary, neither, such, that, them, these, this, those.

EX – existential there: unstressed there, which causes the inversion of the verb in the appropriate form and logical entity. For example: «There was a party in progress».

IN – prepositions or submission unions: among, around, astride, atop, behind, below, by, despite, for, if beside, if like, inside, into, near, next, on, out, pro, throughout, towards, until, upon, whether, within.

LS – list element, marker, numbers and letters that are used as identifiers of elements in the list: A, A., B, B., C, C., D, E, F, First, G, H, I, J, K, One, SP-44001, SP-44002, SP-44005, SP-44007, Second, Third, Three, Two, *, a, b, c, d, first, five, four, one, six, three, two.

MD – modal auxiliary verbs. All verbs do not accept the ending -s in the form of a third person singular: can, could, dare, may, might, must, ought, shall, should, will, would, cannot, couldn't, need, ought, shouldn't.

PDT – prefix determinant. Determinants, as elements, preceding clauses or possessive pronoun: all, both, half, many, quite, such, sure, this. For example: «all his marbles», «quite a mess».

POS – possessive ending of nouns ending in a marker ' or 's.

PRP – personal pronoun: he, her, hers, herself, him, him, himself, hisself, I, it, itself, me, myself, one, oneself, ours, ourselves, ownself, self, she, she, thee, theirs, them, themselves, they, thou, thy, us, you.

PRP$ – possessive pronoun: her, his, its, mine, my, one's, our, ours, their, thy, your.

RP – share. Mostly monosyllabic words, also disyllabic, as adverbs: aboard, about, across, along, apart, around, aside, at, away, back, before, behind, by, crop, down, ever, fast, for, forth, from, go, high, i. e., in, into, just, later, low, more, off, on, open, out, over, per, pie, raising, start, teeth, that, through, under, unto, up, up-pp, upon, whole, with you.

SYM – symbol. Technical characters or expressions that are not words (% & ' '' * + , . < = > @ A[fj] U.S U.S.S.R * ** ***).

TO – literal to, as a preposition or infinitive marker.

UH – interjection: amen, anyways, baby, dammit, diddle, Goodbye, Goody, Gosh, heck, Hey, honey, howdy, Hubba, huh, hush, Jee-sus, Jeepers, Kee-reist, man, my, oh, Oops, please, shucks, sonuvabitch, uh, well, whammo, whodunnit, Wow, yes.

WDT – wh-determinant: that, what, whatever, which, whichever.

WP – wh-pronoun: that, what, whatever, whatsoever, which, who, whom, whosoever.

WP$ – possessive wh-pronoun: whose.

WRB – wh-adverb, including *when*, when used figuratively: how, however, whence, whenever, where, whereby, wherever, wherein, whereof, why.

-LRB- – open bracket.

-RRB- – closed bracket [29–31].

On the removal of words belonging to the list of stop words – this question has already been investigated. The list of such words for English texts is justified and given in [32].

We illustrate the results of the definition of keywords at each step of the method proposed in a small text, consists of two sentences: «Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree».

The found phrases and parts of speech of the corresponding words of the first sentence are given in Table 1, and for the second – in the Table 2.

Table 1

Phrases and parts of speech corresponding words of the first sentence

Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review			
Governor word	Tag of part of speech of governor word (Governor POS)	Dependent word	Tag of part of speech of dependent word (Dependent POS)
graduate	NN	born	VBN
born	VBN	honolulu	NNP
honolulu	NNP	hawaii	NNP
graduate	NN	obama	NNP
graduate	NN	is	VBZ
graduate	NN	a	DT
university	NNP	columbia	NNP
graduate	NN	university	NNP
school	NNP	harvard	NNP
school	NNP	law	NNP
university	NNP	school	NNP
graduate	NN	school	NNP
president	NN	where	WRB
president	NN	he	PRP
president	NN	was	VBD
university	NNP	president	NN
review	NNP	the	DT
review	NNP	harvard	NNP
review	NNP	law	NNP
president	NN	review	NNP

ISSN 2226-3780

Table 2

Phrases and parts of speech corresponding words of the second sentence

He was a community organizer in Chicago before earning his law degree			
Governor word	Tag of part of speech of governor word (Governor POS)	Dependent word	Tag of part of speech of dependent word (Dependent POS)
organizer	NN	he	PRP
organizer	NN	was	VBD
organizer	NN	a	DT
organizer	NN	community	NN
organizer	NN	chicago	NNP
organizer	NN	earning	VBG
degree	NN	his	PRP$
degree	NN	law	NN
earning	VBG	degree	NN

The types of connections between the main and dependent words in the phrases given in the unchanged, basic form of the word are given for the first and second sentences in the Tables 3, 4.

Table 3

Links in the phrases of the first sentence

Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review					
Governor Word	Dependent Word	Dependency Type	Governor Word	Dependent Word	Dependency Type
graduate	bear	vmod	university	school	conj_and
bear	honolulu	prep_in	graduate	school	prep_of
honolulu	hawaius	appos	president	where	advmod
graduate	obama	nsubj	president	he	nsubj
graduate	be	cop	president	be	cop
graduate	a	det	university	president	rcmod
university	columbium	nn	review	the	det
graduate	university	prep_of	review	harvard	nn
school	harvard	nn	review	law	nn
school	law	nn	president	review	prep_of

Table 4

Links in the phrases of the second sentence

He was a community organizer in Chicago before earning his law degree		
Governor Word	Dependent Word	Dependency Type
organizer	he	nsubj
organizer	be	cop
organizer	a	det
organizer	community	nn
organizer	chicago	prep_in
organizer	earn	prepc_before
degree	his	poss
degree	law	nn
earn	degree	dobj

Let's divide the phrases into separate words and count the number of links for each word, that is, in how many phrases the word occurs. Sorting the words by the number of links, let's obtain the results, which are listed in the Table 5.

Table 5

Candidate keywords after dividing phrases

Word	Number of links	Word	Number of links	Word	Number of links
graduate	6	degree	3	hawaius	1
organizer	6	a	2	community	1
president	5	honolulu	2	the	1
university	4	earn	2	his	1
school	4	bear	2	columbium	1
review	4	harvard	2	where	1
be	3	he	2	chicago	1
law	3	obama	1	–	–

Conventionally, the phrase can be designated:

G-[T]->D,

where G – Governor Word; T – Dependency Type; D – Dependent Word.

At the stage of replacing pronouns with their corresponding nouns (replace pronouns):

– the phrase president- [nsubj] -> he is replaced by president- [nsubj] -> obama;

– the phrase organizer- [nsubj] -> he is replaced by organizer- [nsubj] -> obama;

– the phrase degree- [poss] -> his is replaced by degree- [poss] -> obama.

Candidate keywords, after replacing pronouns with their corresponding nouns, are listed in Table 6.

Table 6

Candidate keywords after pronoun substitutions

Word	Number of links	Word	Number of links	Word	Number of links
graduate	6	be	3	harvard	2
organizer	6	law	3	hawaius	1
president	5	degree	3	community	1
university	4	a	2	the	1
obama	4	honolulu	2	columbium	1
school	4	earn	2	where	1
review	4	bear	2	chicago	1

After the replacement of pronouns, the number of candidates for keywords decreased from 23 to 21. Before the replacement of pronouns, the word obama is not enough 1 link, and after – 4 links. Conversely, the words he with 2 bonds and his with a friend after the replacement of pronouns have zero connections, because the phrase with them has been replaced with equivalents with nouns.

Deleting phrases with types of connections that do not carry a significant semantic load (deleting noise relation-

ship). For this text, phrases are deleted: graduate- [det] -> a, review- [det] -> the, organizer- [det] -> a.

As a result, the number of candidates for keywords will decrease to 19, which is reflected in Table 7.

Table 7

Candidate keywords after deleting noise links

Word	Number of links	Word	Number of links
graduate	5	honolulu	2
organizer	5	earn	2
president	5	bear	2
university	4	harvard	2
obama	4	hawaius	1
school	4	community	1
be	3	columbium	1
law	3	where	1
degree	3	chicago	1
review	3	–	–

Withdrawal of words related to noise parts of speech (deleting noise POS keywords). At this point, the word where is deleted with the WRB part speech tag. Candidate keywords will have the form given in Table 8.

Table 8

Candidate keywords after deleting noise parts of speech

Word	Number of links	Word	Number of links	Word	Number of links
graduate	5	be	3	bear	2
organizer	5	law	3	harvard	2
president	5	degree	3	hawaius	1
university	4	review	3	community	1
obama	4	honolulu	2	columbium	1
school	4	earn	2	chicago	1

At the stage of deleting stop words – the stop word is removed and Table 9 contains 17 candidate keywords.

Table 9

Candidate keywords after deleting stop words

Word	Number of links	Word	Number of links	Word	Number of links
graduate	5	law	3	harvard	2
organizer	5	degree	3	hawaius	1
president	5	review	3	community	1
university	4	honolulu	2	columbium	1
obama	4	earn	2	chicago	1
school	4	bear	2	–	–

As a result, after all the proposed steps of the method, it is possible to reduce the number of candidates for keywords from 23 to 17, and also to remove noise words.

Let's now consider a relatively large text in order to determine the quantitative characteristics of the relevance of the results obtained in comparison with analogues. For this, the text «A Workingman's Poet» was chosen, which consists of 3299 words, and the keywords specified by the author are known: american, literature, literature, chicago, poetry, publishing, twentieth century, united states. According to the results of the experiment there are the first ten candidates for keywords found by the developed method: sandburg, poem, write, poet, poetry, book, life, lincoln, learn, speak. The search for keywords in the same text was implemented using similar programs.

The results of finding the keywords developed by the method and analogues are given in Table 10.

Table 10

The results of finding keywords by developed method and analogues

Etalon keywords		Advego		Rise-top		Seotool		Our development	
1	American	–	sandburg	–	sandburg	–	his	–	sandburg
2	Literature	–	that	–	his	–	sandburg	–	poem
3	Books	–	for	–	lincoln	–	lincoln	–	write
4	Chicago	–	poem	5	poetry	–	poems	–	poet
5	Poetry	–	lincoln	–	poems	5	poetry	5	poetry
6	Publishing	5	poetry	–	who	–	who	3	book
7	Twentieth	–	work	1	american	1	american	–	life
8	Century	–	write	–	where	–	where	–	lincoln
9	United	1	american	–	had	–	years	–	learn
10	States	–	where	–	years	–	had	–	speak

The results of the completeness and accuracy of the obtained keywords are given in Tables 11, 12 and in Fig. 6, 7.

Table 11

Keyword completeness results

Name	Advego	Rise-top	Seotool	Our development
Completeness (Jaccard)	0.111111111	0.111111111	0.111111111	0.111111111
Completeness (Absolute)	0.2	0.2	0.2	0.2

Table 12

Keyword accuracy results

Name	Advego	Rise-top	Seotool	Our development
Euclidean distance	0.577061522	0.59749477	0.589067059	0.577061522
Manhattan distance	0.49	0.51	0.49	0.47

The completeness of finding the keywords should be as large as possible, and the distance between the positions of the keywords given by the author and certain programmatically possible is less.

As can be seen from the histograms in Fig. 6, 7 and Tables 11, 12, own development for this text has the same

ISSN 2226-3780

completeness as analogs – 11% and 20%, however, the best quantitative characteristics in terms of accuracy are 57.71% and 47% than analogs of rise-top (59.75%; 51%); and seotool (58.91%; 49%). Also, its own development has the same accuracy for the Euclidean distance, as well as the analogue advego, but, in contrast, the best characteristics for the Manhattan distance.

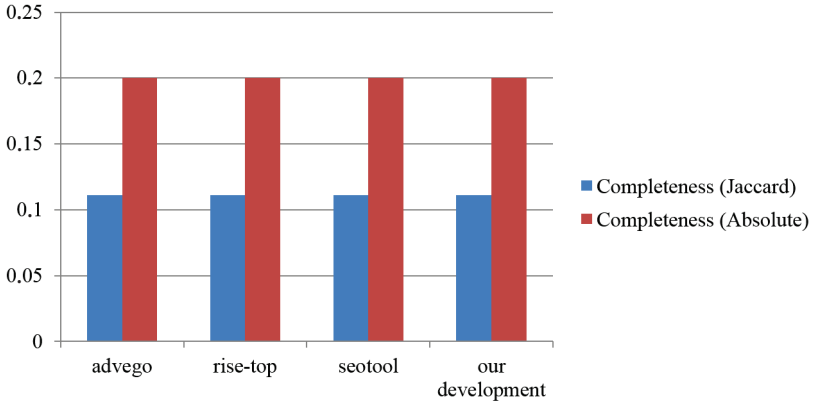

Fig. 6. Histograms of completeness by Jacquard and absolute

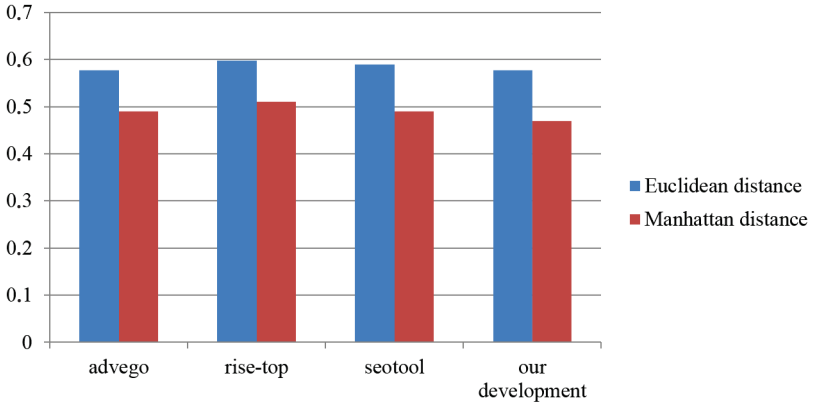

Fig. 7. Histograms of accuracy by Euclidean and Manhattan distances

7. SWOT analysis of research results

Strengths. Compared to analogs, the development is presented, according to the results of the experiment conducted with the text of 3299 words, has the same completeness as the analogs, however, the best quantitative characteristics in terms of accuracy than analogs of rise-top and seotool. Also presented the development has the same accuracy for the Euclidean distance, as well as analogue advego, but, in contrast, the best characteristics for the Manhattan distance. Another advantage compared with analogues is that the presented technology allows to completely eliminate noise words.

Weaknesses. The weaknesses of the method include the speed of its practical implementation by means of DKPro Core, in particular, it is relatively long for the online mode to create a multilevel text markup. But this, in turn, can be corrected through the use of more powerful hardware or cloud computing platforms, allowing you to have a virtual cluster of computers. This is difficult to achieve, because applications for defining keywords and reducing verbal noise are written in Java and can be easily deployed on such platforms.

Opportunities. The opportunity of further research on the definition of keywords is conducting larger-scale experiments for texts of various categories in order to determine additional ways to increase the relevance of the method. It is also advisable to use new linguistic packages that support more languages, including Ukrainian.

Threats. The process of defining keywords by the proposed method is independent of the processes of defining keywords by other methods, therefore there is no threat of a negative impact on the object of study of external factors.

The implementation of the proposed methodology does not require additional costs for the company.

An analogue of the developed method can be SEO optimization sites with the ability to determine keywords.

8. Conclusions

1. A method is proposed for filtering verbal noise in which it is provided by such formalized operations:
   – replacing pronouns with their corresponding nouns;
   – removal of noise connections;
   – removal of noise words;
   – withdrawal of stop words.

The described operations can be used as additional modules that improve the results of finding keywords for the method of determining keywords of English text based on the tools of the DKPro Core packages and also for other algorithms for finding keywords.

2. The calculation of the numerical indicators of the connections between words and the analysis of the results obtained at each stage of the method proposed is illustrated by the example of a text from two sentences. According to the results considered in the example, it is possible to reduce the number of candidates for keywords from 23 to 17, and also to completely eliminate noise words.

3. According to the results of the experiment, a development for text with 3299 words is presented, which has the same completeness as the analogs – 11% and 20%, however, the best quantitative characteristics in terms of accuracy are 57.71% and 47%, than the analogues rise-top (59.75%; 51%) and seotool (58.91%; 49%). The presented development also has the same accuracy for the Euclidean distance, as well as the analogue advego, but, in contrast, the best characteristics for the Manhattan distance.

References

1. Ershov Yu. S. Vydelenie klyuchevykh slov v russkoyazychnykh tekstakh // Molodezhnyy nauchno-tekhnicheskiy vestnik. 2014. Issue FS77-51038. P. 70–79.
2. Grashhenko L. A. O model'nom stop-slovare // Izvestiya Akademii nauk Respubliki Tadzhikistan. Otdelenie fiziko-matematicheskikh, khimicheskikh, geologicheskikh i tekhnicheskikh nauk. 2013. Issue 1 (150). P. 40–46.
3. Modeli i metody avtomaticheskoy klassifikatsii tekstovykh dokumentov / Andreev A. M. et. al. // Vestn. MGTU. Seriia Priborostroenie. 2003. Issue 3. P. 64–94.

4. Abramov E. G. Podbor klyuchevykh slov dlya nauchnoy stat'i // Nauchnaya periodika: problemy i resheniya. 2011. Issue 1 (2). P. 35–40.
5. Darkulova K. N., Ergeshova G. Neobkhodimost' vydeleniya klyuchevykh slov dlya svertyvaniya teksta: Proceedings // Lingvisticheskiy analiz nauchnogo teksta. Yuzhno-Kazakhstanskiy gosudarstvennyy universitet im. Mukhtara Auezova Shymkent, 2014. P. 30–35.
6. Halkidi M., Batistakis Y., Vazirgiannis M. On clustering validation techniques // Journal of intelligent information systems. 2001. Vol. 17, Issue 2-3. P. 107–145. doi: http://doi.org/10.1023/a:1012801612483
7. Barahnin V. B., Tkachev D. A. Clustering of text documents based on composite key terms // Vestnik NSU. Series: Information Technology. 2010. Vol. 8, Issue 2. P. 5–14.
8. Grashhenko L. A. O model'nom stop-slovare // Izvestiya Akademii nauk Respubliki Tadzhikistan. Otdelenie fiziko-matematicheskikh, khimicheskikh, geologicheskikh i tekhnicheskikh nauk. 2013. Issue 1 (150). P. 40–46.
9. Guo A., Tao Y. Research and Improvement of Feature Words Weight Based on TFIDF Algorithm // 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference. Chongqing, 2016. doi: http://doi.org/10.1109/itnec.2016.7560393
10. Sifting Micro-blogging Stream for Events of User Interest / Grineva M. et. al. // Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. Boston, 2009. P. 327–333. doi: http://doi.org/10.1145/1571941.1572157
11. TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams / Reed J. et. al. // 2006 5th International Conference on Machine Learning and Applications. Orlando, 2006. P. 258–263. doi: http://doi.org/10.1109/icmla.2006.50
12. Mihalcea R., Csomai A. Wikify!: linking documents to encyclopedic knowledge // Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. Lisbon, 2007. P. 233–242. doi: http://doi.org/10.1145/1321440.1321475
13. Astrakhantsev N. Automatic term acquisition from domain-specific text collection by using Wikipedia // Proceedings of the Institute for System Programming of RAS. 2014. Vol. 26, Issue 4. P. 7–20. doi: http://doi.org/10.15514/ispras-2014-26(4)-1
14. Özgür A., Hur J., He Y. The Interaction Network Ontology-supported modeling and mining of complex interactions represented with multiple keywords in biomedical literature // BioData Mining. 2016. Vol. 9, Issue 1. doi: http://doi.org/10.1186/s13040-016-0118-0
15. Wong W., Liu W., Bennamoun M. Ontology learning from text // ACM Computing Surveys. 2012. Vol. 44, Issue 4. P. 1–36. doi: http://doi.org/10.1145/2333112.2333115
16. Korobkin D. M., Fomenkov S. A., Kolesnikov S. G. Method of ontology-based extraction of physical effect description // Vestnik Komp'iuternykh i Informatsionnykh Tekhnologii. 2015. P. 28–35. doi: http://doi.org/10.14489/vkit.2015.02.pp.028-035
17. Besplatnyy onlayn-generator klyuchevykh slov s teksta. URL: http://seotool.by/analiz/seo/keywordstext.php
18. Generator klyuchevykh slov s teksta. URL: http://www.rise-top.com
19. Advego. URL: http://wiki.advego.ru/index.php/Адвего
20. Natural Language Processing: Integration of Automatic and Manual Analysis. 2014. URL: http://tuprints.ulb.tu-darmstadt.de/4151/1/rec-thesis-final.pdf
21. Method of determining of keywords in English texts based on the DKPro Core / Bisikalo O. V. et. al. // Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016. 2016. doi: http://doi.org/10.1117/12.2249225
22. Determiner. URL: http://universaldependencies.org/u/dep/det.html
23. Expletive and Reflexives. URL: http://universaldependencies.org/u/dep/expl.html
24. Welo E. Null Anaphora // Encyclopedia of Ancient Greek Language and Linguistics. 2013. doi: http://doi.org/10.1163/2214-448x_eagll_com_00000254
25. Manning C., de Marneffe M. Stanford typed dependencies manual. 2016. URL: https://nlp.stanford.edu/software/dependencies_manual.pdf
26. Fixed multiword. URL: http://universaldependencies.org/u/dep/fixed.html
27. Punctuation. URL: http://universaldependencies.org/u/dep/punct.html
28. Root. URL: http://universaldependencies.org/u/dep/root.html
29. Taylor A., Marcus M., Santorini B. The Penn Treebank: An Overview // Text, Speech and Language Technology. 2003. P. 5–22. doi: http://doi.org/10.1007/978-94-010-0201-1_1
30. Penn Treebank II Constituent Tags: Word level. URL: http://www.surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html#Word
31. Alphabetical list of part-of-speech tags used in the Penn Treebank Project. URL: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
32. Bougé K. Lists of stop words. URL: https://sites.google.com/site/kevinbouge/stopwords-lists

Bisikalo Oleg, *Doctor of Technical Sciences, Professor, Department of Automation and Computer-Integrated Technologies, Vinnitsa National Technical University, Ukraine, e-mail: obisikalo@gmail.com, ORCID: http://orcid.org/0000-0002-7607-1943*

Yahimovich Alexander, *Postgraduate Student, Department of Automation and Computer-Integrated Technologies, Vinnitsa National Technical University, Ukraine, e-mail: yahimovich.olexandr@gmail.com, ORCID: http://orcid.org/0000-0001-6960-5823*

Yahimovich Yaroslav, *Postgraduate Student, Department of Electronics and Nanosystems, Vinnitsa National Technical University, Ukraine, e-mail: yaroslavyahimovich@gmail.com, ORCID: http://orcid.org/0000-0003-2101-2791*