TF-IDF metrics, coupling strength of words and formation of units for knowledge representation in open tests

Mikhaylov D., Kozlov A., Emelyanov G.

Yaroslav-the-Wise Novgorod State University

All-Russian Conference with International Participation
«Mathematical Methods for Pattern Recognition» (MMPR-18),

October 9–13, 2017

Taganrog, Rostov region, Russian Federation

Knowledge unit estimated by means of open form test assignment

Is defined by a *set of natural-language phrases equivalent-by-sense* (i. e. *semantically equivalent*) relatively to the subject area considered.

Optimal sense transfer

Is provided by *those phrases* from initial set of equivalent-by-sense which are of *minimal character length* under a *maximum of words most frequently used* in all initial phrases.

Main problems

- to extract knowledge units from the texts of topical corpus;
- to select texts for the corpus by analyzing the relevance to initial phrase;
- completeness of reflection of revealed actual knowledge in initial phrases.

Forming units of expert knowledge for open tests

Research subject

Methods and algorithms for formation of knowledge on the basis of text corpus.

Expert tasks to be automated

1. Search for semantically equivalent forms for description of reality fragment in the given natural language.
2. Comparison of knowledge of given expert with the closest knowledge fragments of another experts.

Requirements for the solution

1. Revelation of concepts and relations between them in a given text.
2. Extraction from texts of corpus the usage contexts of general vocabulary by means of which synonymic paraphrases can be formed.

The image of initial phrase in the analyzed text

- In analyzed text a fragment, which corresponds to image component, can be identified with some semantic relation of words in initial phrase.
- The coupling strength of words of each such fragment is always greater than between any word from given fragment and a word not related to it.
- For terms prevailing in corpus, a combinations with a general vocabulary can be related to the extracted image component only at presence of fragments with a greater coupling strength of words.
- Generally not be required the presence of strictly predetermined part of components of image of initial phrase in text.
- The links of words of different phrases from the set of initial mutually equivalent or complementary in sense and related to the same image are allowable.

Main problems

- consideration of word combinations only within bigrams and syntactic dependences for given natural language;
- precision of revelation of knowledge fragment in the form of concepts and relations between them when initial phrase is single.

Closest ideas

- syntactic n-grams [Grigori Sidorov, 2013];
- chunking of sentences in Russian based on conditional random fields [Kudinov M. S., 2013].

Basic assumptions

- routs in dependency trees or constituent trees as a basis for n-gram revelation should be measured not from tree root, but from the word combinations with a greatest values of coupling strength;
- chunks can contain prepositions and conjunctions.

Estimation chosen for coupling strength of words concerning given document

Estimation for coupling strength of words applied in Distributive-Statistical Method of Thesaurus Construction [Moskovich W., 1971]:

$$K_{AB} = \frac{k}{a + b - k}, \tag{1}$$

where a, b and k are the numbers of document phrases containing the words A, B and A simultaneously with B, respectively.

According to classic definition, TF-IDF is the product of two statistics:
term frequence (TF) and inverse document frequency (IDF).

Term frequency estimates the significance of word t_i within the document d and can be defined as

$$\mathrm{tf}\,(t_i, d) = \frac{n_i}{\sum_k n_k}, \tag{2}$$

where n_i is the number of times that t_i occurs in document d,
and denominator contains the total number of words for d.

The value of IDF is unique for each unique word in corpus D and can be determined as follows:

$$\mathrm{idf}\,(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right), \tag{3}$$

where numerator represents the total number of documents in corpus,
and $|D_i \subset D|$ is a number of documents where the word t_i appears.

Clustreing the vocabulary of initial phrase by TF-IDF metrics: basic assumptions

1. The words, which are the most unique in document and have the largest values of TF*IDF, must be related to terms of document's topical area.
2. The fact that the term has synonyms at the same document means the decrease of TF metrics for this word relatively to given document.
3. For words of general vocabulary and for those terms which are prevail in corpus the value of IDF tends to zero.
4. Synonyms, unique for some documents of corpus, will have a higher values of IDF.

   For example: general-vocabulary words which are define the conversive replacements, like *«приводить ⇔ являться следствием»* (in Russian).

Let

D be an initial text set considered as a topical corpus.

X be an ordered descending sequence of $\mathrm{tf}(t_i, d) \cdot \mathrm{idf}(t_i, D)$ values for all words t_i of initial phrase relatively to document $d \in D$.

$H_1, \ldots H_r$ be the sequence of clusters as a result of splitting the initial X by means of algorithm close to FOREL class taxonomy algorithms.

As the mass center of cluster H_i the arithmetic mean of all $x_j \in H_i$ is taken.

For revelation of links *the most significant* words are related to the clusters:

$H_1(X)$ — the *terms* from initial phrase which are the *most unique* for d;

$H_{r/2}(X)$ — *general vocabulary* as a basis of *synonymic paraphrases*, and those *terms* which have *synonyms*.

Definition 1

Let's name further a pair of words *as pairwise related* by TF-IDF if the value of TF-IDF at least for one of them is related to either $H_1(X)$ or $H_{r/2}(X)$.

Let $d \in D$ be some document and $L(d)$ is a *sequence of bigrams* which are the *pairs of* initial phrase's *words* (A, B) *related* according to chosen method for links revelation either *syntactically or by TF-IDF*. The bigrams from $L(d)$ are *ordered descending* the *coupling strength*, $\{(A_1, B_1), (A_2, B_2)\} \subset L(d)$.

Definition 2

A bigrams (A_1, B_1) and (A_2, B_2) be a part of the same n-gram $T \subseteq L(d)$ if

$$\big((A_1 = A_2) \vee (B_1 = B_2) \vee (A_1 = B_2) \vee (B_1 = A_2)\big) = \text{true}.$$

The *significance* of n-gram T for rank estimation of d concerning the corpus D

$$\mathrm{N}(T, d) = \frac{\sqrt{\sum_{i=1}^{\mathrm{len}(T)} \left[S_i(d)\right]^2}}{\sigma\big(S_i(d)\big) + 1}, \tag{4}$$

where $S_i(d)$ is the coupling strength of words of i-th bigram relatively to d;
$\sigma\big(S_i(d)\big)$ is the root-mean-square deviation of mentioned value;
$\mathrm{len}(T)$ is the length of n-gram T (in bigrams).

Let's denote further the set of n-grams $\{T\colon\ T \subseteq L(d)\}$ as $\mathbb{T}(d)$.

Ranking function and forming the annotation

The *rank for document* d relatively to topical corpus D:

$$W(d) = \mathrm{N}_{\max}(d) \cdot \log_{10}\Big(\max_{T \in \mathbb{T}(d)} \mathrm{len}\,(T)\Big) \cdot \log_{10}\Big(\big|\mathbb{T}\,(d)\big|\Big), \tag{5}$$

where $\mathrm{N}_{\max}(d) = \max_{T \in \mathbb{T}(d)} \mathrm{N}(T, d)$.

Let $D' \subset D$ be the cluster of greatest values of estimation (5).

Similarly, according to the values of (4) the set $\mathbb{T}\,(d)$ for $\forall\, d \in D'$ is splitted.
Let $\mathbb{T}'\,(d)$ be the cluster of greatest values of estimation (4) for given d.

Herewith for phrase s from $d \in D'$ two variants of estimation are possible:

$$\mathrm{N}\,(s) = \Big|\big\{w \in b\colon\ \exists\, T \in \mathbb{T}'\,(d)\,,\, b \in T\big\}\Big| \tag{6}$$

and, correspondingly,

$$\mathrm{N}\,(s) = \Big|\big\{b\colon\ \exists\, T \in \mathbb{T}'\,(d)\,,\, b \in T\big\}\Big| \tag{7}$$

as a basis of clustering the whole set $\big\{s\colon\ s \in d \,\big|\, d \in D'\big\}$.

Annotation phrases

Form the *first cluster* from obtained according to the values of $\mathrm{N}\,(s)$.

Let

Ts be a *group of initial phrases* mutually equivalent or complementary in sense and determined some *knowledge unit*.

The *relevance estimation* of text corpus D to knowledge unit and *situation of natural language usage* associated with Ts on the basis of revealed n-grams *can be determined as*

$$\mathbb{W}(D) = \frac{1}{\left|D'\right|} \sum_{d \in D'} \left[\frac{\left|\left\{w \in b\colon\ \exists T \in \mathbb{T}'(d),\, b \in T\right\}\right|}{\left|\left\{w\colon\ \exists Ts_i \in \mathrm{Ts},\, w \in Ts_i\right\}\right|} \sum_{T \in \mathbb{T}'(d)} \mathrm{N}(T, d) \right], \quad (8)$$

where $\mathrm{N}(T, d)$ is the significance estimation for n-gram T according to (4);

$\mathbb{T}'(d)$ is the cluster of greatest values of estimation (4) for given d;

$D' \subset D$ is the cluster of greatest values of estimation (5).

The main criteria

1. The initial phrases should be formulated independently from each other by different experts.
2. The initial text sets should allow for comparison the initial phrase's images extracted in analyzed texts:
   - for separate initial phrases and their sets with the respect of possible links among phrases;
   - by estimation the coupling strength of words within a pair and for sequences of such pairs within n-grams;
   - by analysis of syntactic dependences and with application of TF-IDF metrics at revelation of links of words.
3. The fullest and evident illustration of extraction from texts the usage contexts both for terms, and general vocabulary by means of which synonymic paraphrases of initial phrase can be formed.

Initial text set for experiment: the first variant

- Vestnik of the Plekhanov Russian University of Economics (VPRUE, 1 paper);
- The annual «Filosofija nauki» (Philosophy of Science) (PhSc, 1 paper);
- materials of the 4^{th} All-Russian conference of students, post-graduates and young scientists «Artificial Intelligence: Philosophy, Methodology, Innovations» (AI PhMI, 2010, 3 papers in Part 1 and 1 paper in Part 2);
- matherials of the 7^{th} Conference AI PhMI (2013, 2 sectional reports and 1 plenary report);
- matherials of the 8^{th} Conference AI PhMI (2014, 1 plenary report);
- matherials of the 9^{th} Conference AI PhMI (2015, 1 paper);
- Taurida journal of computer science theory and mathematics (TJCSTM, 1 paper).

Remark

The number of words in documents of initial set varied here from 618 to 3765, and the number of phrases per document varied between 38 and 276.

Initial Russian phrases for expert knowledge units formation

№	Initial phrase
1	*Определение модели представления знаний накладывает ограничения на выбор соответствующего механизма логического вывода.*
2	*Под знанием понимается система суждений с принципиальной и единой организацией, основанная на объективной закономерности.*
3	*С точки зрения искусственного интеллекта знание определяется как формализованная информация, на которую ссылаются или используют в процессе логического вывода.*
4	*Факты обычно указывают на хорошо известные обстоятельства в данной предметной области.*
5	*Эвристика основывается на собственном опыте специалиста в данной предметной области, накопленном в результате многолетней практики.*
6	*Метазнания могут касаться свойств, структуры, способов получения и использования знаний при решении практических задач искусственного интеллекта.*
7	*Однородность представления знаний приводит к упрощению механизма управления логическим выводом и упрощению управления знаниями.*
8	*Отличительными чертами логических моделей являются единственность теоретического обоснования и возможность реализации системы формально точных определений и выводов.*
9	*Язык представления знаний на основе фреймовой модели наиболее эффективен для структурного описания сложных понятий и решения задач, в которых в соответствии с ситуацией желательно применять различные способы вывода.*

Initial text set for experiment: the second variant

- Taurida journal of computer science theory and mathematics (TJCSTM, 3 papers);
- Proceedings of International conferences «Intelligent Information Processing» IIP-8 and IIP-9 (2 papers);
- Proceedings of All-Russian Conference with International Participation on Mathematical Methods for Pattern Recognition (MMPR-15, 1 paper);
- Proceedings of the Conference MMPR-13 (2 papers);
- Proceedings of the Conference MMPR-16 (14 papers);
- Proceedings of the Conference IIP-10 (2 papers);
- the text of a scientific report prepared in 2003 by Dmitry Mikhaylov.

Remark

The number of words in documents of initial set varied here from 218 to 6298, and the number of phrases per document varied between 9 and 587.

The scope of selected papers for the second variant of initial text corpus

- mathematical methods for learning by precedents (K. Vorontsov, M. Khachay, E. Djukova, N. Zagoruiko, Yu. Dyulicheva, I. Genrikhov, A. Ivakhnenko);
- methods and models of pattern recognition and forecasting (V. Mottl, O. Seredin, A. Tatarchuk, P. Turkov, M. Suvorov, A. Maysuradze);
- intelligent processing of experimental information (S. Dvoenko, N. Borovykh);
- image processing, analysis, classification and recognition (A. Zhiznyakov, K. Zhukova, I. Reyer, D. Murashov, N. Fedotov, V. Martyanov, M. Kharinov).

Some technical details

- To calculate the offered estimations the lemmatization of words was performed by the function $getNormalForms$ from the *Russian Morphology for lucene*.
- The syntactic links are extracted according to the rules employed in paper [Tsarkov S., Natural and Technical Sciences, 2012, № 6].
- Sentence boundary detection by a punctuation character marks was implemented with attraction of pre-trained model of classifier created by means of *Apache OpenNLP*.
- Training data for sentence boundary detector were the tagged sentences from Russian newspaper texts represented in *Leipzig Corpora* (2010, total 10^6 phrases).

Initial Russian phrases for expert knowledge units formation

№	Initial phrase
1	*Переобучение приводит к заниженности эмпирического риска.*
2	*Переподгонка приводит к заниженности эмпирического риска.*
3	*Переподгонка служит причиной заниженности эмпирического риска.*
4	*Заниженность эмпирического риска является результатом нежелательной переподгонки.*
5	*Переусложнение модели приводит к заниженности средней ошибки на тренировочной выборке.*
6	*Переподгонка приводит к увеличению частоты ошибок дерева принятия решений на контрольной выборке.*
7	*Переподгонка приводит к заниженности оценки частоты ошибок алгоритма на контрольной выборке.*
8	*Заниженность оценки ошибки распознавания связана с выбором правила принятия решений.*
9	*Рост числа базовых классификаторов ведёт к практически неограниченному увеличению обобщающей способности композиции алгоритмов.*

software implementation and experimental results

Initial Russian phrases in groups

№	Group of initial phrases	
1	*Нежелательная переподгонка является причиной заниженности средней величины ошибки алгоритма на обучающей выборке.*	
	Переобучение приводит к заниженности эмпирического риска.	(2.1)
2	*Определение модели представления знаний накладывает ограничения на выбор соответствующего механизма логического вывода.*	(1.1)
	Однородность представления знаний приводит к упрощению механизма управления логическим выводом и упрощению управления знаниями.	(1.7)
3	*Эвристика основывается на собственном опыте специалиста в данной предметной области, накопленном в результате многолетней практики.*	(1.5)
	Метазнания могут касаться свойств, структуры, способов получения и использования знаний при решении практических задач искусственного интеллекта.	(1.6)

Remark

The *first digit* in a number to the right from phrase *denotes the topical area* (1 — Philosophy and Methodology of Knowledge Engineering, 2 — Mathematical Methods for Learning by Precedents), *the second denotes* the *number according to table* for initial phrase (see *Slides 14* and *17*).

go to the examples

Selection the relevant phrases for groups of initial ones on the basis of n-grams[1]

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
using TF-IDF and estimation (6)								*syntactic rules and estimation* (6)						
2	1	0	0	1	0	0	1	16	2	0	3	0	0	3
3	1	1	1	1	1	3	2	3	1	1	2	1	1	3
using TF-IDF and estimation (7)								*syntactic rules and estimation* (7)						
2	3	1	0	1	1	0	1	4	0	0	2	0	0	1
3	2	1	2	2	1	5	5	2	0	0	1	0	0	1

Here:

N is the total number of selected phrases;

N_1 is the number of phrases representing the linguistic expressional means for initial ones;

N_2 is the number of phrases representing synonyms;

N_3 is the number of phrases representing conceptual relations at the topical area;

N_1^1 is the number of linguistic expressional means which were represented in resulted phrases;

N_2^1 is the number of synonyms found in resulted phrases;

N_3^1 is the number of found concept relations from mentioned in initial phrases.

[1] Here and below on *Slides 20–29* combinations with prepositions and conjunctions are respected too

Selection the relevant phrases for groups of initial ones on the basis of n-grams

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
Selection the relevant for separate phrases from groups №2 and №3														
using TF-IDF and estimation (6)								*syntactic rules and estimation (6)*						
1.1	2	0	0	2	0	0	1	5	0	0	3	0	0	3
1.5	3	0	1	2	0	1	2	3	0	2	1	0	2	1
1.6	3	0	0	1	0	0	1	3	1	0	2	1	0	2
1.7	3	0	0	1	0	0	2	3	0	0	2	0	0	2
using TF-IDF and estimation (7)								*syntactic rules and estimation (7)*						
1.1	1	0	0	1	0	0	1	4	0	0	2	0	0	2
1.5	1	0	1	1	0	1	1	1	0	1	1	0	1	1
1.6	10	1	0	4	1	0	3	1	1	0	1	1	0	1
1.7	3	0	0	1	0	0	2	1	0	0	0	0	0	0

N is the total number of selected phrases; N_2 is the number of phrases representing synonyms;

N_1 is the same for linguistic expressional tools; N_3 is the same for conceptual relations;

N_1^1 is the number of linguistic expressional means which were represented in resulted phrases;

N_2^1 is the number of synonyms found in resulted phrases;

N_3^1 is the number of found concept relations from mentioned in initial phrases.

Selection of relevant by the number of «most strong» links for phrases from groups №2 and №3

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
	with attraction of TF-IDF							*on the basis of syntactic rules*						
1.1	1	0	0	1	0	0	1	2	0	0	1	0	0	1
1.5	2	2	2	0	2	2	0	4	0	0	2	0	0	5
1.6	6	1	1	1	1	1	1	1	1	1	0	1	1	0
1.7	6	0	0	2	0	0	3	6	0	0	2	0	0	3

Here:

N is the total number of selected phrases;

N_1 is the number of phrases representing the linguistic expressional means for initial ones;

N_2 is the number of phrases representing synonyms;

N_3 is the number of phrases representing conceptual relations at the topical area;

N_1^1 is the number of linguistic expressional means which were represented in resulted phrases;

N_2^1 is the number of synonyms found in resulted phrases;

N_3^1 is the number of found concept relations from mentioned in initial phrases.

Selection the relevant phrases for groups of initial ones

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
using TF-IDF and estimation (6)								*syntactic rules and estimation (6)*						
1	3	1	1	1	1	1	2	1	0	0	1	0	0	2
using TF-IDF and estimation (7)								*syntactic rules and estimation (7)*						
1	1	0	1	1	0	1	2	1	0	0	1	0	0	2
Selection the relevant for separate phrases from the group №1														
using TF-IDF and estimation (6)								*syntactic rules and estimation (6)*						
2.1	1	1	0	0	1	0	0	1	1	0	0	1	0	0
using TF-IDF and estimation (7)								*syntactic rules and estimation (7)*						
2.1	1	1	0	0	1	0	0	1	1	0	0	1	0	0
by the number of «most strong» links from revealed														
with attraction of TF-IDF								*on the basis of syntactic rules*						
2.1	2	0	0	1	0	0	1	1	1	0	0	1	0	0

N is the total number of selected phrases; N_2 is the number of phrases representing synonyms;

N_1 is the same for linguistic expressional tools; N_3 is the same for conceptual relations;

N_1^1 is the number of linguistic expressional means which were represented in resulted phrases;

N_2^1 is the number of synonyms found in resulted phrases;

N_3^1 is the number of found concept relations from mentioned in initial phrases.

Example of extraction the constituents of image for the group of initial phrases №3

On the basis of n-grams using TF-IDF and estimation (6):

Selected phrase	Expressed relations	Words of the most significant n-grams
Эвристика может пониматься как: — *научно-прикладная дисциплина, изучающая творческую деятельность;* — *приёмы решения проблемных (творческих, нестандартных, креативных) задач в условиях неопределённости, которые обычно противопоставляются формальным методам решения, опирающимся, например, на точные математические алгоритмы;* — *метод обучения;* — *один из способов создания компьютерных программ — эвристическое программирование.*	Relation of the concept group *heuristics – knowledge* with the concept for *methods of solving tasks*, periphrase *в результате* $\Longleftrightarrow$ *как результат*, synonyms *способ* $\Longleftrightarrow$ *приём*, *опираться* $\Longleftrightarrow$ *основываться*, *практический* $\Longleftrightarrow$ *прикладной*	*эвристика, в, задача, на, способ, решение, мочь*

Selection the relevant ones for the phrase (1.6) by the number of «most strong» links from revealed by TF-IDF:

Selected phrase	Expressed relations	The «most strong» links
Стремление преодолеть узость алгоритмического подхода привело к возникновению эвристического направления в разработке проблем искусственного интеллекта, где эвристика понимается как термин, противостоящий понятию алгоритма, который представляют собой «набор инструкций или четко сформулированных операций, составляющих определенную процедуру».	Relation of the concept of *artificial intelligence* mentioned in initial phrase with the concept of *heuristics*	*искусственный – интеллект*

On the basis of n-grams using TF-IDF and estimation (6):

Selected phrase	Expressed relations
При этом модель знания понималась как формализованная в соответствии с определенными структурными планами информация, сохраняемая в памяти, и которая может быть им использована в ходе решения задач на основании заранее запрограммированных схем и алгоритмов.	**Relation of the concept of** *knowledge* mentioned in initial phrase **with the concept of** *knowledge model*, **periphrase** *определяется как* $\Longleftrightarrow$ *понимается как*

By the number of the «most strong» links from revealed with attraction of TF-IDF:

Selected phrase	Expressed relations
Согласно Дж. фон Нейману, информация имеет двоякую природу: она может трактоваться как программа или алгоритм по работе с данными и как информация об объектах, т. е. те данные, с которыми программа работает.	**Conceptual relationships** for the concept of *information*
Информация представляет собой закодированное в эксплицитной форме знание, по которому человек способен творчески его воссоздать.	
При этом модель знания понималась как формализованная в соответствии с определенными структурными планами информация, сохраняемая в памяти, и которая может быть им использована в ходе решения задач на основании заранее запрограммированных схем и алгоритмов.	**periphrase** *определяется как* $\Longleftrightarrow$ *понимается как*

Estimating the relevance of text corpus to initial knowledge units

No. of initial phrase or their group[2]	taking into account of prepositions/conjunctions/interjections	excluding prepositions/conjunctions/interjections
	Philosophy and Methodology of Knowledge Engineering	
	for separate initial phrases	
1	0,1443376	0,0861601
2	0,1423988	0,0643456
3	0,3995547	0,5083567
4	0,1513025	0,1650242
5	0,6166341	0,3633269
6	0,1591293	0,1621076
7	0,2127629	0,0326510
8	0,2393714	0,1471097
9	0,5758868	0,3178877
	for groups of initial phrases	
3	0,3120782	0,4472640
	Mathematical Methods for Learning by Precedents	
	for separate initial phrases	
1	0,6517818	0,2905786
2	0,5433360	0,2905786
3	0,2066957	0,2066957
4	0,1962131	0,1962131
5	0,3398426	0,0599116
6	0,2031058	0,2676248
7	0,2507539	0,3768646
8	0,2621604	0,2166871
9	0,1825379	0,1977494

[2] Revelation of links of words here is carried out without application of syntactic rules

Comparison of n-grams and links most significant for phrases selection (estimating by the number of words without application of syntactic rules)

No. of initial phrase	Words which are not entered in most significant links	n-grams
	Philosophy and Methodology of Knowledge Engineering	
2	*знание*	*и, на, с*
3	*знание, с, или, использовать*	
4		*на*
5	*на, собственный, опыт, область*	
6	*и*	
7	*представление, и*	
8	*реализация, система, и, возможность*	
9	*с, понятие, структурный, соответствие, представление, в, ситуация*	*различный*
	Mathematical Methods for Learning by Precedents	
2	*заниженность*	
3	*заниженность, причина*	
4	*заниженность, являться*	
5	*к, средний*	
6	*приводить, к*	*принятие, решение*
8		*принятие*

In given illustration the comparison is made for those documents which were related to the *most relevant* for initial phrase at ranking both on the basis of n-grams, and by the «most strong» links.

Comparison of n-grams and links most significant for phrases selection (estimating by the number of words without application of syntactic rules)

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
	by maximization of number of the «most strong» links for K_{AB}									*by analysis of n-grams on the found links of words*								
Philosophy and Methodology of Knowledge Engineering																		
N	1	2	11	1	2	6	6	6	1	2	4	4	2	3	3	3	2	3
N_1	0	0	1	1	2	1	0	0	0	0	0	1	1	0	0	0	1	0
N_2	0	0	2	0	2	1	0	1	0	0	0	0	1	1	0	0	0	0
N_3	1	0	5	1	0	1	2	3	1	2	2	3	0	2	1	1	2	3
Mathematical Methods for Learning by Precedents																		
N	2	1	15	15	5	1	6	1	1	1	1	2	1	2	1	9	2	6
N_1	0	1	3	2	0	0	0	0	1	1	1	2	1	0	1	1	0	1
N_2	0	1	2	2	1	0	1	0	1	0	1	2	1	1	1	0	0	0
N_3	1	0	7	4	0	0	0	1	0	0	0	1	0	1	1	5	1	4

Here:

N is the total number of selected phrases;

N_1 is the number of phrases representing the linguistic expressional tools;

N_2 is the number of phrases representing synonyms;

N_3 is the number of phrases representing conceptual relations.

Comparison of n-grams and links most significant for phrases selection (estimating by the number of words with application of syntactic rules)

No. of initial phrase	Words which are not entered in most significant links	n-grams
	Philosophy and Methodology of Knowledge Engineering	
1	*знание, выбор*	
2	*основать, организация*	
5	*в, на, специалист, результат, практика, область*	*опыт, накопить*
8	*определение, возможность*	
	Mathematical Methods for Learning by Precedents	
2	*заниженность*	
3	*заниженность*	
4	*заниженность, являться*	
5	*средний*	
6	*приводить, к*	
9	*алгоритм, к*	

As in previous, in the given illustration the comparison is made for documents which were related to the *most relevant* for initial phrase at ranking both on the basis of n-grams, and by the «most strong» links.

Comparison of n-grams and links most significant for phrases selection (estimating by the number of words with application of syntactic rules)

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
	by maximization of number of the «most strong» links for K_{AB}									*by analysis of n-grams on the found links of words*								
Philosophy and Methodology of Knowledge Engineering																		
N	2	4	1	3	4	1	6	1	5	5	2	19	7	3	3	3	1	1
N_1	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	1	0
N_2	0	0	0	2	0	1	0	0	0	0	0	3	0	2	0	0	0	0
N_3	1	2	0	1	2	0	2	0	2	3	1	4	4	1	2	2	0	1
Mathematical Methods for Learning by Precedents																		
N	1	1	15	15	5	11	1	1	1	1	1	2	1	2	1	9	4	1
N_1	1	1	3	2	0	0	0	0	1	1	1	2	1	0	0	1	0	0
N_2	0	1	2	2	1	9	0	0	1	0	1	2	1	2	1	0	0	0
N_3	0	0	7	4	0	4	0	1	0	0	0	1	0	2	0	3	0	0

Here:

N is the total number of selected phrases;

N_1 is the number of phrases representing the linguistic expressional tools;

N_2 is the number of phrases representing synonyms;

N_3 is the number of phrases representing conceptual relations.

Alternative solution: the search of relevant phrases on a ready syntactically marked text corpus

Words and their combinations for selection of phrases from *Russian National Corpus*:

№	Words and their combinations
Philosophy and Methodology of Knowledge Engineering	
1	*модель – представление – знание, механизм – логический – вывод*
2	*система – суждение, объективный – закономерность*
3	*процесс – логический – вывод*
4	*данный – предметный – область*
5	*эвристика, данный – предметный – область*
6	*метазнания, свойство – знание, структура – знание, способ – получение – знание, способ – использование – знание, задача – искусственный – интеллект*
7	*представление – знание, управление – вывод, механизм – логический – вывод, управление – знание*
8	*теоретический – обоснование – модель, логический – модель, система – вывод, система – определение, точный – вывод*
9	*язык – представление – знание, фреймовый – модель, способ – вывод*

№	Words and their combinations
Mathematical Methods for Learning by Precedents	
1	*переобучение, эмпирический – риск*
2	*эмпирический – риск*
3	*эмпирический – риск*
4	*эмпирический – риск*
5	*ошибка – средний*
6	*частота – ошибка, контрольный – выборка*
7	*оценка – частота, контрольный – выборка*
8	*ошибка – распознавание, правило – принятие – решение*
9	*базовый – классификатор*

Selection the relevant phrases from texts of Russian National Corpus

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
	Philosophy and Methodology of Knowledge Engineering									*Mathematical Methods for Learning by Precedents*								
N	13	73	2	15	83	33	79	224	20	56	1	1	1	24	17	21	5	2
N_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N_3	2	5	0	1	5	3	3	2	2	0	0	0	0	0	0	0	1	0
N_3^1	2	6	0	2	4	3	3	2	2	0	0	0	0	0	0	0	1	0

Here:

N is the total number of selected phrases;

N_1 is the number of phrases representing the linguistic expressional means for initial ones;

N_2 is the number of phrases representing synonyms;

N_3 is the number of phrases representing conceptual relations at the topical area;

N_3^1 is the number of found concept relations from mentioned in initial phrases.

Conclusions

1. The main *result* of current work is the *formation method* for topical corpus of texts relevant at described knowledge fragments to the group of initial phrases with extraction of its image components expressed in words and their combinations.
2. In comparison with the search of such components on a syntactically marked text corpus covering all given natural language the *method* for text selection *offered in this work enables* a **17**-*times reduction (on average)* in the *output of phrases* which are *irrelevant to initial ones* in terms of either the described knowledge fragment or its linguistic expression forms.
3. The proposed *variant of contextual annotation* is primarily oriented *to search the forms of expression for conceptual relationships* in texts related to so topical area where the percentage of general vocabulary and terms are comparable.
4. The open problem is *the speed and precision of morphological analysis*. Here of interest is the Python-implementation of offered method with attraction of *NLTK* library and morphological analyzer *Pymorphy* as an alternative to realized solution based on *Russian morphology framework*.