Coupling strength of words and estimation of text relevance to unit of knowledge in open tests

Mikhaylov D., Kozlov A., Emelyanov G.

Yaroslav-the-Wise Novgorod State University

11^{th} International Conference
on Intelligent Data Processing: Theory and Applications,

October 10–14, 2016

Barcelona, Spain

Open test and rational sense transfer

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.

The main problems of formation of topical corpus for open tests

1. The selection of texts for the corpus, as a rule, is subjective and depends on expert.
2. When choosing criterion for the selection of texts it is necessary to respect both the level of difficulty of the text and its significance for the formation of test on the specified fragments of expert knowledge (for example, from the point of view of thematic rubrication).
3. In general case, the significance of text in the problem to be solved is unrelated to the image representing the initial phrase in the analyzed texts, and may guide the selection of the measure of affinity to initial phrase.
4. Itself, the initial phrase only in a few cases meet the standard for comparison.

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 image extraction for initial phrase in texts selected to corpus

- Analysis of occurrence both for separate words, and for their combinations.
- Estimation of coupling strength of words relatively to text and corpus.

Main purpose of research

To study variants of numerical estimation for coupling strength of words and how to apply them for extraction of image components for initial phrase.

Extraction of word combinations

Instruments:

- frequency of L-grams (according to $C.\ Shannon$);
- frequency and filtering by tags;
- expected value and dispersion.

Estimation methods for statistical significance of word combinations:

- Student's t-test;
- Pearson's chi-squared test (χ^2);
- likelihood ratio test.

Problems

- The syntactically marked text corpus is required here to extract bigrams associated with the word combinations.
- A syntactic marking of corpus texts cannot be fully automatized and requires considerable time.
- Existing corpora in most cases are not contain the required information about bigrams from analyzed texts.

Estimation variants for coupling strength of words

1. Significance estimation for word combinations extracted from a text containing n phrases [Biemann C., 2004]:

$$\operatorname{sig}(A, B) = x - k \log(x) + \log k!, \tag{1}$$

where $x = \frac{ab}{n}$; a, b and k are the numbers of phrases containing the words A, B and A simultaneously with B, respectively.

Disanvantages:

- for correctly application of the estimation (1) each of words from pair (A, B) must be at least in one phrase from analyzed text;
- as ideologically close to G-test for *Poisson distributions*, the estimation (1) may be inaccurate if the expected number of phrases in document is less then 5.

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

$$K_{AB} = \frac{k}{a + b - k}. \tag{2}$$

Remark

To prevent divising by zero when A and B not occur in phrases of analyzed text separately one from another the value in denominator of formula (2) has to be increased on 1.

Document ranking by relevance to the initial phrase

Let

D be an initial text set.

X be an ordered descending sequence of nonzero $\operatorname{sig}(A, B)$ or K_{AB} relatively to document $d \in D$ for pairs of words (A, B), to which a syntactical links in initial phrase are correspond.

$H_1, \ldots H_r$ 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.

The *document ranking function* here can be defined as

$$W(d) = K_\Sigma(d) \frac{K_1(d)}{K_\Sigma(d)} = K_1(d), \tag{3}$$

where $K_\Sigma(d)$ is the total «strength» of all links (A, B) that were found in the initial phrase relatively to d;

$K_1(d)$ is the total «strength» of links related to the cluster H_1 of greatest values of coupling strength.

Forming the annotation

Let

D be clustered by analogy with X, but according to the values of function (3);

$D' \subset D$ be the cluster of greatest values of (3).

Using the terminology of information retrieval, let's name further the search of phrases closest to initial in documents $d \in D'$, as the *forming of annotation*.

Variants for phrase selection to annotation

- using the number of links found in the phrase and related to cluster H_1;
- using the total «strength» of mentioned links.

Closest approaches

- Search for fuzzy duplicates of documents, where the *similarity measure* for pair of documents is calculated as the *ratio* of number of common fixed length substrings (in our case this length would be equal to two) to document size (in words) [Manber U., 1994; Heintze N., 1996].
- Unlike *Yandex* algorithms of contextual annotation [Yandex, 2008], one annotation is formed here *for several documents at once*.

Let L be a sequence of *bigrams* which are the pairs of *syntactically linked* words (A, B) of initial phrase, *ordered descending* by the value of *coupling strength* relatively to some document $d \in D$, $\{(A_1, B_1), (A_2, B_2)\} \subset L(d)$.

Definition 1

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 *total coupling strength* for words of T relatively to d can be estimated as

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

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 $\mathbb{T}(d)$.

Let D be an initial text set.

Ranking function for documents $d \in D$, which estimates the found n-grams:

$$W(d) = \frac{1}{\left|\mathbb{T}(d)\right|}\Big[\sum_{T\in\mathbb{T}(d)} \mathrm{N}\big(T,d\big)\Big]\Big[\left|\mathbb{T}(d)\right| - \max_{T\in\mathbb{T}(d)} \mathrm{len}\,(T)\Big]\frac{\min_{T\in\mathbb{T}(d)} \mathrm{N}\big(T,d\big)}{\max_{T\in\mathbb{T}(d)} \mathrm{N}\big(T,d\big)}. \quad (5)$$

The set D is clustered according to the values of function (5).

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.

For each phrase s of each document $d \in D'$ the estimation

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

is entered 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 estimation (6).

The main criteria

- The initial phrases should be formulated independently from each other by different experts.
- The initial text sets should allow for comparison the initial phrase's images extracted in analyzed texts on the basis of coupling strength and TF-IDF of words of initial phrase that are found in phrases of these texts.
- 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.
- The number of phrases in text document must be at least 5.

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 set

- 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

Example for initial phrase №8, Mathematical Methods for Learning by Precedents

Selected phrase	Expressed relations	The «most strong» links	Kind of estimation
	M. Khachay, MMPR-16		
Системы ограничений, возникающие в задачах принятия решений, оптимизации, распознавания образов и анализа часто являются несовместными, подразумевающими те или иные подходы к их коррекции, связанной с обобщением классического понятия решения	**Relation between** *generalization of classic conception of decision* **and** *a choice of decision-making rule*	*распознавание – с, принятие – решение*	K_{AB}
Современная теория комитетных решений и тесно связанных с ними комитетных методов обучения распознаванию опирается на фундаментальные результаты, полученные Вл. Д. Мазуровым	**Relation of the concept of** *recognition* **mentioned in initial phrase with the concept of** *(machine) learning*		K_{AB}
	Yu. Dyulicheva, TJCSTM 2003 №2		
Эмпирический решающий лес повысил эффективность распознавания объектов, не участвовавших ранее в обучении, по сравнению с одним решающим деревом, при использовании одного и того же критерия ветвления	**Consideration of** *decision tree* **and** *decision forest* **as a ways to represent** *decision-making rules*	*распознавание – с, принятие – решение*	$\mathrm{sig}\,(A, B)$

Were colored here:

K_{AB} — if phrases were found only on a maximum of the «most strong» links;

K_{AB}, $\mathrm{sig}\,(A, B)$ — if phrases were found both on a maximum of these links, and on their total strength.

Clusters by TF-IDF for phrases selection		Estimation	The «most strong» links
K. Vorontsov, TJCSTM 2004 №1, words presented in clusters		Yu. Dyulicheva, MMPR-13	
H_1	*алгоритм*, *обобщать*, *способность*	K_{AB}	*увеличение – обобщать*, *увеличение – способность*, *обобщать – способность*
$H_{r/2}$	*к*, *классификатор*, *увеличение*		
H_r	*вести*		
K. Vorontsov, MMPR-15, words presented in clusters		K. Vorontsov, TJCSTM 2004 №1	
H_1	*алгоритм*	$\mathrm{sig}\,(A, B)$	*обобщать – способность*
$H_{r/2}$	*рост*, *композиция*		
H_r	*неограниченный*, *базовый*, *увеличение*		

For comparison: the phrase selected by TF-IDF and not revealed by $\mathrm{sig}\,(A, B)$**:** *Наиболее общая теория алгоритмических композиций разработана в алгебраическом подходе к построению корректных алгоритмов, предложенном академиком РАН Ю. И. Журавлёвым и активно развиваемом его учениками.*

Not related to the «most strong» links here: *композиция – алгоритм*, *вести – к*

Selection the relevant phrases: comparison with the decision based on TF-IDF

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
phrases selection according to TF-IDF of words of initial phrase										*selection by the number of the «most strong» links for* $\mathrm{sig}(A,B)$								
N	1	1	1	1	3	2	4	1	40	1	1	11	11	5	20	9	10	19
N_1	1	1	1	1	0	0	0	0	7	1	1	1	2	0	1	0	0	2
N_2	0	1	1	1	3	0	0	0	6	0	1	1	1	1	1	1	0	1
N_3	0	0	0	0	1	1	1	0	8	0	0	4	4	0	0	5	1	7
selection by the number of the «most strong» links for K_{AB}										*selection by the total «strength» of the «most strong» links for* $\mathrm{sig}(A,B)$								
N	1	1	15	15	5	11	1	1	1	9	9	1	1	1	1	6	3	8
N_1	1	1	3	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0
N_2	0	1	2	2	1	9	0	0	1	0	0	0	0	0	0	0	0	0
N_3	0	0	7	4	0	4	0	1	0	0	0	0	1	0	0	1	1	2
selection by the total «strength» of the «most strong» links for K_{AB}																		
N	10	9	2	2	8	6	2	2	1									
N_1	0	0	0	0	1	0	0	0	1									
N_2	0	0	0	0	1	4	0	0	1									
N_3	1	0	1	0	1	2	0	2	0									

N is the total number of selected phrases;

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

N_2 is the same for synonyms;

N_3 is the same for conceptual relations.

Clusters for phrases selection according to TF-IDF of words of initial phrase:

A. Yankovskaya, TJCSTM 2004 №1, words presented in clusters	
H_1	различный
$H_{r/2}$	применять, модель, наиболее, ситуация, соответствие
H_r	с, решение, понятие, сложный, который, вывод, фреймовый, на, задача, в, и, основа, для, знание

Documents which are the best in criterion (3), and links of words from initial phrase:

Estimation	The «most strong» links for phrases selection
V. Rusanov, VPRUE 2012 №1	
K_{AB}	язык – на, язык – сложный, на – основа, представление – с, язык – фреймовый, представление – в, представление – для, представление – понятие, язык – основа
$\mathrm{sig}\,(A, B)$	язык – на, на – основа, язык – сложный, язык – фреймовый, основа – с
V. Lektorskiy, AI PhMI, 2014	
K_{AB}	язык – задача, представление – способ, основа – модель, модель – для, модель – применять, сложный – понятие, в – знание, описание – применять, решение – различный, на – описание
H. Krymskaya, AI PhMI, 2010	
K_{AB}	решение – задача, решение – с, задача – в, на – решение, решение – для
A. Yankovskaya, TJCSTM 2004 №1	
$\mathrm{sig}\,(A, B)$	на – основа, решение – задача

Selected phrase	Expressed relations	Estimation
V. Rusanov, VPRUE 2012 №1		
Специфика структурно-фреймовой организации состоит в том, чтобы во фрейме (а он представляет собой достаточно сложную концептуальную конструкцию, записанную средствами программной части вычислительной (информационной) системы) все понятия, относящиеся к охватываемой данным фреймом предметной области, имели внутреннюю интерпретацию, т. е. были наделены смыслом на соответствующем языке представления знаний	**Relations among the groups of concepts** *complex conceptual construction – complex concept – inner interpretation* **and** *structural description – knowledge representation language*	K_{AB}, $\mathrm{sig}\,(A, B)$
Фреймовые структуры реализуются на базе языков программирования высокого уровня, позволяющих человеку работать с информационной системой, используя лингвистические средства, близкие к языку межчеловеческого общения	**Relation among the concepts of** *structural description* **and** *high-level programming language*; **periphrase** *на основе* $\Longleftrightarrow$ *на базе*	K_{AB}

Were colored here:

$\mathrm{sig}\,(A, B)$ — if phrases were found only on a maximum of the «most strong» links;

K_{AB} — if phrases were found only on a total strength of the «most strong» links;

K_{AB} — if phrases were found both on a maximum of these links, and on their total strength.

Selection the relevant phrases: comparison with the decision based on TF-IDF

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
phrases selection according to TF-IDF of words of initial phrase										*selection by the total «strength» of the «most strong» links for* K_{AB}								
N	5	8	14	9	1	1	29	5	10	1	12	15	1	1	2	2	1	11
N_1	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	1
N_2	0	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	2
N_3	2	1	0	1	0	0	1	0	0	0	1	1	0	0	0	1	0	4
selection by the number of the «most strong» links for K_{AB}										*selection by the number of the «most strong» links for* $\mathrm{sig}\,(A, B)$								
N	2	4	1	3	2	1	6	1	5	3	2	32	1	2	1	18	1	3
N_1	0	1	0	1	2	1	0	0	0	0	0	0	0	0	0	1	0	0
N_2	0	0	0	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0
N_3	1	2	0	0	0	0	2	0	1	1	2	1	0	0	0	2	0	1

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 (maximum of number of the «most strong» links for K_{AB})

No. of initial phrase	Words which are not entered in most significant links	n-grams
	Philosophy and Methodology of Knowledge Engineering	
3	*с, информация, который, на*	*точка, зрения*
4		*в, факт, данный*
6	*при*	
9	*который, вывод, структурный, соответствие, различный, способ, ситуация*	
	Mathematical Methods for Learning by Precedents	
3	*заниженность*	
4	*заниженность, являться*	

Remark

In given illustration the comparison is made for those documents which were related to the most relevant for initial phrase at usage of both (3), and (5) variant for ranking function.

Comparison of n-grams and links most significant for phrases selection (maximum of number of the «most strong» links for K_{AB})

№	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	2	1	6	1	5	2	1	2	4	6	1	6	2	1
N_1	0	1	0	1	2	1	0	0	0	0	0	0	1	1	0	1	0	0
N_2	0	0	0	2	2	1	0	0	0	0	0	0	2	4	0	0	0	0
N_3	1	2	0	0	0	0	2	0	1	0	1	2	2	5	1	2	0	1
Mathematical Methods for Learning by Precedents																		
N	1	1	15	15	5	11	1	1	1	2	4	1	1	3	1	2	1	1
N_1	1	1	3	2	0	0	0	0	1	0	1	1	1	0	0	0	0	0
N_2	0	1	2	2	1	9	0	0	1	0	0	1	1	3	1	0	0	0
N_3	0	0	7	4	0	4	0	1	0	1	2	0	0	0	1	0	1	1

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 phrases selection 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*

№	1	2	3	4	5	6	7	8	9
Philosophy and Methodology of Knowledge Engineering									
N	13	67	2	15	29	30	79	224	20
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	2	5	0	1	1	2	3	2	2
Mathematical Methods for Learning by Precedents									
N	56	1	1	1	24	17	21	5	2
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	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 tools;
N_2 is the number of phrases representing synonyms;
N_3 is the number of phrases representing conceptual relations.

Main conclusions

1. The main *result* of current work is the *formation method* for topical corpus of texts relevant at described knowledge fragments to initial phrase 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, the *method* for text selection *offered in this work enables* a **15**-*times reduction (on average)* in the *output of phrases* which are *irrelevant to the initial one* in terms of either the described knowledge fragment or its expression forms in a given natural language.
3. The *improving of offered method by extraction of n-grams* on the found links of words *increases* the output of phrases representing *conceptual relations* if the percentage of general vocabulary and terms of subject area are comparable.

What requires the separate research ?

1. Extraction the image components of initial phrase from texts by analysis of occurrence of words from the cluster of greatest values of TF-IDF together with the n-grams on the found links of words.
2. How to interpret the TF-IDF metrics for mentioned n-grams ?
3. Estimation of precision of sentence boundary detection for different variants of classifier training.