Relevance of textual set to knowledge unit and estimation of affinity to sense standard for its linguistic expressional means

Mikhaylov D., Kozlov A., Emelyanov G.

Yaroslav-the-Wise Novgorod State University

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

October 8–12, 2018

Gaeta, Italy

Knowledge unit

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 (with the respect of possible synonyms).

Just such phrases represent a *sense standard*.

Main problems

- completeness of extraction of knowledge units from the texts of topical corpus by analyzing the relevance to initial phrase;
- search for the most rational linguistic variant which corresponds to the *sense standard* to describe the revealed knowledge fragment.

Sense standard

Is defined *by the set* of textual units and their links *necessary and enough* for representation of knowledge unit.

Main problems

- constancy of structure of initial set of semantically equivalent (SE) phrases under constructing the annotation;
- *precision* of standard's revelation is *essentially dependent* on completeness of description of *linguistic expressional forms* for knowledge unit;
- it is necessary *to parse* (fully or partially) *the initial SE-phrases* to search the most significant links and to calculate the distance statistics for linked words within separate phrases;
- *compatibility* of morphological characteristics and their tags used by different programs implemented morphological analysis.

Purpose of research

To find a *compromise* between the *accuracy of revelation of word relationships* most significant for linguistic representation of knowledge unit and the *number of initial SE-forms* of its description by expert.

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

$$\mathrm{tf}\,(t_i, d) = \frac{n_i}{\sum_k n_k}, \tag{1}$$

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:

$$\mathrm{idf}\,(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right), \tag{2}$$

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: 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).

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]:

$$K_{AB} = \frac{k}{a + b - k}, \tag{3}$$

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

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 and calculate the estimation (3) for them 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 (RMSD) 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 $\mathbb{T}(d)$.

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

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

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

and n-grams in $\mathbb{T}\,(d)$ are ordered descending the value of $\mathrm{N}(T, d) \cdot \mathrm{len}\,(T)$.

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*

$$\mathbb{W}(D) = \frac{1}{\left|D'\right|} \sum_{d \in D'} \left[\frac{\left|\left\{w \in b\colon\ \exists T \in \mathbb{T}'\,(d)\,,\, b \in T\right\}\right|}{\left|\left\{w\colon\ \exists Ts_i \in \mathrm{Ts},\, w \in Ts_i\right\}\right|} \sum_{T \in \mathbb{T}'(d)} \mathrm{N}(T, d) \right], \tag{6}$$

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).

Initial topical 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 topical 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 digit denotes* the *number according to the table* for initial phrase (see *Slides* 10 and 13).

Estimating the relevance of text corpus to initial knowledge units

No of init. phrase or phrase group[1]		taking into account of Preps/Conjs	excluding Preps/Conjs
	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	
2		0,1259124	0,0666667
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
		for groups of initial phrases	
1		0,5516767	0,6094707

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

Adding phrases of annotation to the set of initial ones

Estimation (6) for relevance of text corpus to knowledge unit and situation of natural language usage	
taking into account of Preps and Conjs	excluding Preps and Conjs
Initial phrase №9, Philosophy and Methodology of Knowledge Engineering	
0,5758868	0,3178877
Step 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».	
1,5439084	1,8877527
Step 2	
«Каждое выражение, входящее во фрейм, каждый знак в нём, несущий самостоятельную информационную нагрузку, являются интерпретированными, т. е. заключают в себе смысл, заложенный человеком с помощью соответствующей программы или же сконструированный системой».	
0,9197011	3,8257502
Attempt for Step 3	
«Перспективы развития эффективных систем представления знаний на основе естественного языка, т. е. вербальных знаковых систем, понятных человеку без особо сложной выучки, сегодня во многом связаны с построением так называемых фреймов».	
3,5916774	0,2568324

Numerical estimation for affinity of phrase to sense standard: basic empirical considerations

1. The affinity to standard for phrase should be evaluated basing on results of classifying of its words according to their TF-IDF together with estimation of coupling strength for combinations of words in this phrase.
2. The estimation of coupling strength of words should be calculated concerning not separate texts, but all considered topical text set (corpus).
3. Respecting the requirement of minimization of phrase length, actual here is to consider only those links that are syntactical in nature.
4. The division of words of initial phrase into general vocabulary and terms according to their TF-IDF should be expressed as much as possible.
5. In clusters which were formed for words of initial phrase according to their TF-IDF relatively to some corpus document the words must be distributed more or less evenly.

Let $H_1, \dots H_r$ be the sequence of clusters formed by TF-IDF values of words of initial phrase relatively to the document d of corpus D.

Documents $d \in D$ are sorted descending the product of estimations:

$$val_1 = -\frac{1}{\log_{10}\left[\sqrt{\Sigma^2_{H_1} + \Sigma^2_{H_{r/2}} + \Sigma^2_{H_r}}\right]} \tag{7}$$

and, correspondingly,

$$val_2 = 10^{-\sigma\left(\left|H_i,\, i=1,\dots,r\right|\right)}, \tag{8}$$

where $\Sigma^2_{H_1}$, $\Sigma^2_{H_{r/2}}$ and $\Sigma^2_{H_r}$ are the squares of the sums of TF-IDF values for words related, correspondingly, to the clusters H_1, $H_{r/2}$ and H_r;

$\sigma\left(\left|H_i,\, i = 1, \dots, r\right|\right)$ is the RMSD of the number of cluster elements.

Remark

Besides H_1 and $H_{r/2}$, the meaningful interest here also have the cluster H_r, to which *the terms prevailing in corpus* are correspond.

To estimate the affinity to standard for each initial phrase Ts_i of group Ts a pair of values (val_1, val_2) is taken concerning $d \in D$ with maximal $val_1 \cdot val_2$. Then val_1 and val_2 are divided by their maximums in Ts and transform to $[0, 1]$.

Normalized val_1 and val_2 let's denote further as val'_1 and val'_2.

Coupling strength of words and estimation of affinity to standard for phrase

Let's enter the next variant of estimation (3) for coupling strength of words, K'_{AB}:

- the link between the words A and B must be of a syntactic nature;
- values of a, b and k in (3) must be calculated relatively to whole corpus D;
- the estimation is calculated only at presence of TF-IDF value at least one of the words (A, B) either in cluster $H_1(X)$, or in $H_{r/2}(X)$.

Let

$R(Ts_i, d)$ be the set of those links of words in a phrase Ts_i, for which the estimation K'_{AB} is defined relatively to the document d;

$R_1(Ts_i, d)$ be the set of links related to the cluster of greatest values of mentioned estimation;

$K_1(Ts_i, d)$ be the sum of values of estimation K'_{AB} for the links from $R_1(Ts_i, d)$.

Then the estimation of affinity to standard for phrase Ts_i relatively to $d \in D$ is defined by analogy with the document ranking by relevance to initial phrase as

$$W^R(Ts_i, d) = K_1(Ts_i, d) \frac{\left|R_1(Ts_i, d)\right|}{\left|R(Ts_i, d)\right|}. \qquad (9)$$

Let's designate further as val'_3 the maximal value of estimation (9) for Ts_i over all $d \in D$ transformed to $[0, 1]$ by division by its maximum for all $Ts_i \in \mathrm{Ts}$.

Let Val' be a triple of values $val'_1 \cdot val'_2$, val'_3 with taking and val'_3 without taking into account of prepositions/conjunction for initial phrase Ts_i from group Ts.

Let's enter into consideration the RMSD, difference ($\max - \min$) and quotient ($\max / \min$) of greatest and least value in Val' (further — the *RMSD-estimations*).

The phrase cannot be related to the «standard» if:

- by to one of the values $val'_1 \cdot val'_2$ or val'_3 the phrase refers to the cluster of greatest, but according to another — to cluster of least values; herewith one of its RMSD-estimations is related to the cluster of greatest values;
- simultaneously by the values of $val'_1 \cdot val'_2$ and val'_3 (both with and without taking into account of prepositions and conjunctions) the phrase is related to the cluster of least values;
- according to any of the RMSD-estimations the phrase can be related, but none of the values $val'_1 \cdot val'_2$ and val'_3 is suitable for relating the phrase to the cluster of greatest values.

Eventually, *the standard* will be *defined* by the frases from related to the clusters of greatest values of $val'_1 \cdot val'_2$ and val'_3, herewith if the phrase satisfies one of the three aforementioned conditions then it will be *exclude from consideration*.

№ Russian phrase from the set of equivalent by sense

1 *Нежелательная переподгонка является причиной заниженности средней величины ошибки алгоритма на обучающей выборке.*

2 *Нежелательное переобучение приводит к заниженности эмпирического риска.*

3 *Заниженность эмпирического риска является следствием нежелательного переобучения.*

4 *Нежелательное переобучение служит причиной заниженности эмпирического риска.*

5 *Заниженность эмпирического риска является результатом нежелательного переобучения.*

6 *Заниженность эмпирического риска связана с переобучением.*

7 *Заниженность эмпирического риска относится к следствию нежелательного переобучения.*

8 *Заниженность эмпирического риска связана с нежелательным переобучением.*

9 *Нежелательное переобучение является причиной заниженности эмпирического риска.*

10 *Нежелательная переподгонка приводит к заниженности эмпирического риска.*

11 *Нежелательная переподгонка, следствием которой является заниженность эмпирического риска.*

12 *Эмпирический риск, к заниженности которого ведёт нежелательная переподгонка.*

13 *Риск, заниженный как следствие переподгонки.*

Initial data for estimating the affinity to the sense standard

Phrase No.	Maximum values for the text corpus			
	estimation (7)	estimation (8)	estimation (9) without taking into account of Preps/Conjs	estimation (9) with taking into account of Preps/Conjs
1	0,4671832	0,0802703	0,3294206	0,8005671
2	0,4314354	0,3162278	0,2000000	0,5555556
3	0,4317240	0,2646365	0,1428571	0,1428571
4	0,4314354	0,3162278	0,1666667	0,1666667
5	0,4102928	0,3906175	0,1428571	0,1428571
6	0,4313110	0,2646365	0,3333333	0,2500000
7	0,4318168	0,1525820	0,1428571	0,1666667
8	0,4313110	0,3162278	0,3333333	0,2500000
9	0,4232485	0,2833201	0,1666667	0,1666667
10	0,4314424	0,3162278	0,0444445	0,2000000
11	0,4188872	0,1275184	0,0555556	0,0555556
12	0,4317911	0,1525820	0,2222222	0,1666667
13	0,3896160	0,3162278	0,0416667	0,0416667

Clustering of phrases from represented on the Slide 21

Cluster No.	*Numbers on the Slide 21 for phrases related to the cluster (descending order of estimation)*
on the base of $val'_1 \cdot val'_2$	
1	5, 10, 2, 4, 8, 13, 9, 3, 6
2	7, 12, 11
3	1
on the base of val'_3 *with taking into account of prepositions and conjunctions*	
1	1, 2
2	6, 8
3	10, 4, 7, 9, 12, 3, 5
4	11, 13
on the base of $\max - \min$ *in* Val'	
1	1
2	5, 6, 8, 13, 10, 4, 12, 9, 3
3	11, 7
4	2

Cluster No.	*Numbers on the Slide 21 for phrases related to the cluster (descending order of estimation)*
on the base of RMSD for values of elements in Val'	
1	1
2	5, 8, 6, 13, 10, 4, 12, 9, 3
3	7, 11
4	2
on the base of val'_3 *without taking into account of prepositions and conjunctions*	
1	6, 8, 1
2	12
3	2, 4, 9, 3, 5, 7
4	11, 10, 13
on the base of $\max / \min$ *in* Val'	
1	13
2	10, 5, 1
3	11, 4, 3, 6, 8, 12, 9
4	7, 2

Iterative selection of phrases and including them into the set of initial phrases for the example on the Slide 16

Phrase No.	Maximum values for the text corpus			
	estimation (7)	estimation (8)	estimation (9) without taking into account of Preps/Conjs	estimation (9) with taking into account of Preps/Conjs
1	*«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».*			
	0,4809750	$1{,}480 \cdot 10^{-4}$	0,4866667	0,1041667
2	*«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».*			
	0,5740613	$2{,}854 \cdot 10^{-4}$	0,1477941	0,1322368
3	*«Каждое выражение, входящее во фрейм, каждый знак в нём, несущий самостоятельную информационную нагрузку, являются интерпретированными, т. е. заключают в себе смысл, заложенный человеком с помощью соответствующей программы или же сконструированный системой».*			
	0,5709775	$7{,}693 \cdot 10^{-4}$	2,1250000	0,2222222

Clustering of phrases from represented on the Slide 24

Cluster No.	*Numbers on the Slide 21, for phrases related to the cluster (descending order of estimation)*
on the base of $val'_1 \cdot val'_2$	
1	3
2	2
3	1
on the base of val'_3 *with taking into account of prepositions and conjunctions*	
1	3
2	2, 1
on the base of $\max - \min$ *in* Val'	
1	2
2	1
3	3

Cluster No.	*Numbers on the Slide 21, for phrases related to the cluster (descending order of estimation)*
on the base of RMSD for values of elements in Val'	
1	2, 1
2	3
on the base of val'_3 *without taking into account of prepositions and conjunctions*	
1	3
2	1
3	2
on the base of $\max / \min$ *in* Val'	
1	2
2	1
3	3

Reducing of volume of textual information necessary for expression the knowledge unit

Estimated as $(l_1 \cdot n_1)/(l_2 \cdot n_2)$, where n_1 and n_2 are the number of phrases representing the knowledge unit and defining the standard, respectively; l_1 and l_2 are the maximal length of phrase (in words) form representing the knowledge unit and from defining the standard.

Conclusions

1. The main *result* of current work is the *method* for estimation of affinity to sense standard for natural-language phrase relatively to the knowledge unit represented by it.
2. The evident *advantage* of the offered method is that *there is no need* to describe as many equivalent-by-sense expressional forms as possible for corresponding knowledge unit in a given natural language.
3. The offered method for revelation of sense standard gives *at least a two-fold* reduction *in the textual information* necessary for lossless-in-sense *transmission of knowledge unit* in a given natural language.
4. The open problem is that *the results given by* the proposed solutions *significantly depend* from the selection of texts into corpus by expert. This takes into account the level of complexity of the text selected to corpus, and its significance in solved task.
5. It is of interest *to study* the dynamics of changing the estimation (3) concerning the different corpus documents for syntactically related words of initial phrase.