Sense's standards and transmission of knowledge for their estimation applying the open-form tests

Dmitry Mikhaylov and Gennady Emelianov

Yaroslav-the-Wise Novgorod State University

11^{th} International Conference
«Pattern Recognition and Image Analysis:
New Information Technologies» (PRIA-11-2013),

23–28 September, 2013

Samara, Russian Federation

Research subject, purpose, and considered problem

Research subject

Methods and algorithms for formation of knowledge about synonymy in Natural Language (NL).

Considered problem

To transmit a knowledge represented by texts on some natural language between its native speakers (experts and trainees, correspondingly).

Main purpose of research

Theoretical reasoning of structure of knowledge about synonymy.
Development of methods and algorithms for forming these knowledge and application them for the family of tasks of:

- sense-similarity's estimation of texts in subject-oriented natural language;
- computer-aided filling and compression of language's and subject-area's knowledge base;
- seeking a most rational plan for sense's transfer among different native speakers;
- coordination of knowledge units which were formed by various experts.

Usage situation for natural language

Definition 1.4

Usage Situation for Natural Language (USNL) is the description of new social experience (the content of joint actions) by means of this natural language.

The NL-context accumulated by some USNL S can be represented by a triple:

$$S = (O, R, Ts)\,, \tag{1}$$

where O is a set of symbols associated with reality concepts;
Ts is a set of description forms for S in some sign system;
$R \subset O^n$, where $n \in 1, \ldots, |O|$.

Let $Synt$ be a surjective function determined by syntax of given NL;

Then for $\forall\ Ts_i \in Ts\ \exists\ Tr_i : Ts_i = Synt\,(Tr_i)$, where Tr_i is a marked tree.

Thus if $O = M \cup V,\ M \cap V \neq \emptyset$, then for $\forall\ o_j \in M$ it will be found $o_k \in V$ with the following correspondence in the tree Tr_i:

- to the concept o_j there corresponds a child node with the label w_j,
- to the concept o_k there corresponds a parent node with the label w_k.

Formal Concept Analysis and formation of knowledge units

Let's represent a single USNL by a Formal Context (FC):

$$K = (G, M, I)\,, \tag{2}$$

where $\forall g \in G$ is a stem of word syntactically submitted to any other word from some $Ts_i \in Ts$ in model (1).

The attributes's set M comprises the subsets designated by corresponding bottom indexes:

- M_1 is the set of indications to the syntactically main word's stem;
- M_2 is the set of indications to the main word's inflection;
- M_3 is the set of «stem–inflection» relations for a syntactically main word;
- M_4 is the set of combinations of inflections of dependent and main words. After an inflection via the colon, a preposition is shown (if any) that provides a relation with a dependent word;
- M_5 is the set of indications for dependent word's inflection.

It is required: to form $I \subseteq G \times M$ by analysis of symbolic structure and choice of $Ts_i \in Ts$ with a *minimal* length and *maximum* of words most generally *used* in different phrases from Ts (taking possible synonyms into account).

Problem statement

Let

W_{ij} be the ordered sequence of symbols constituting the word w_{ij},

Wc_{ij} be the ordered sequence of symbols for invariant part (i.e. stem) of word,

Wf_{ij} be the ordered sequence of symbols for word's inflection,

$\odot$ be the designation for concatenation operations.

It is given:

$Ts = \left\{ Ts_i \colon Ts_i = \bigodot_j W_{ij} \right\}$ is a set of Semantically Equivalent (SE) phrases defining the USNL.

It is required to find:

$Pw_i = \left\{ \left(Wc_{ij}, Wf_{ij} \right) : Wc_{ij} \odot Wf_{ij} = W_{ij} \right\}$ for all $i = 1, \dots, \left| Ts \right|$.

Occurrence of symbols for given positions in the word concerning the USNL

Let's identify concepts: with a stem Wc_{ij} — the «prefix»
and with inflection Wf_{ij} — the «suffix» as accepted in informatics.

These are paramount procedures and functions of algorithm:

- $pref.show\,(w_{ij})$ return the current value of prefix for the word w_{ij};
- $pref.inc\,(w_{ij})$ increments the length of prefix for the word w_{ij} by 1;
- $prefs$ forms the lists by grouping of wordforms similar in prefix with sorting them by length's decrease;
- $pref.check\,(Prf)$ for group of wordforms with common prefix Prf carry out analysis of absolute frequencies of occurrence of characters on various positions concerning the word's front and end. Note that the frequency ν_p of occurrence of symbol which is the first at the left is always maximal as well as for symbols in Prf. Relative to word's end the search of common suffix's symbols having the occurrence's frequency ν_p, is carry out with including them into word's inflection. *The total length of common prefix and suffix must be, at least, a one third of word's length, and lengths's difference for a pair of words having a common prefix is always less than half a length of smaller word (independently from suffix!).*

pseudocode description

software realization

continue

Algorithm revealing the stems and inflections for words concerning the USNL

```
Require: $Ts$;
Ensure: $Pw = \bigcup_{i=1}^{|Ts|} Pw_i$;
 1: $Pw := \varnothing$;
 2: for all $W_{ij}$: $\underset{j}{\odot} W_{ij} = Ts_i$, where $Ts_i \in Ts$ do
 3:     $Wc_{ij} := \{W_{ij}[1]\}$; $Wf_{ij} := \overset{|W_{ij}|}{\underset{k=2}{\odot}} W_{ij}[k]$;
 4: end for // initialization of stems and inflections
 5: $prefs(PrfsTmp)$;
 6: if $PrfsTmp = \varnothing$ then
 7:     return $Pw$ and exit the algorithm;
 8: else
 9:     take $Prf$ from $PrfsTmp$;
10:     if $pref.check(Prf) = \text{true}$ then
11:         $Pw := Pw \cup \left\{ \left(Prf, Wf_{ij}(Prf)\right) \middle| pref.show(w_{ij}) = Prf \right\}$;
12:         $PrfsTmp := PrfsTmp \setminus \{Prf\}$;
13:         go to the Step 6;
14:     else
15:         for all $w_{ij}$: $pref.show(w_{ij}) = Prf$ do
16:             $pref.inc(w_{ij})$;
17:         end for
18:         go to the Step 5
19:     end if
20: end if
```

Models of linear structures for NL-phrases defining the standard

Let Ts be a set of SE-phrases defining some USNL according to (1),

J be an index set for invariant parts of words of phrases from Ts.

Definition 3.2

The ordered sequence of indexes of invariant parts of words for some $Ts_i \in Ts$ let's name as Model of its Linear Structure (MLS), $Ls\,(Ts_i)$.

Let $\left\{J_1, J_2\right\}$ be the pair of sequences of indexes in $Ls\left(Ts_i\right)$, where $J_1 = \left\{j_1^1, \dots\ , j_2^1\right\}$, $J_2 = \left\{j_1^2, \dots\ , j_2^2\right\}$, and both $\left(j_1^1, j_2^1\right)$ and $\left(j_1^2, j_2^2\right)$ correspond to the syntactic links.

The sense standard for USNL is defined by those $Ts_i \in Ts$, in MLS of which

$$(J_1 \subset J_2) \vee (J_2 \subset J_1) \vee \left(\mid J_1 \cap J_2 \mid = 1\right) \vee \left(J_1 \cap J_2\right) = \varnothing, \qquad (3)$$

and summary length of sequences of mentioned kind for all syntactic links revealed on Ts_i has to be *minimum*.

Sense standard for USNL and word's occurrence frequency

Let $fr\left(w_j\right|$ be a frequency of occurrence of word w_j in all $Ts_i \in Ts$.

So the most informative words in Ts are forming a cluster $Clust$:

- the word with a maximal value of this frequency will be in $Clust$;
- for $\forall\ \{w_j, w_k\} \subset Clust$ and $\forall\ w_l \notin Clust$ is *true* that

$$\Big(\big|\ fr\left(w_j\right) - fr\left(w_k\right)\big| < \big|\ fr\left(w_j\right) - fr\left(w_l\right)\big|\Big) \wedge \\ \wedge \Big(\big|\ fr\left(w_j\right) - fr\left(w_k\right)\big| < \big|\ fr\left(w_k\right) - fr\left(w_l\right)\big|\Big) = \text{true} \qquad (4)$$

The basis of standard are made the phrases with maximum of words in $Clust$.

Here for words from $Clust$ possible synonyms and different orders in a phrase are taking into account.

Let LS be a set of linear structures's models given on J for sentenses from Ts.

Lemma 5.1

The pair of indexes $\{j_1, j_2\} \subset J$ corresponds to synonymic words and can be replaced by one index from $(\mathbb{N} \backslash J)$ if $\exists\ \{Ls\left(Ts_1\right), Ls\left(Ts_2\right)\} \subseteq Ls$:

$$Ls\left(Ts_1\right) = J_1 \odot \{j_1\} \odot J_2 \text{ and } Ls\left(Ts_2\right) = J_1 \odot \{j_2\} \odot J_2,$$

where $J_1 \subset J$, $J_2 \subset J$, and $\odot$ is the concatenation operation at the set J.

Problem statement

Let J_{Cl} be a set of indexes of words related to the cluster of most informative concerning to USNL given by the set of SE-phrases Ts;

$frq\left(\left(j,k\right),LS\right)$ be the frequency of occurrence of the pair (j,k) in the models from LS taking into account that $(j,k) \Leftrightarrow (k,j)$.

Then USNL's standard is defined by phrases with MLSs belonging to the set

$$LC = \bigcup_i LS_i : LS_i \subset LS, \ \exists \ \{Ts_i, Ts_j\} \in Ts:$$

$$Ls\left(Ts_i\right) \in LS_i$$

$$\left|Ls\left(Ts_i\right) \cap J_{Cl}\right| \to \text{max}$$

$$\Big(\left(Ls\left(Ts_j\right) \in LS_i\right) \wedge \left(Ts_j \neq Ts_i\right)\Big) \to \left(Ls\left(Ts_i\right) \cap J_{Cl}\right) \subset Ls\left(Ts_j\right),$$

and attributes set's forming for USNL's standard in a form of FC (2) requires:

- to find index pairs $(j,k) : frq\left(\left(j,k\right),LS\right) > 1$, which satisfy the condition (3), for all linear structure's models from LC;
- to define the direction of syntactic link for each found pair (j,k);
- to eliminate from $\forall \ LS_i \subset LC$ any MLS containing indexes which not appeared in any found link.

False links and links revealed earlier

There are three stages to find $Dir\,(j,k)$, $Dir \in \{\leftarrow, \rightarrow\}$, namely:

- checking the link corresponding to (j,k) on falsity's condition's fulfilment;
- an attempt to identify with the links revealed earlier;
- if there are no identification with known links then interview with expert.

Let $St\,(j)$, $St\,(k)$ and $St\,(l)$ are the word's stems corresponding to j, k and l.

For given USNL the link for (j,k) is identified as false if $j,k,l \in Ls\,(Ts_i)$ in some $Ts_i \in Ts$, but another USNL has false link for $St\,(j)$ and $St\,(k)$, and true link either between $St\,(j)$ and $St\,(l)$ or between $St\,(k)$ and $St\,(l)$.

Let Lnk be a set of links revealed earlier, each of which is represented by:

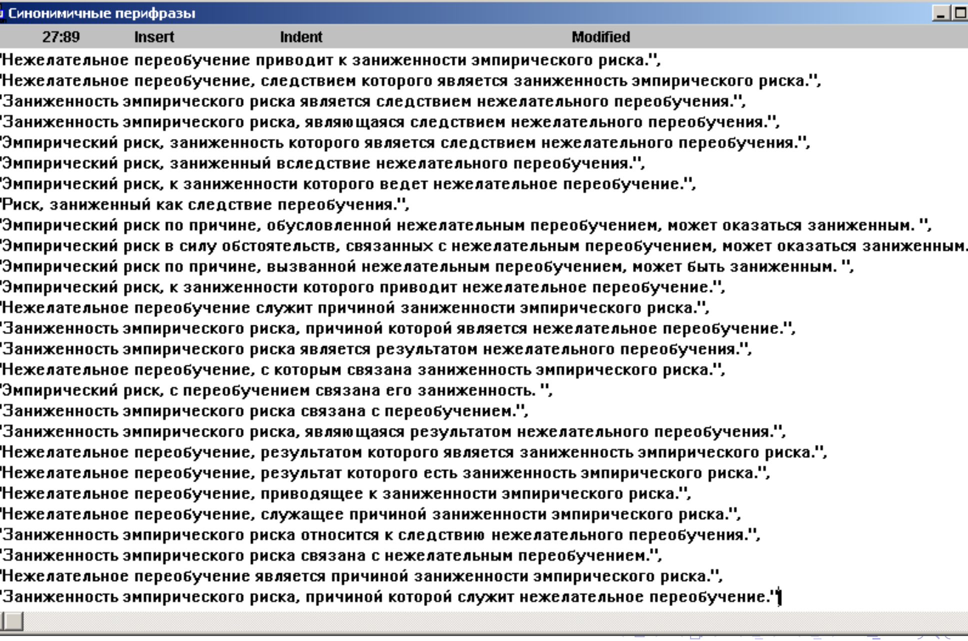
- an ID number of USNL (Id);
- a main word's stem (St_1);
- a stem for dependend word (St_2);
- a list of inflections combinations «main word–dependent word» (FCm).

A pair (j,k) is put in conformity of link $((j,k)\,,\rightarrow)$ if for some other USNL

$\exists\,(Id, St_1, St_2, FCm) \in Lnk$:

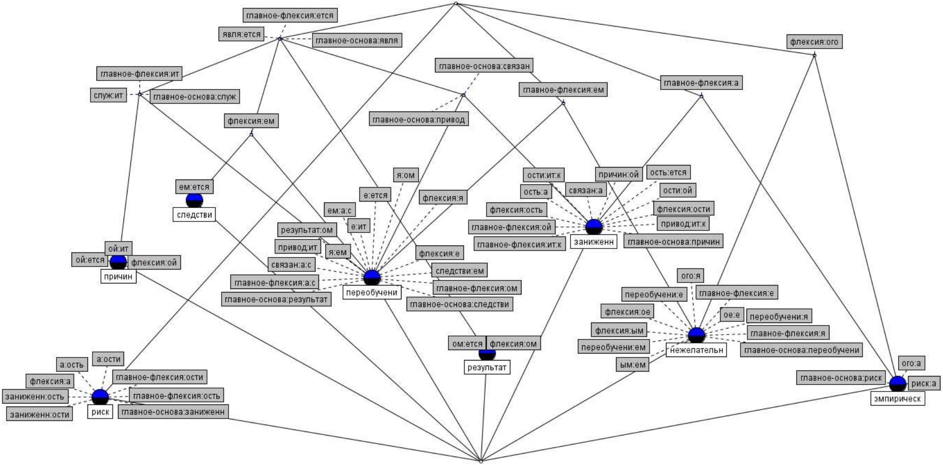
$$St\,(j) = St_1,\;\; St\,(k) = St_2 \text{ and } \Big(Fl\,(j)\,, Fl\,(k)\Big) \in FCm.$$

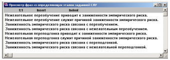
Example of initial set of semantically equivalent phrases defining the USNL



"Нежелательное переобучение приводит к заниженности эмпирического риска.",
"Нежелательное переобучение, следствием которого является заниженность эмпирического риска.",
"Заниженность эмпирического риска является следствием нежелательного переобучения.",
"Заниженность эмпирического риска, являющаяся следствием нежелательного переобучения.",
"Эмпирический риск, заниженность которого является следствием нежелательного переобучения.",
"Эмпирический риск, заниженный вследствие нежелательного переобучения.",
"Эмпирический риск, к заниженности которого ведет нежелательное переобучение.",
"Риск, заниженный как следствие переобучения.",
"Эмпирический риск по причине, обусловленной нежелательным переобучением, может оказаться заниженным. ",
"Эмпирический риск в силу обстоятельств, связанных с нежелательным переобучением, может оказаться заниженным.
"Эмпирический риск по причине, вызванной нежелательным переобучением, может быть заниженным. ",
"Эмпирический риск, к заниженности которого приводит нежелательное переобучение.",
"Нежелательное переобучение служит причиной заниженности эмпирического риска.",
"Заниженность эмпирического риска, причиной которой является нежелательное переобучение.",
"Заниженность эмпирического риска является результатом нежелательного переобучения.",
"Нежелательное переобучение, с которым связана заниженность эмпирического риска.",
"Эмпирический риск, с переобучением связана его заниженность. ",
"Заниженность эмпирического риска связана с переобучением.",
"Заниженность эмпирического риска, являющаяся результатом нежелательного переобучения.",
"Нежелательное переобучение, результатом которого является заниженность эмпирического риска.",
"Нежелательное переобучение, результат которого есть заниженность эмпирического риска.",
"Нежелательное переобучение, приводящее к заниженности эмпирического риска.",
"Нежелательное переобучение, служащее причиной заниженности эмпирического риска.",
"Заниженность эмпирического риска относится к следствию нежелательного переобучения.",
"Заниженность эмпирического риска связана с нежелательным переобучением.",
"Нежелательное переобучение является причиной заниженности эмпирического риска.",
"Заниженность эмпирического риска, причиной которой служит нежелательное переобучение."

Resulting formal context of sense standard and NL-phrases defining standard



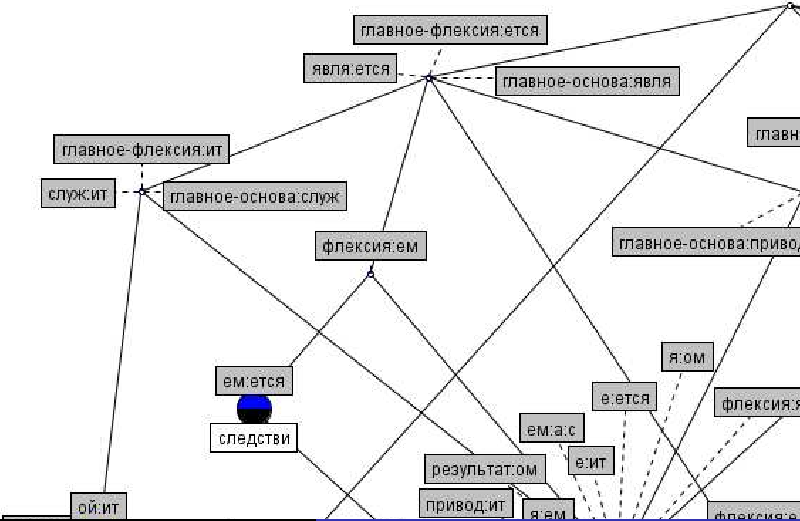


Нежелательное переобучение приводит к заниженности эмпирического риска.
Нежелательное переобучение служит причиной заниженности эмпирического риска.
Заниженность эмпирического риска связана с переобучением.
Заниженность эмпирического риска связана с нежелательным переобучением.
Нежелательная переподгонка приводит к заниженности эмпирического риска.
Нежелательная переподгонка служит причиной заниженности эмпирического риска.
Заниженность эмпирического риска связана с переподгонкой.
Заниженность эмпирического риска связана с нежелательной переподгонкой.

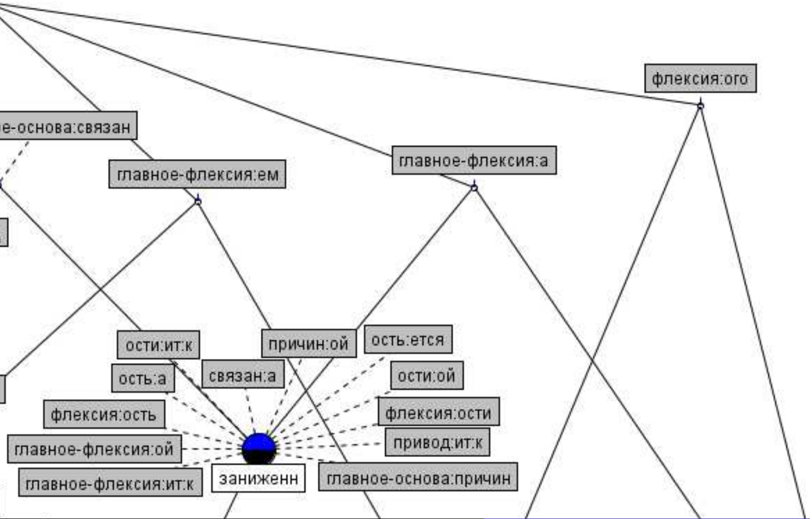
go back

continue

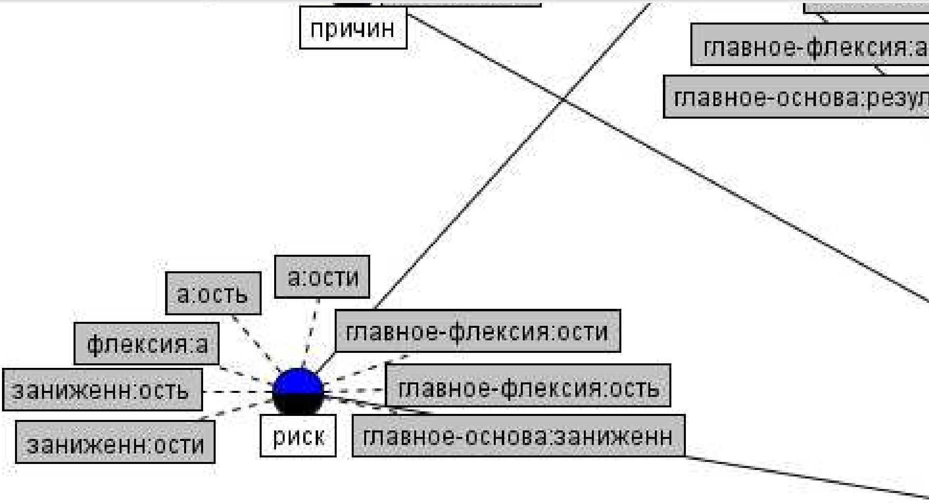
Resulting formal context of sense standard and NL-phrases defining standard



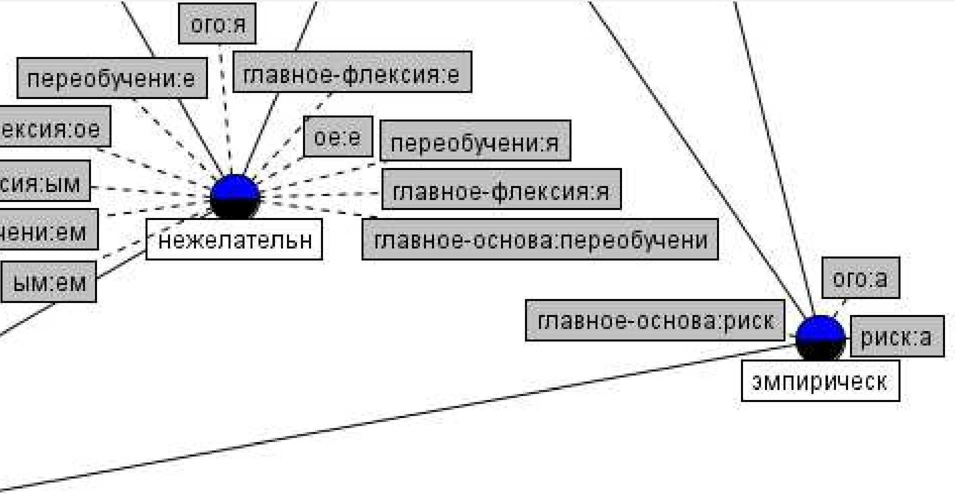
Resulting formal context of sense standard and NL-phrases defining standard



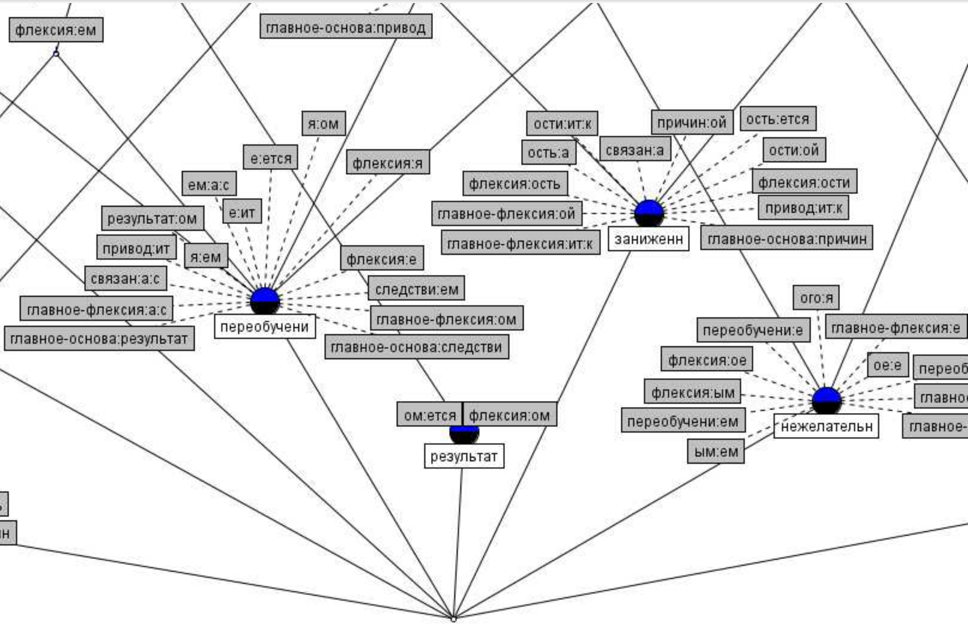
Resulting formal context of sense standard and NL-phrases defining standard



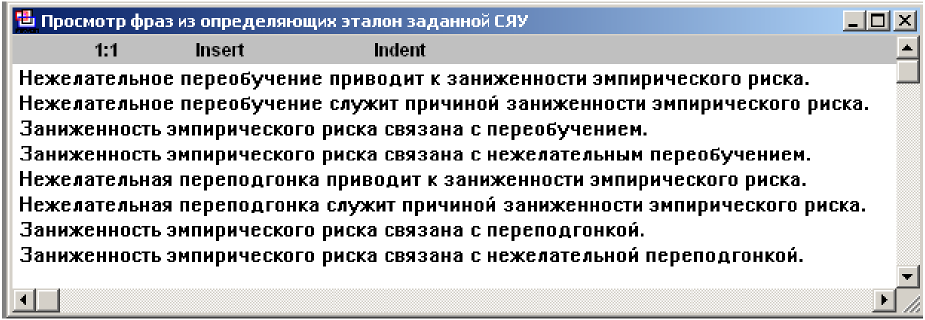
Resulting formal context of sense standard and NL-phrases defining standard



Resulting formal context of sense standard and NL-phrases defining standard



Resulting formal context of sense standard and NL-phrases defining standard



Просмотр фраз из определяющих эталон заданной СЯУ

Нежелательное переобучение приводит к заниженности эмпирического риска.
Нежелательное переобучение служит причиной заниженности эмпирического риска.
Заниженность эмпирического риска связана с переобучением.
Заниженность эмпирического риска связана с нежелательным переобучением.
Нежелательная переподгонка приводит к заниженности эмпирического риска.
Нежелательная переподгонка служит причиной заниженности эмпирического риска.
Заниженность эмпирического риска связана с переподгонкой.
Заниженность эмпирического риска связана с нежелательной переподгонкой.

Sense standards and description of USNL by means of objects and attributes

Serial number of USNL, i	1	2	3	4	5	6
Number of SE-phrases defining USNL	56	28	29	30	6	10
including representatives of standard	8	9	7	9	1	2
Initial number of objects for USNL	18	17	15	13	12	14
Initial number of attributes for USNL	177	186	173	162	94	81
Number of standard's objects	9	12	12	11	8	12
Number of standard's attributes	82	90	80	69	35	53

i What does the situation of language usage represents in Russian ?

1 Связь переобучения с эмпирическим риском

2 Связь переусложнения модели с заниженностью средней ошибки на тренировочной выборке

3 Влияние переподгонки на частоту ошибок дерева принятия решений

4 Причина заниженности оценки обобщающей способности алгоритма

5 Зависимость оценки ошибки распознавания от выбора решающего правила

6 Зависимость обобщающей способности логического алгоритма классификации от числа закономерностей алгоритмической композиции

example
coordination of knowledge
estimating the amount of memory

Estimating the amount of memory for storing NL-phrase

i	1	2	3	4	5	6
n	12	15	16	17	10	14
$vol\left(n\right)$	$4.790 \cdot 10^{8}$	$1.308 \cdot 10^{12}$	$2.092 \cdot 10^{13}$	$3.557 \cdot 10^{14}$	$3.629 \cdot 10^{6}$	$8.718 \cdot 10^{10}$
$vol_1\left(n\right)$	648	795	416	442	20	42
$vol_2\left(n\right)$	168	225	80	187	20	42

Here:

i is the serial number of USNL;
n is the maximal number of words in a phrase;
$vol\left(n\right) = n!$ is the estimation which is taken usually;
vol_1 and vol_2 are the estimations received with application of method and algorithms of NL-usage's situation's standard's revelation.

Numerically:

$vol_1\left(n\right) = l_1 \cdot n$ is the upper estimation, l_1 is the number of SE-phrases defining the USNL;
$vol_2\left(n\right) = l_2 \cdot n$ is the lower estimation, l_2 is the number of SE-phrases defining the standard of USNL.

Thesaurus and knowledge coordination for different situations of NL-usage

Let thesaurus to be represented in the form of formal context

$$Kth = (Gth, Mth, Ith)\,, \tag{5}$$

where Gth consists of symbolic labels of individual NL-usage's situations;

Mth includes the attributes of formal context (2) for each $gth \in Gth$.

In addition, in Mth one can distinguish the following subsets:

- M_6 is the set of indications to objects of formal contexts (2) generated for individual $gth \in Gth$;
- M_7 is the set of «stem–inflection» combinations for a syntactically dependent word;
- M_8 contains combinations of stems of the dependent and main word.

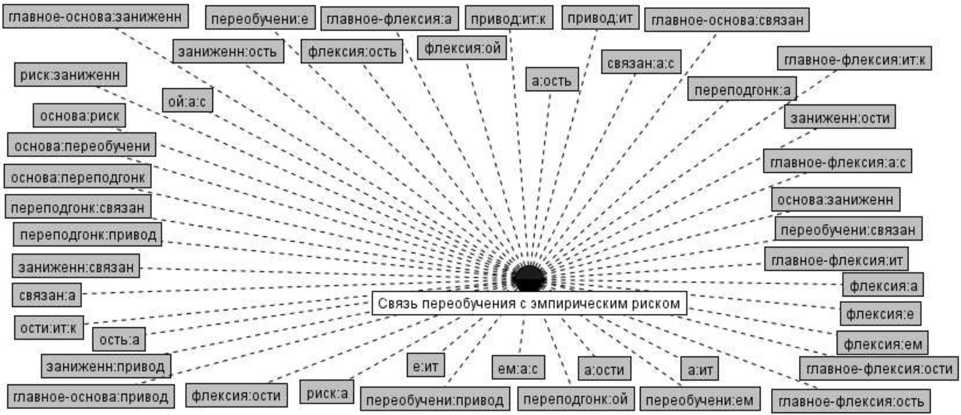
By analogy with the formal context (2) of individual USNL $Ith \subseteq Gth \times Mth$.

example of representation of individual USNL in the thesaurus's formal context

In this case the numerical estimation of similarity of NL-usage situations is determined by the number of attributes be shared by objects of compared situations concerning the formal context of thesaurus.

coordination of knowledge concerning different situations of NL-usage

USNL as an object of the thesaurus's formal context



главное-основа:заниженн
переобучени:е
главное-флексия:а
привод:ит:к
привод:ит
главное-основа:связан
заниженн:ость
флексия:ость
флексия:ой
связан:а:с
главное-флексия:ит:к
риск:заниженн
а:ость
переподгонк:а
ой:а:с
основа:риск
заниженн:ости
основа:переобучени
главное-флексия:а:с
основа:переподгонк
основа:заниженн
переподгонк:связан
переобучени:связан
переподгонк:привод
главное-флексия:ит
заниженн:связан
флексия:а
связан:а
Связь переобучения с эмпирическим риском
флексия:е
ости:ит:к
ость:а
флексия:ем
заниженн:привод
е:ит
ем:а:с
а:ости
а:ит
главное-флексия:ости
главное-основа:привод
флексия:ости
риск:а
переобучени:привод
переподгонк:ой
переобучени:ем
главное-флексия:ость

go back to definition of thesaurus's formal context

continue

Coordination of knowledge about synonymy concerning different situations of natural language's usage

Let

St be the designation for word's *invariant part* identified with the *stem*;
Fl be the designation for word's *inflection*;
S_1 and S_2 be the some situations of givel NL's usage.

Let's suppose that some Wrd can be represented as $St_1 \odot Fl_1$ concerning S_1, and as $St_2 \odot Fl_2$ — concerning S_2. At that $St_1 = St_2 \odot Sf$, where Sf contains one symbol as minimum, and $\odot$ is the operation of strings's concatenation.

Then concerning S_1 the following replacements can be implemented: the stem St_1 is replaced with St_2, and inflection Fl_1 — with $Fl_3 = Sf \odot Fl_2$ only if the frequencies of occurrence of inflections Fl_3 and Fl_2 in all lexico-syntactic links represented by the formal context (5) for given subject area won't decrease at fulfillment of these changes.

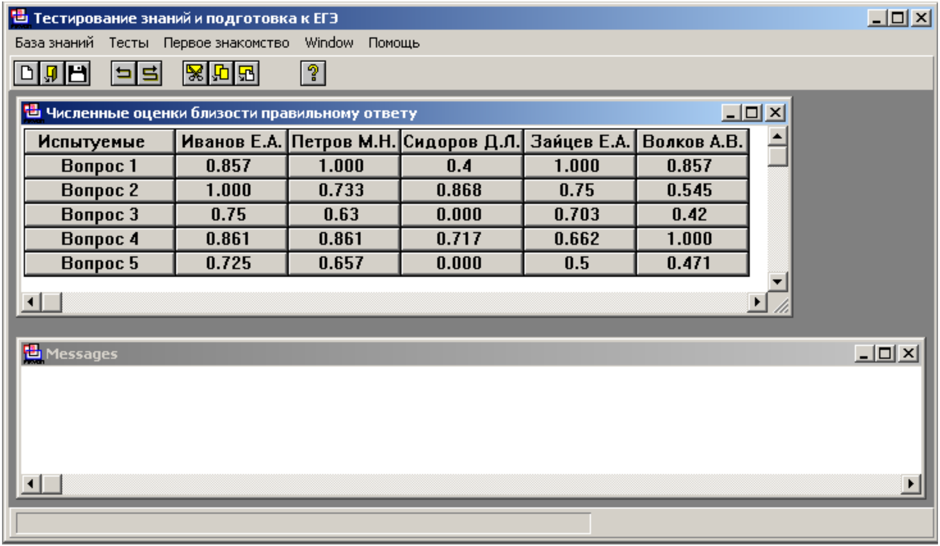
Example (in Russian).

USNL №3, $St_1 =$ «является», $Fl_1 =$ «»,
USNL №1, $St_2 =$ «явля», $Fl_2 =$ «ется», $Sf =$ «ется».
Concerning the USNL №3 the replacement of Fl_1 to $Fl_3 =$ «ется» is fulfilled.

go back

continue

Knowledge-control system as the practical issue of developed approach


Тестирование знаний и подготовка к ЕГЭ
База знаний Тесты Первое знакомство Window Помощь
Численные оценки близости правильному ответу
Messages

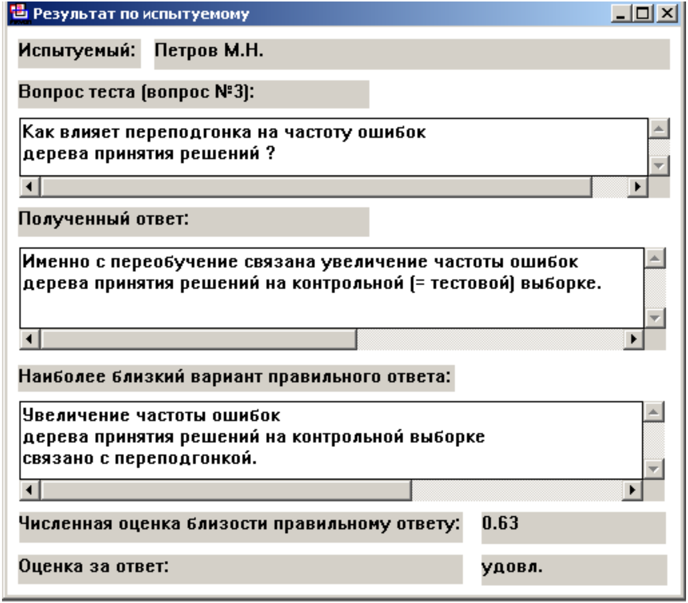
Испытуемые	Иванов Е.А.	Петров М.Н.	Сидоров Д.Л.	Зайцев Е.А.	Волков А.В.
Вопрос 1	0.857	1.000	0.4	1.000	0.857
Вопрос 2	1.000	0.733	0.868	0.75	0.545
Вопрос 3	0.75	0.63	0.000	0.703	0.42
Вопрос 4	0.861	0.861	0.717	0.662	1.000
Вопрос 5	0.725	0.657	0.000	0.5	0.471

Demo-release of system is presented
on the personal webpage of author at www.machinelearning.ru.

more details

continue

Results for a single trainee



Испытуемый: Петров М.Н.

Вопрос теста (вопрос №3):

Как влияет переподгонка на частоту ошибок
дерева принятия решений ?

Полученный ответ:

Именно с переобучение связана увеличение частоты ошибок
дерева принятия решений на контрольной (= тестовой) выборке.

Наиболее близкий вариант правильного ответа:

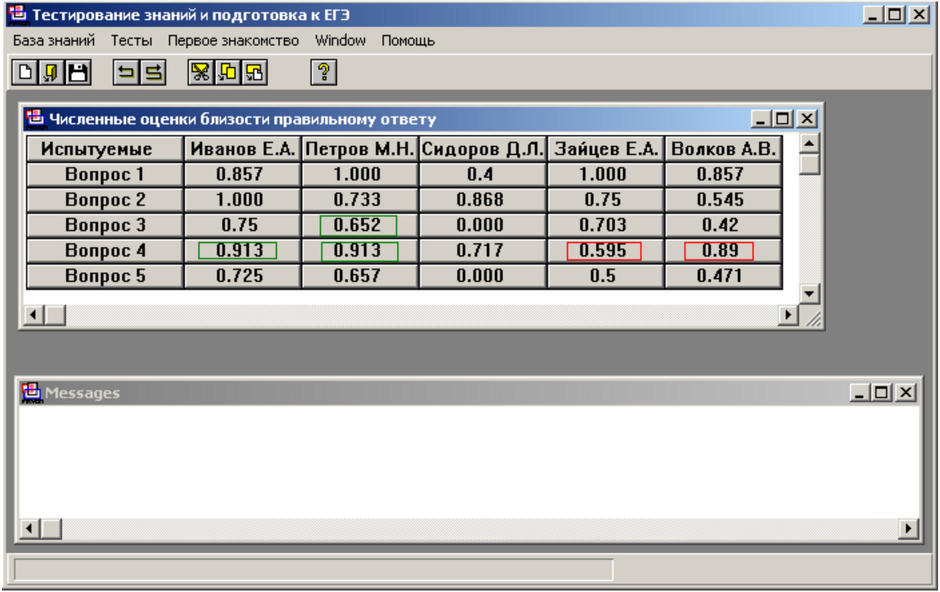
Увеличение частоты ошибок
дерева принятия решений на контрольной выборке
связано с переподгонкой.

Численная оценка близости правильному ответу: 0.63

Оценка за ответ: удовл.

go back continue

Group testing's results after the coordination of knowledge about synonymy concerning the different situations of Russian language's usage


Тестирование знаний и подготовка к ЕГЭ
База знаний
Тесты
Первое знакомство
Window
Помощь
Численные оценки близости правильному ответу
Messages

Испытуемые	Иванов Е.А.	Петров М.Н.	Сидоров Д.Л.	Зайцев Е.А.	Волков А.В.
Вопрос 1	0.857	1.000	0.4	1.000	0.857
Вопрос 2	1.000	0.733	0.868	0.75	0.545
Вопрос 3	0.75	0.652	0.000	0.703	0.42
Вопрос 4	0.913	0.913	0.717	0.595	0.89
Вопрос 5	0.725	0.657	0.000	0.5	0.471

Possible cases of trainee's answer and their interpretation

Case 1. *Incomplete answer* when for all words and their combinations from trainee's answer the prototypes in the most similar «correct» variant were found but *for some words of correct answer no prototypes in the trainee's answer* were found.

Not-nil value of similarity with the object from the correct answer's USNL's formal context will be *only for missed word syntactically submitted to some other word presented both in analyzed and «correct» variant*.

Case 2. *Orthographic errors (which are admissible)* when a word from trainee's answer and a word from the variant of correct answer are the same word's *different forms admissible within the frameworks of the same known lexico-syntactic link*.

Case 3. *«Excess» words* when the analyzed answer has a words which hasn't prototypes in «correct» answer's «variant».

In this case the trainee's *answer* will *not be* considered as *incorrect* only if the *«excess» words don't appear in any* lexico-syntactic *link* presented in system's knowledge base.

What requires the separate research ?

- In offered USNL's conception all kinds of links between main and dependent word were assumed as equally significant.

  To apply such estimations in the tasks of testing of knowledge relatively to concrete subject areas it is necessary to *re-define the affinity of NL-usage's situations from viewpoint of fuzzy logic*.
- Here the *systems analysis of structure of professional knowledge* for the specific area is necessary for the *description of membership functions of fuzzy sets*.
- *Duquenne–Guigues set of implictions* of NL-usage situation's formal context can be a basis of *development of strategies and rules of syntactic analysis*.
- The offered conception of phrase's linear structure's model can be more versatile at applying the *probabilities of coexistence of words* in texts related to given subject area and genre.