Смысловые эталоны и передача знаний в задаче их оценки на основе открытых тестов

Михайлов Д. В., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

11-я Международная конференция
«Распознавание образов и анализ изображений:
новые информационные технологии» (РОАИ-11-2013),

23–28 сентября 2013 г.

г. Самара

Предмет и цель исследований, исследуемая проблема

Предмет исследования

Методы и алгоритмы формирования знаний о синонимии.

Исследуемая проблема

Передача знаний, представляемых текстами на Естественном Языке (ЕЯ), между его носителями (экспертами и обучаемыми).

Основная цель исследований

Разработка и теоретическое обоснование структуры знаний о синонимии, а также методов и алгоритмов их формирования и использования для совокупности задач:

- оценки схожести смыслов текстов предметно-ограниченного ЕЯ;
- автоматизации пополнения и компрессии баз языковых и предметных знаний;
- поиска наиболее рационального плана передачи заданного смысла между носителями заданного ЕЯ;
- согласования знаний, формируемых разными экспертами.

Определение 1.4

Ситуация Языкового Употребления (СЯУ) — описание нового социального опыта (содержания совместных действий) средствами заданного ЕЯ.

Фиксируемый СЯУ S языковой контекст представляется тройкой:

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

где O — множество символов, обозначающих понятия действительности;
Ts — множество форм описания S в некоторой знаковой системе;
$R \subset O^n$, где $n \in 1, \ldots, |O|$.

Пусть $Synt$ — сюръективная функция, определяемая синтаксисом языка.

Тогда для $\forall\, Ts_i \in Ts\ \ \exists\, Tr_i : Ts_i = Synt\,(Tr_i)$, Tr_i — помеченное дерево.

При этом если $O = M \cup V$, $M \cap V \neq \emptyset$, то для $\forall\, o_j \in M$ найдётся $o_k \in V$ такое, что понятию o_j соответствует дочерний узел с пометкой w_j, а понятию o_k — родительский узел с пометкой w_k в дереве Tr_i.

Представим СЯУ посредством формального контекста (ФК):

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

где $\forall g \in G$ — основа слова, синтаксически подчинённого другому слову из некоторой $Ts_i \in Ts$ в составе тройки (1).

Множество признаков M включает подмножества, обозначаемые далее посредством соответствующего нижнего индекса и содержащие:

- указания на основу синтаксически главного слова (индекс 1);
- указания на флексию главного слова (индекс 2);
- связи «основа–флексия» для главного слова (индекс 3);
- сочетания флексий зависимого и главного слова (индекс 4). При этом после флексии главного слова через двоеточие указывается предлог (если такой имеется) для связи главного слова с зависимым;
- указания на флексию зависимого слова (индекс 5).

Задача: построить $I \subseteq G \times M$ анализом буквенного состава и отбором фраз $Ts_i \in Ts$ *минимальной* длины с *наибольшим* числом слов, наиболее *употребимых* в различных фразах из Ts (с учётом синонимов).

Пусть

W_{ij} — последовательность символов слова w_{ij},
Wc_{ij} — неизменная часть (основа),
Wf_{ij} — изменяемая (флективная) часть,
$\odot$ — обозначение для операций типа конкатенации.

Дано:

$Ts = \left\{ Ts_i : Ts_i = \bigodot_j W_{ij} \right\}$ — множество СЭ-фраз, задающих СЯУ.

Найти:

$Pw_i = \left\{ \left(Wc_{ij}, Wf_{ij} \right) : Wc_{ij} \odot Wf_{ij} = W_{ij} \right\}$ для всех $i = 1, \ldots, \left|Ts\right|$.

Будем отождествлять с основой Wc_{ij} и флексией Wf_{ij} принятые в информатике понятия «префикс» и «суффикс».

Ключевые процедуры и функции алгоритма:

- $pref.show\,(w_{ij})$ возвращает текущее значение префикса слова w_{ij};
- $pref.inc\,(w_{ij})$ увеличивает длину префикса слова w_{ij} на 1;
- $prefs$ объединяет словоформы в группы (списки) по сходству префикса, сортируя их при этом по убыванию длины;
- $pref.check\,(Prf)$ для группы словоформ с общим префиксом Prf анализирует частоты (абсолютные) встречаемости букв на разных позициях относительно начала и конца слова. При этом частота ν_p встречаемости первого слева символа и букв в составе Prf всегда максимальна. Относительно конца слова также производится поиск символов общего суффикса (включаются во флективную часть) с частотой встречаемости ν_p. Суммарная длина общих префикса и суффикса при этом должна составлять минимум треть длины слова, а разность длин у пары слов с общим префиксом (независимо от суффикса) всегда меньше половины длины меньшего слова.

описание алгоритма программная реализация далее

Алгоритм выделения основ и флексий для слов в рамках СЯУ

Вход: Ts;

Выход: $Pw = \bigcup_{i=1}^{|Ts|} Pw_i$;

1: $Pw := \varnothing$;
2: **для всех** W_{ij}: $\bigodot\limits_{j} W_{ij} = Ts_i$, где $Ts_i \in Ts$
3: $\quad Wc_{ij} := \{W_{ij}[1]\}$; $Wf_{ij} := \bigodot\limits_{k=2}^{|W_{ij}|} W_{ij}[k]$;
4: **конец для** // инициализации основ и флексий
5: $prefs\,(PrfsTmp)$;
6: **если** $PrfsTmp = \varnothing$ **то**
7: $\quad$ выдать Pw и выйти из алгоритма;
8: **иначе**
9: $\quad$ взять очередной Prf из $PrfsTmp$;
10: $\quad$ **если** $pref.check\,(Prf) =$ true **то**
11: $\quad\quad Pw := Pw \cup \left\{ \left(Prf, Wf_{ij}\,(Prf)\right) \;\middle|\; pref.show\,(w_{ij}) = Prf \right\}$;
12: $\quad\quad PrfsTmp := PrfsTmp \setminus \{Prf\}$;
13: $\quad\quad$ перейти к шагу 6;
14: $\quad$ **иначе**
15: $\quad\quad$ **для всех** $w_{ij} : pref.show\,(w_{ij}) = Prf$
16: $\quad\quad\quad pref.inc\,(w_{ij})$;
17: $\quad\quad$ **конец для**
18: $\quad\quad$ перейти к шагу 5
19: $\quad$ **конец если**
20: **конец если**

Пусть Ts — множество СЭ-фраз, задающих некоторую СЯУ согласно (1),

J — множество индексов неизменных частей слов фраз в составе Ts.

Определение 3.2

Последовательность индексов неизменных частей слов некоторой $Ts_i \in Ts$ назовём моделью её линейной структуры (МЛС), $Ls\,(Ts_i)$.

Пусть $\left\{J_1, J_2\right\}$ — пара последовательностей индексов в $Ls\left(Ts_i\right)$, где $J_1 = \left\{j_1^1, \ldots\ , j_2^1\right\}$, $J_2 = \left\{j_1^2, \ldots\ , j_2^2\right\}$, а парам $\left(j_1^1, j_2^1\right)$ и $\left(j_1^2, j_2^2\right)$ соответствуют синтаксические связи.

Для формирования эталона отбираются $Ts_i \in Ts$, в МЛС которых

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

а суммарная длина всех последовательностей указанного вида для всех связей, выявленных на Ts_i, минимальна.

Пусть $fr\left(w_j\right|$ — частота появления слова w_j во всех $Ts_i \in Ts$.

Тогда наиболее информативные в Ts слова составят кластер $Clust$:

- слово с максимальным значением данной частоты войдёт в $Clust$;
- для $\forall\ \{w_j, w_k\} \subset Clust$ и $\forall\ w_l \notin Clust$ верно то, что

$$\left(\left|\ fr\left(w_j\right) - fr\left(w_k\right)\ \right| < \left|\ fr\left(w_j\right) - fr\left(w_l\right)\ \right|\right) \wedge$$

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

Основу эталона составляют фразы с максимумом слов, вошедших в $Clust$.

При этом для слов из $Clust$ учитываются различные порядки следования их во фразе и возможные синонимы.

Пусть LS — множество моделей линейных структур ЕЯ-фраз из Ts на J.

Лемма 5.1

Пара индексов $\{j_1, j_2\} \subset J$ соответствует словам-синонимам и могут быть заменены одним индексом из $(\mathbb{N} \backslash J)$, если $\exists\ \{Ls\left(Ts_1\right), Ls\left(Ts_2\right)\} \subseteq Ls$:

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

где $J_1 \subset J$, $J_2 \subset J$, а $\odot$ есть операция типа конкатенации над J.

Пусть J_{Cl} — множество индексов слов кластера наиболее информативных относительно СЯУ, задаваемой множеством СЭ-фраз Ts;

$frq\left((j,k),LS\right)$ — частота появления пары (j,k) в моделях из множества LS с учётом того, что $(j,k) \Leftrightarrow (k,j)$.

Тогда эталон СЯУ определяют фразы, МЛС которых входят в множество

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

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

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

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

а построение признакового множества ФК (2) эталона СЯУ требует:

- найти пары индексов $(j,k) : frq\left((j,k),LS\right) > 1$, отвечающие условию (3), для всех МЛС множества LC;
- каждой найденной (j,k) задать направление синтаксической связи;
- из $\forall\ LS_i \subset LC$ исключить МЛС с индексами, не входящими ни в одну из найденных связей.

Поиск $Dir\,(j,k)$, $Dir \in \{\leftarrow, \rightarrow\}$, идёт в три этапа:

- проверка на признак ложной связи;
- попытка отождествления с ранее выделенными связями;
- при отсутствии ассоциации с известными связями — опрос эксперта.

Пусть $St\,(j)$, $St\,(k)$ и $St\,(l)$ — основы слов, отвечающие индексам j, k и l.

В заданной СЯУ для пары (j,k) *связь ложная, если* $j,k,l \in Ls\,(Ts_i)$ *в некоторой* $Ts_i \in Ts$ *и имеется СЯУ, где связь* $St\,(j)$ *и* $St\,(k)$ *ложная, но есть связь либо между* $St\,(j)$ *и* $St\,(l)$*, либо между* $St\,(k)$ *и* $St\,(l)$*.*

Пусть Lnk — множество ранее найденных связей, каждая представлена:

- идентификационным номером СЯУ (Id);
- основой главного слова (St_1);
- основой зависимого слова (St_2);
- списком пар «флексия главного слова–флексия зависимого» (FCm).

Паре (j,k) соответствует связь $((j,k)\,,\rightarrow)$, если для некоторой СЯУ $\exists\,(Id, St_1, St_2, FCm) \in Lnk$:

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

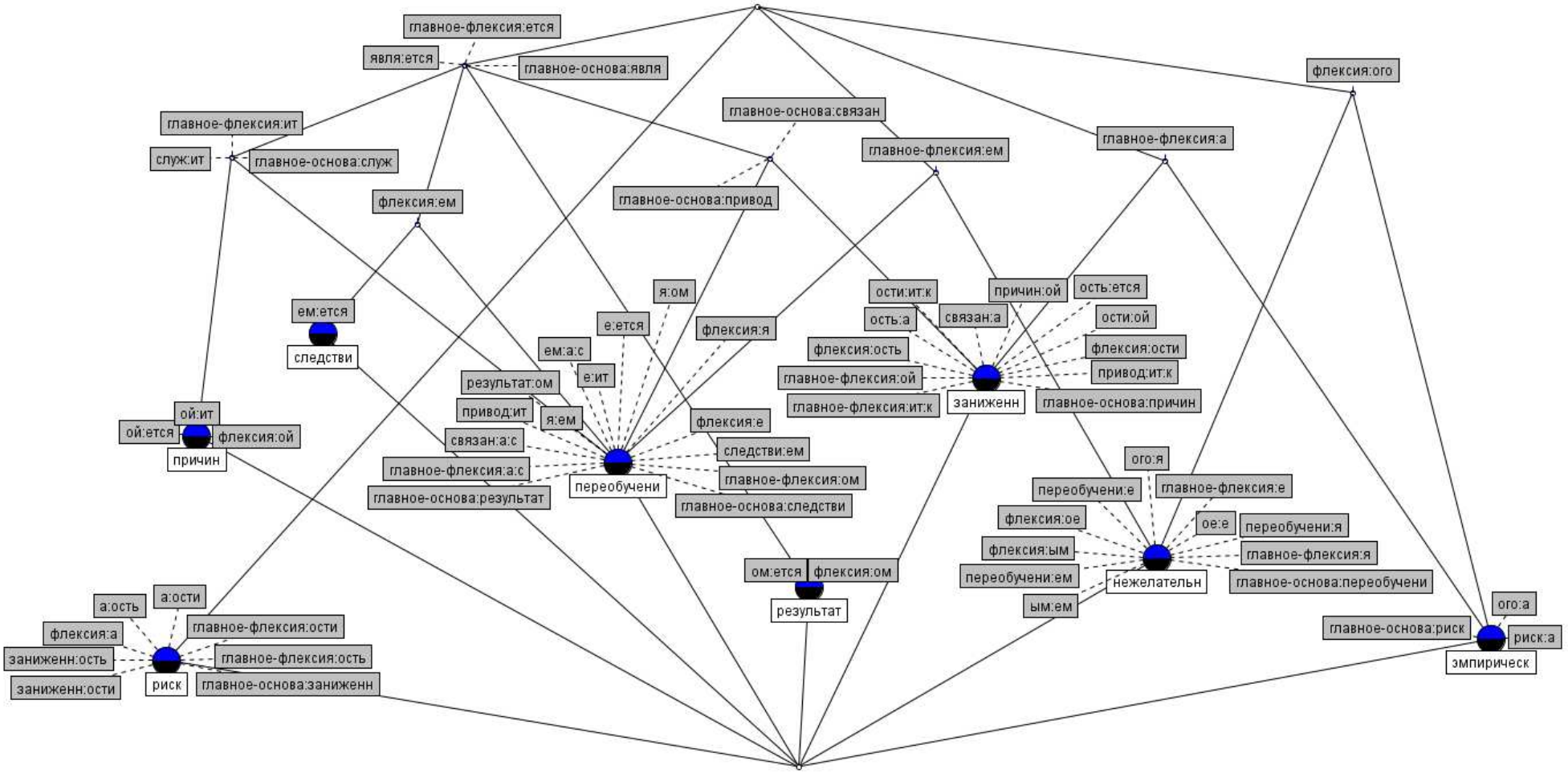
Пример: исходное множество семантически эквивалентных фраз

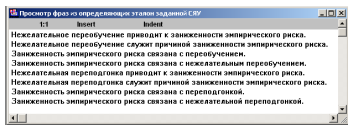
Синонимичные перифразы

27:89 Insert Indent Modified

"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."

Результат: ЕЯ-фразы смыслового эталона и его формальный контекст

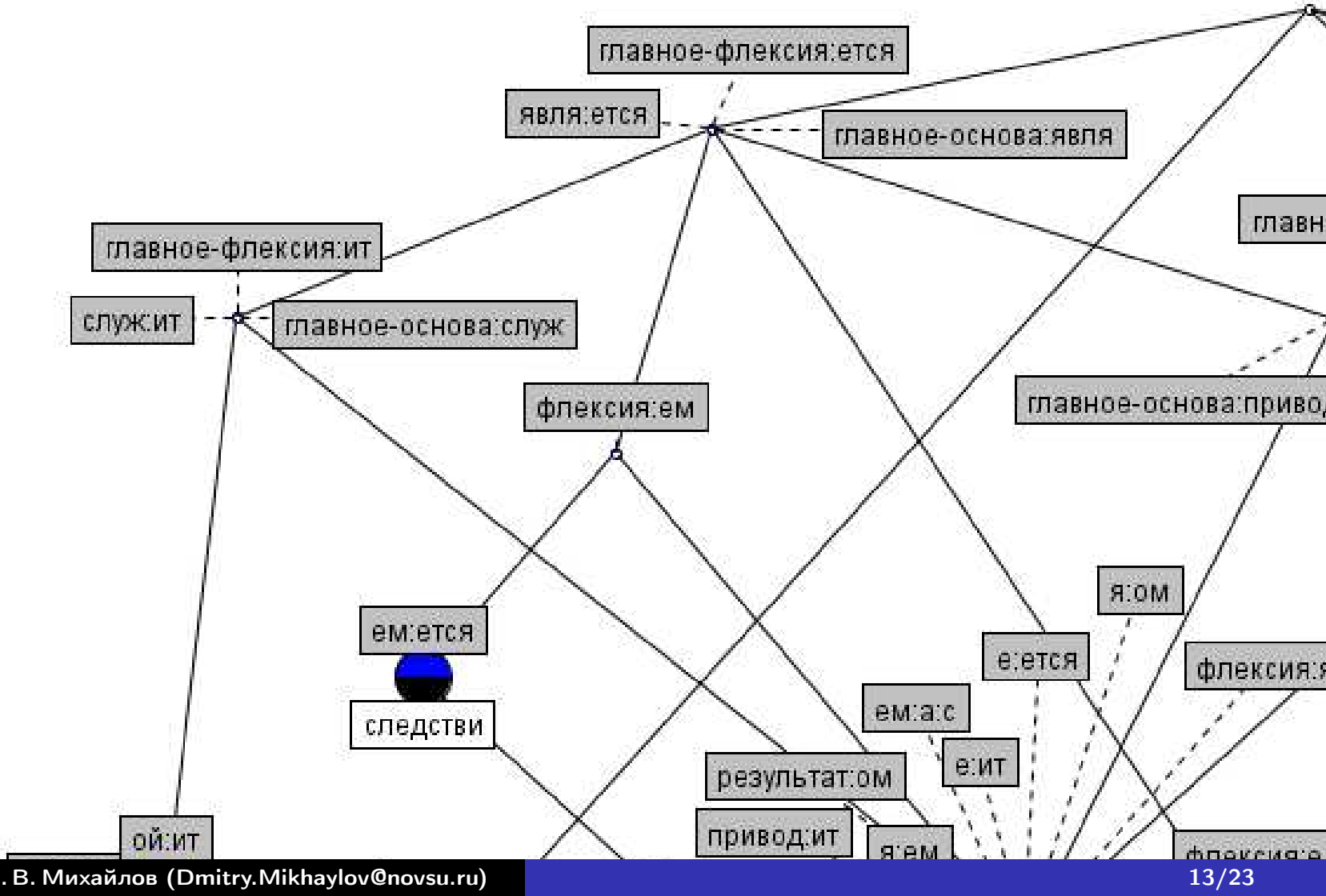




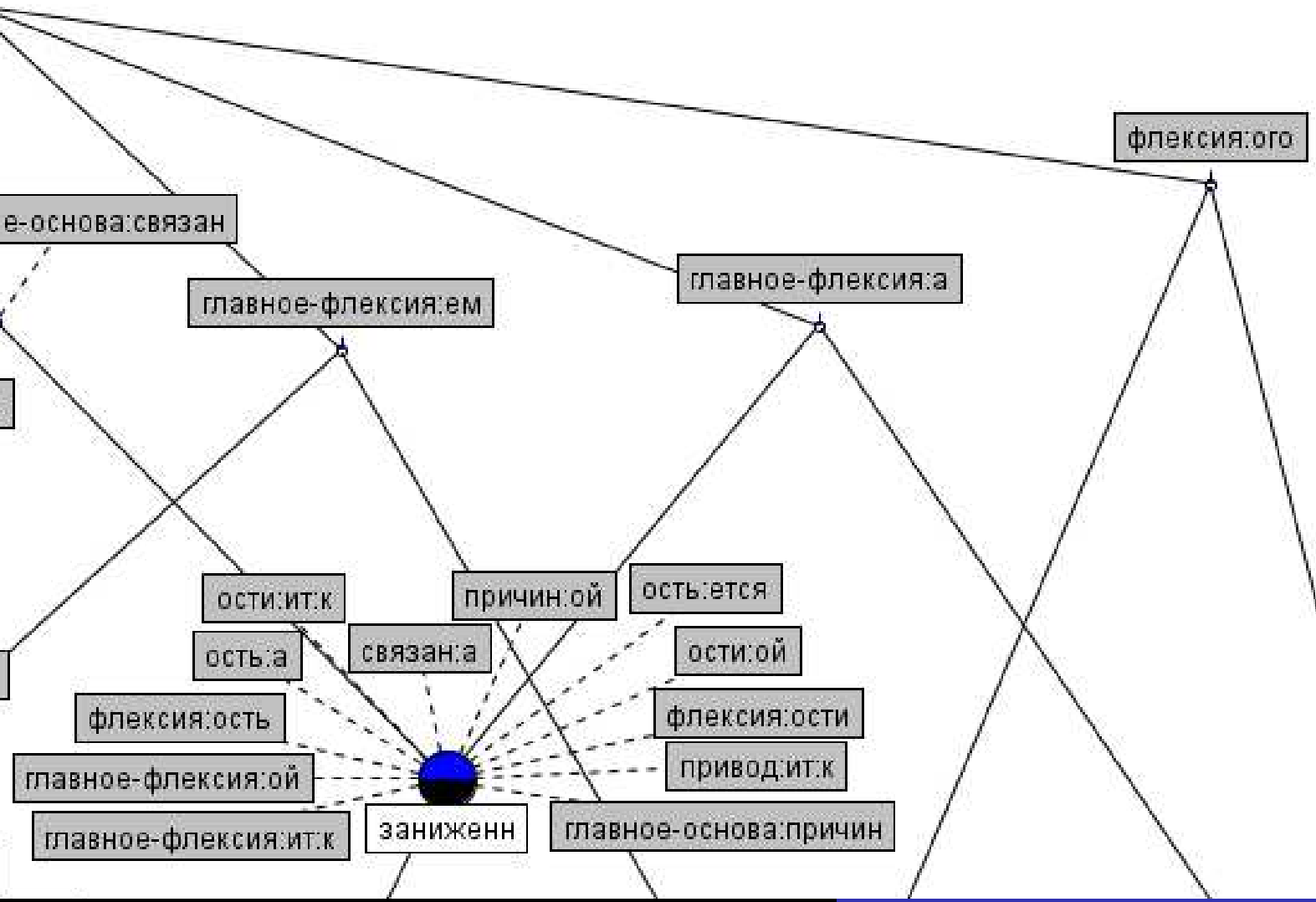
назад

далее

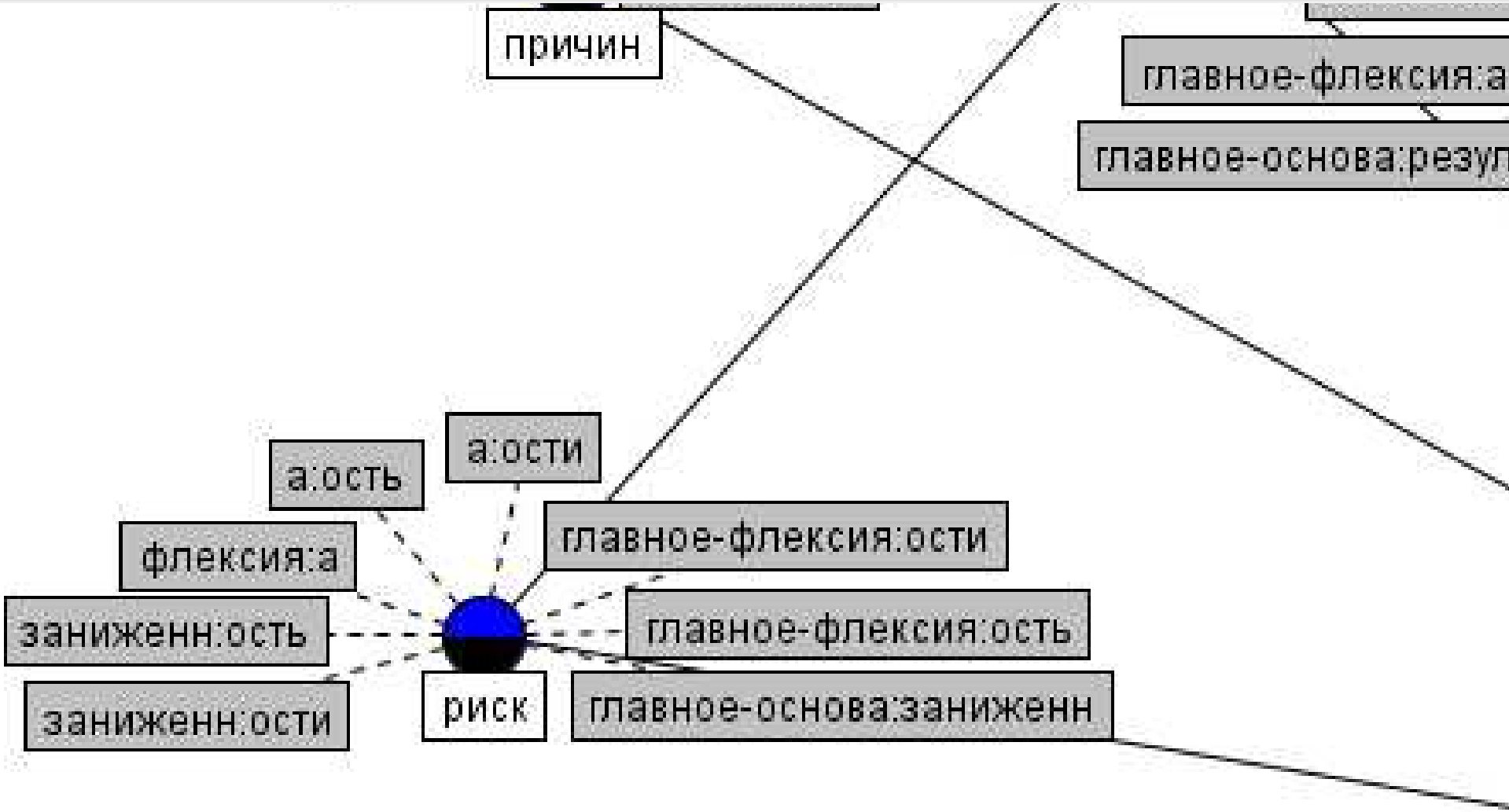
Результат: ЕЯ-фразы смыслового эталона и его формальный контекст



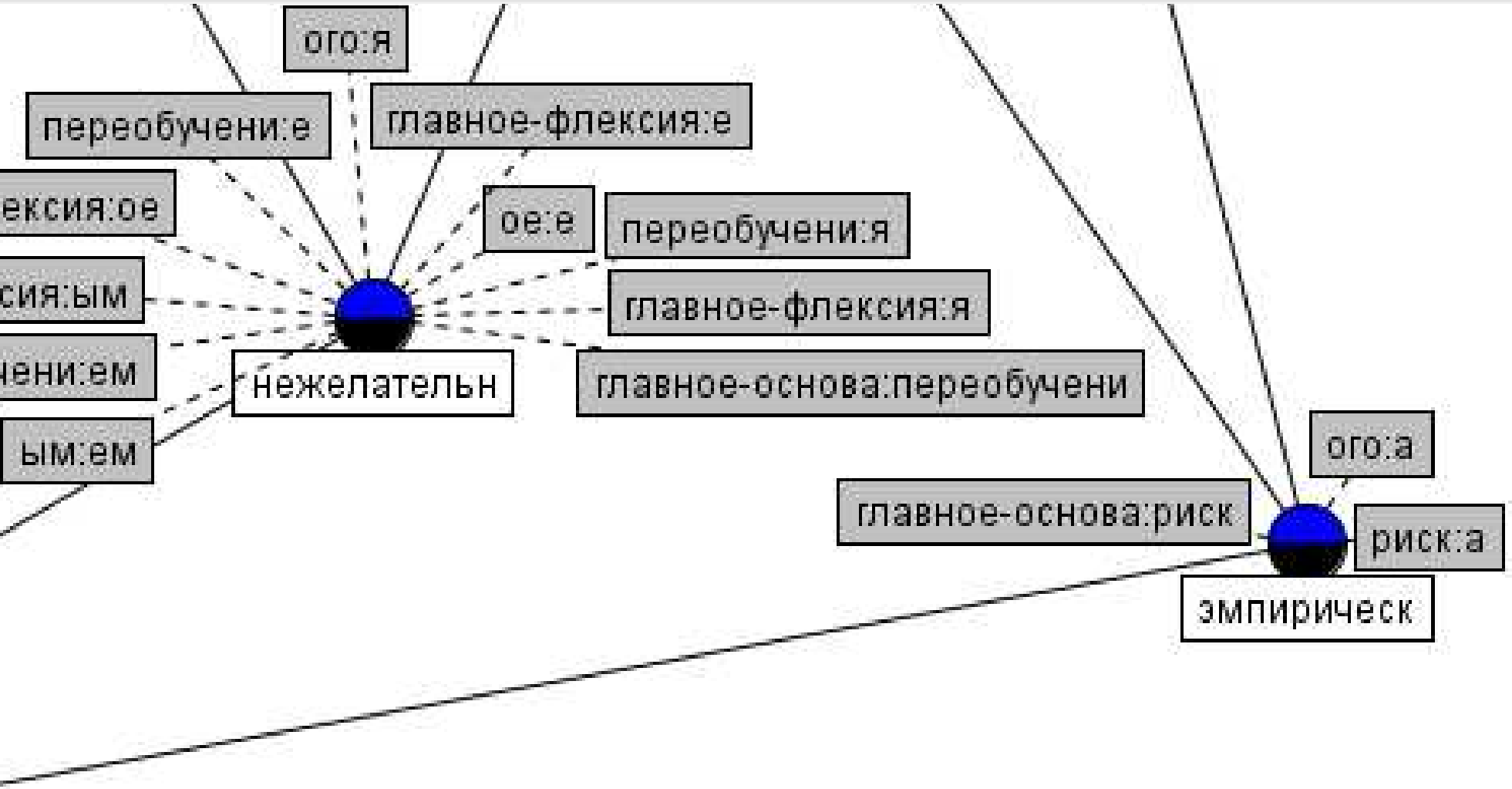
Результат: ЕЯ-фразы смыслового эталона и его формальный контекст



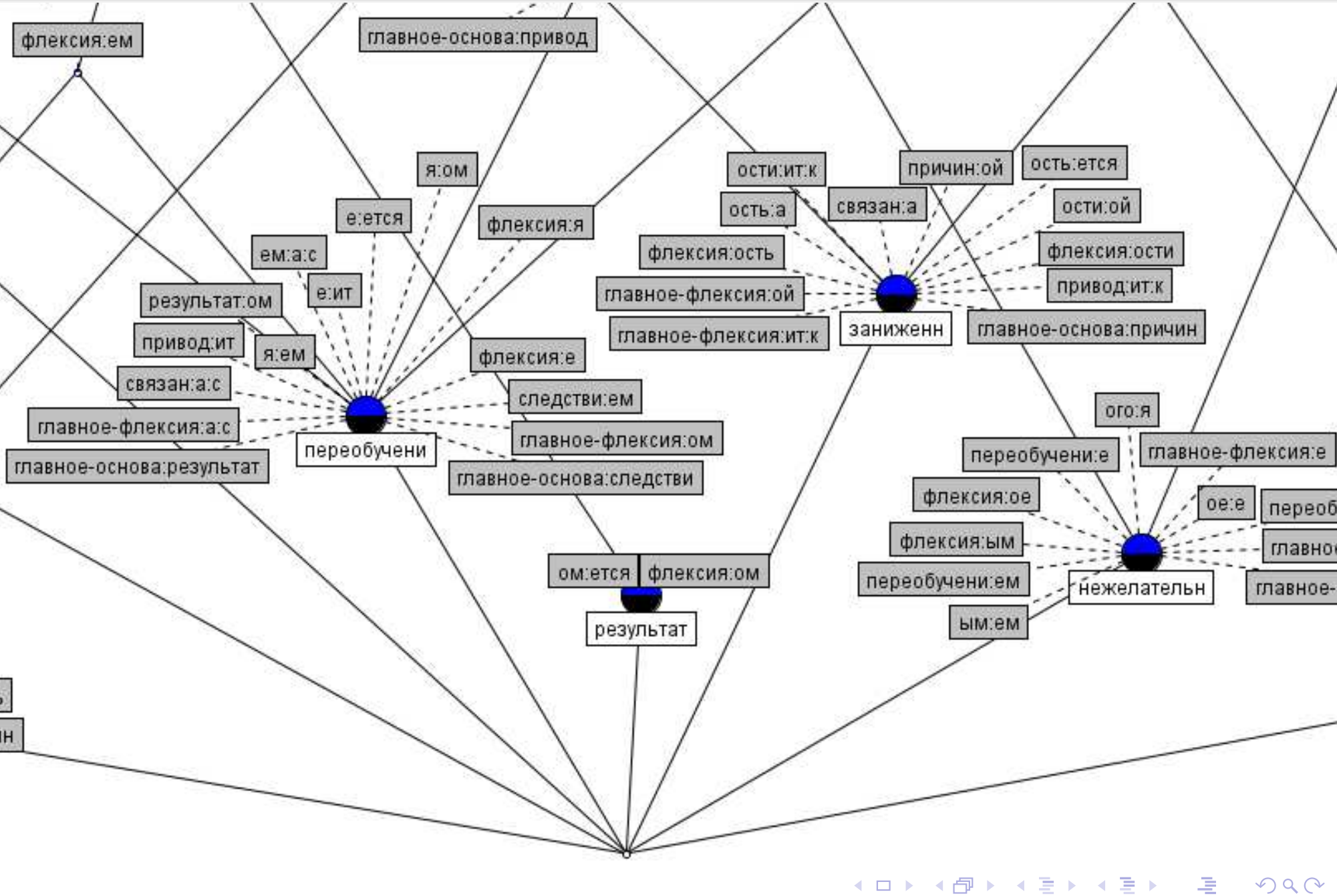
Результат: ЕЯ-фразы смыслового эталона и его формальный контекст



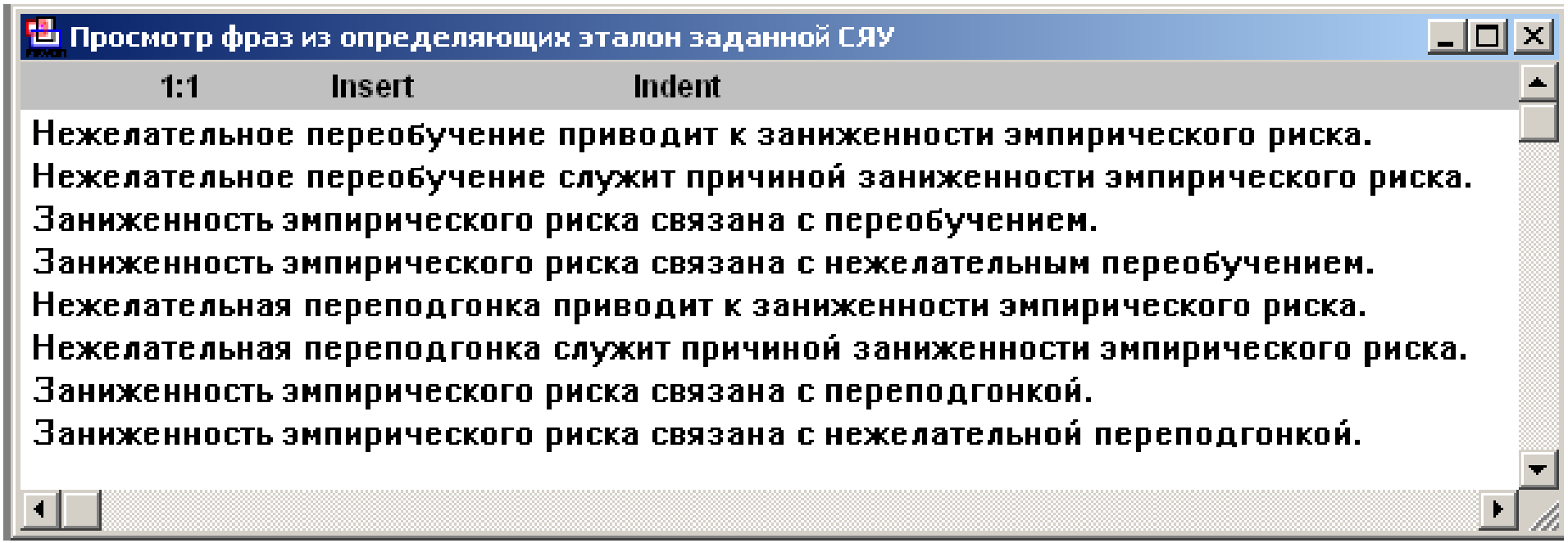
Результат: ЕЯ-фразы смыслового эталона и его формальный контекст



Результат: ЕЯ-фразы смыслового эталона и его формальный контекст



Результат: ЕЯ-фразы смыслового эталона и его формальный контекст



Порядковый номер СЯУ, i	1	2	3	4	5	6
Число фраз, задающих СЯУ	56	28	29	30	6	10
из них представляют эталон	8	9	7	9	1	2
Исходное число объектов СЯУ	18	17	15	13	12	14
Исходное число признаков СЯУ	177	186	173	162	94	81
Число объектов эталона	9	12	12	11	8	12
Число признаков эталона	82	90	80	69	35	53

i	Ситуация языкового употребления
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

пример | согласование знаний о синонимии | оценка требуемого объёма памяти

Оценка требуемого объёма памяти для хранения ЕЯ-фразы

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

Здесь:

i — порядковый номер СЯУ;
n — максимальное число слов во фразе;
$vol\,(n) = n!$ есть традиционно используемая оценка;
vol_1 и vol_2 — оценки с применением метода и алгоритмов выделения эталона СЯУ.

При этом:

$vol_1\,(n) = l_1 \cdot n$ есть оценка сверху, l_1 — число фраз, определяющих СЯУ;
$vol_2\,(n) = l_2 \cdot n$ есть оценка снизу, l_2 — число фраз, определяющих эталон СЯУ.

Рассмотрим представление тезауруса формальным контекстом

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

где Gth состоит из символьных пометок отдельных СЯУ;

Mth содержит признаки ФК вида (2) всех $gth \in Gth$.

Кроме того, в составе Mth выделяются подмножества:

- M_6 — указаний на объекты формальных контекстов вида (2) отдельных $gth \in Gth$;
- M_7 — множество связей «основа–флексия» для синтаксически зависимого слова;
- M_8 — множество сочетаний основ зависимого и главного слова.

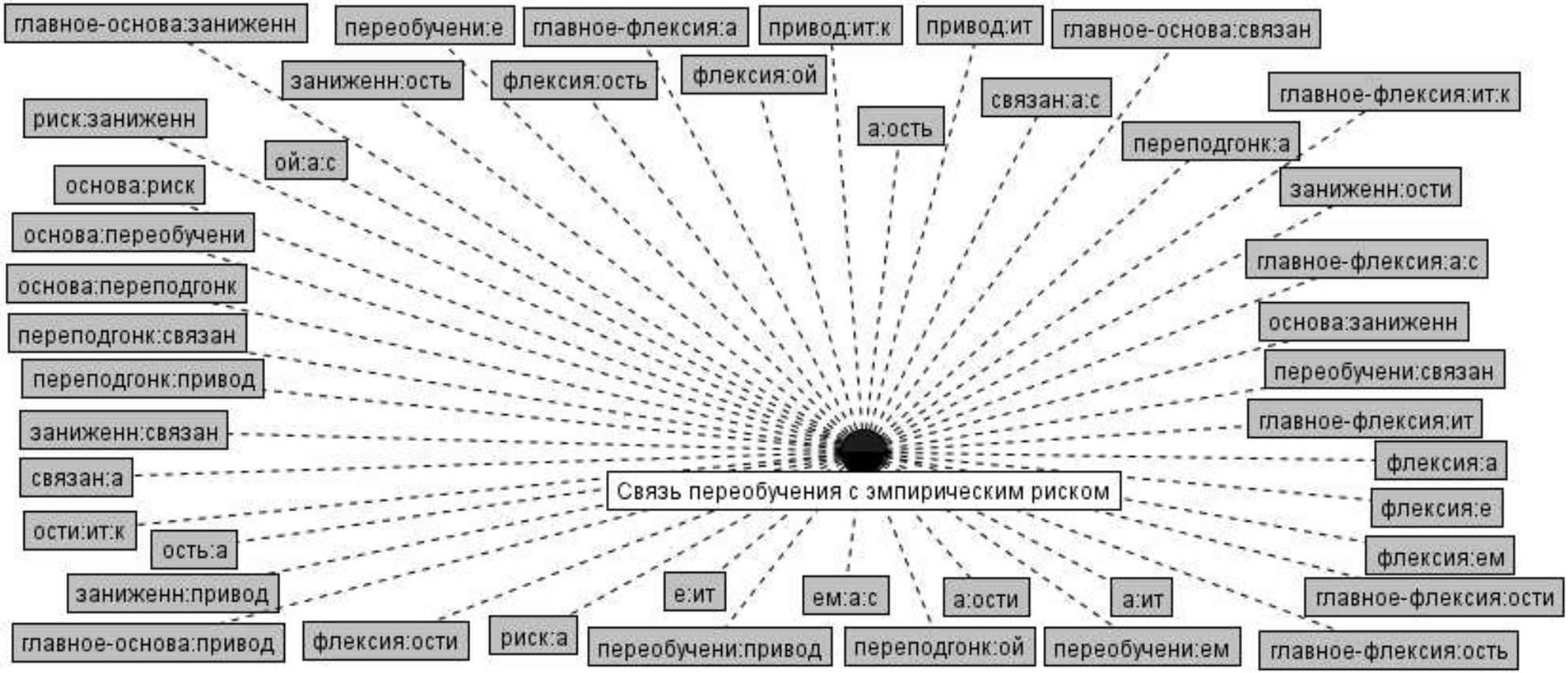
По аналогии с ФК (2) отдельной СЯУ имеем $Ith \subseteq Gth \times Mth$.

пример представления СЯУ в формальном контексте тезауруса

При этом численная оценка схожести СЯУ определяется числом признаков, которые разделяются объектами сравниваемых ситуаций относительно формального контекста тезауруса.

согласование знаний о синонимии относительно разных СЯУ

Пример представления СЯУ в формальном контексте тезауруса



назад к определению ФК тезауруса

далее

Пусть

St — основа слова (его неизменная часть);
Fl — его флексия;
S_1 и S_2 — некоторые СЯУ.

Предположим, что некоторое слово Wrd относительно S_1 представляется как $St_1 \odot Fl_1$, а относительно S_2 — как $St_2 \odot Fl_2$, причём $St_1 = St_2 \odot Sf$, где Sf содержит минимум один символ, а $\odot$ есть операция конкатенации символьных строк.

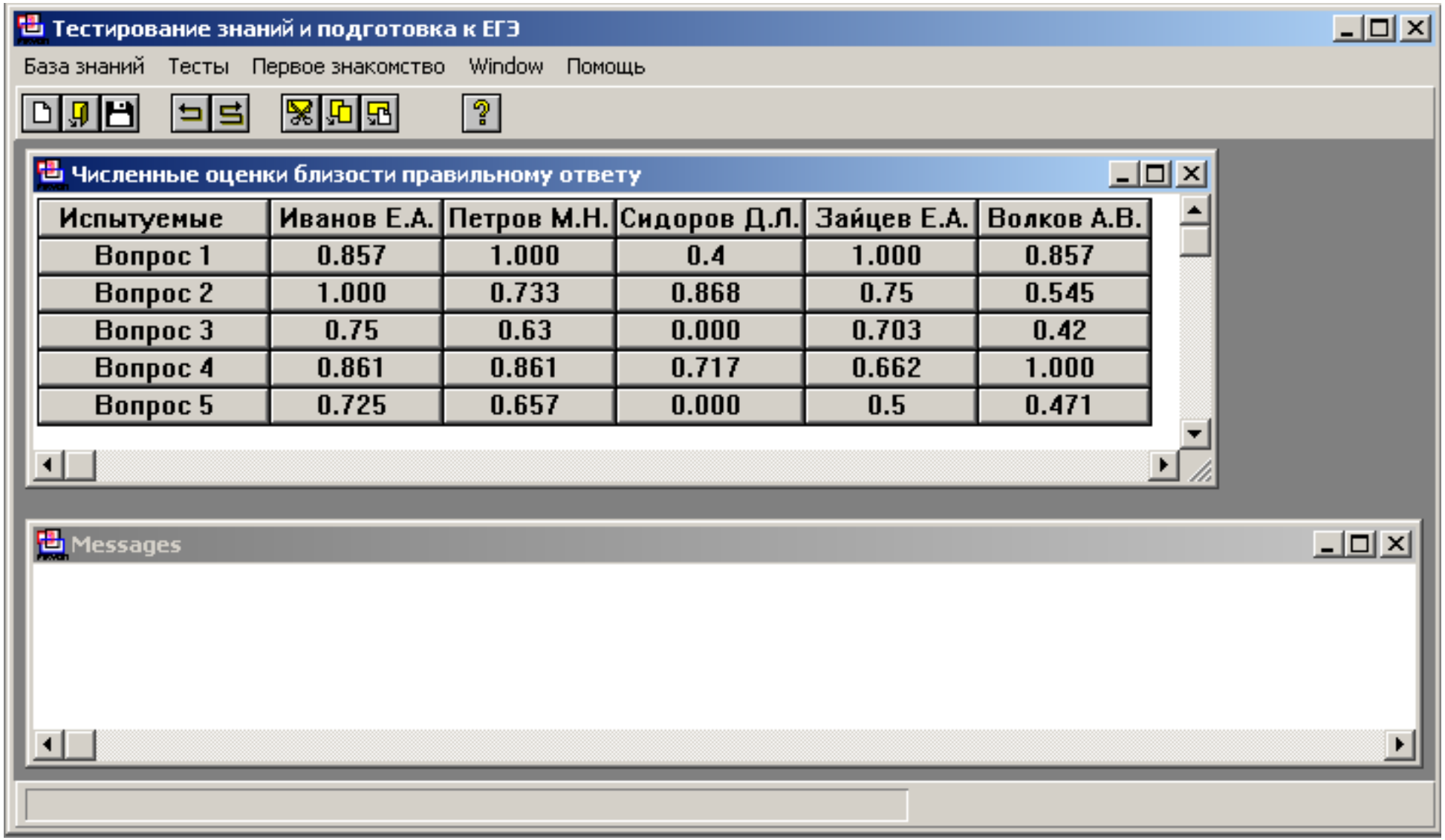
Тогда относительно S_1 основа St_1 будет заменена на St_2, флексия Fl_1 заменяется на $Fl_3 = Sf \odot Fl_2$, но только в том случае, если частоты встречаемости флексий Fl_3 и Fl_2 во всех лексико-синтаксических связях, представляемых формальным контекстом вида (5) для заданной предметной области, не уменьшаются при выполнении указанных замен.

Пример.

СЯУ №3, $St_1 =$ «является», $Fl_1 =$ «»,
СЯУ №1, $St_2 =$ «явля», $Fl_2 =$ «ется», $Sf =$ «ется»,
а относительно СЯУ №3 производится замена: Fl_1 — на $Fl_3 =$ «ется».

назад далее

Практическое приложение: система тестирования знаний

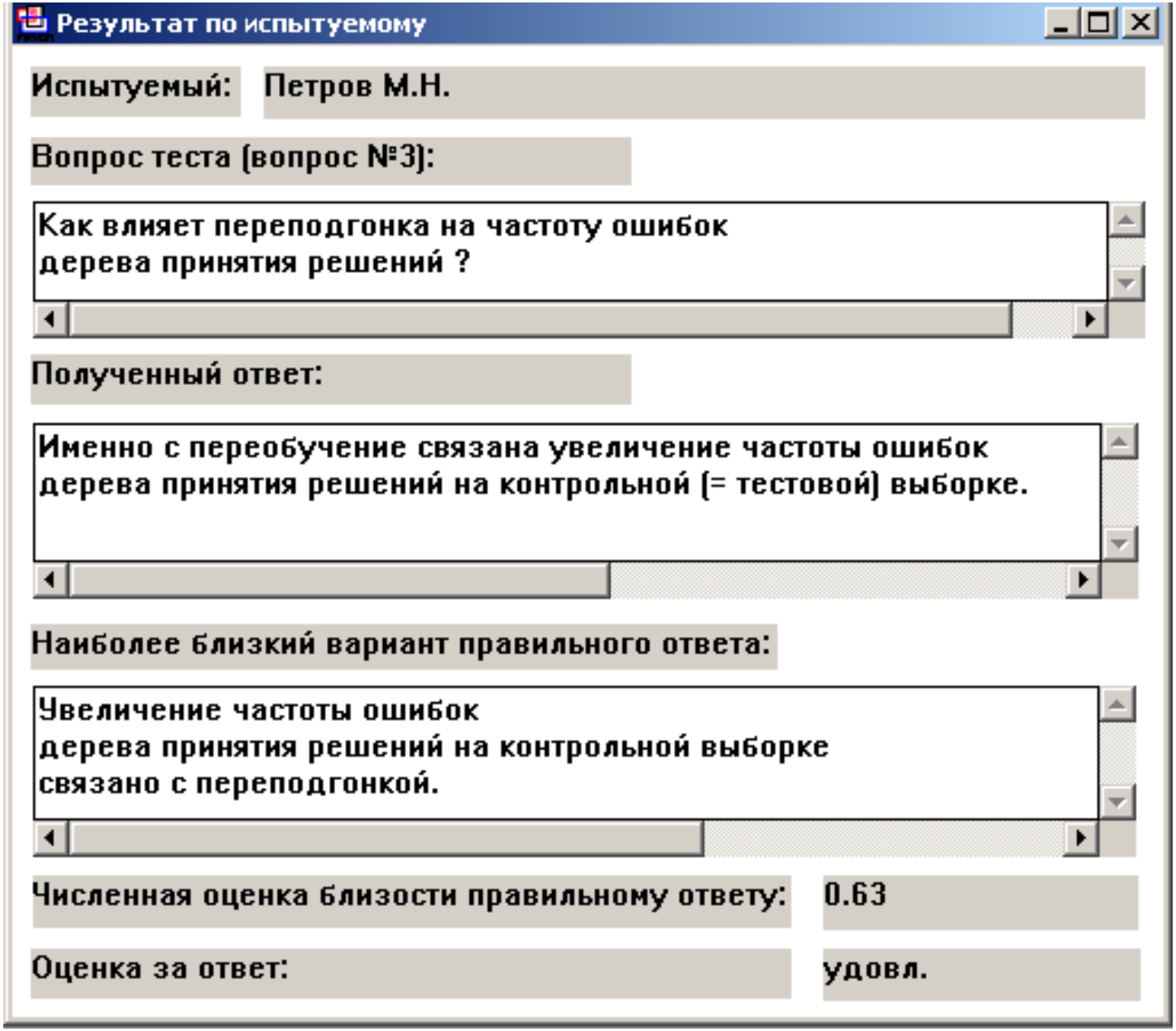


Испытуемые	Иванов Е.А.	Петров М.Н.	Сидоров Д.Л.	Зайцев Е.А.	Волков А.В.
Вопрос 1	0.857	1.000	0.4	1.000	0.857
Вопрос 2	1.000	0.733	0.868	0.75	0.545
Вопрос 3	0.75	0.63	0.000	0.703	0.42
Вопрос 4	0.861	0.861	0.717	0.662	1.000
Вопрос 5	0.725	0.657	0.000	0.5	0.471

Демо-версия системы представлена
на www.machinelearning.ru, на персональной странице автора.

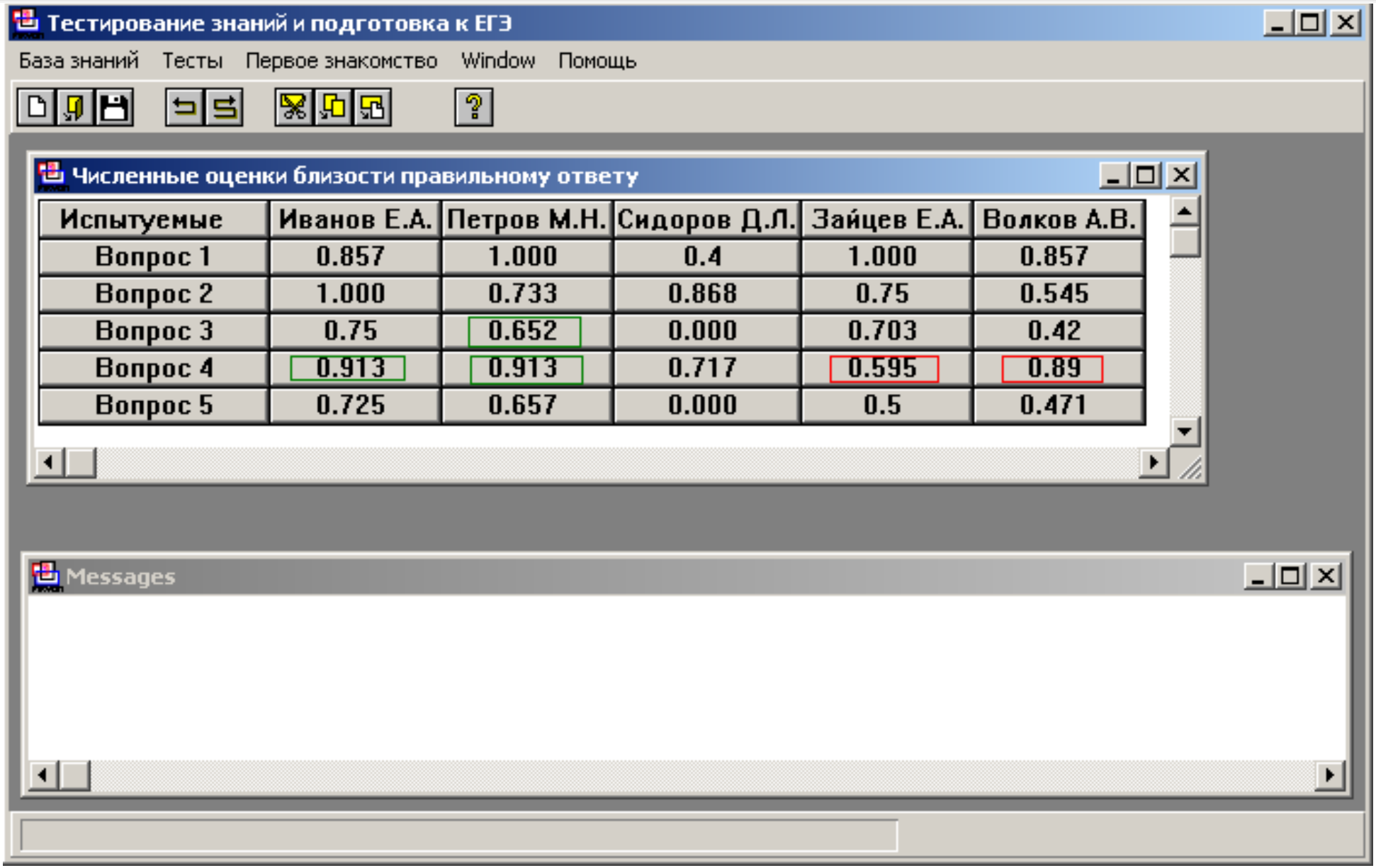
подробнее
далее

Результаты по отдельному испытуемому



назад далее

Результаты группового тестирования после автоматического согласования знаний о синонимии относительно различных СЯУ



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

Возможные случаи ответа и их интерпретация

Случай 1. *Неполный ответ* — все слова и словосочетания из ответа испытуемого нашли прообразы в наиболее близком варианте правильного ответа, но *часть слов правильного ответа не нашла прообразов в ответе испытуемого*.

Ненулевое значение оценки схожести с объектом из ФК СЯУ правильного ответа будет *для упущенного слова, которое в «правильном» варианте синтаксически зависимо относительно некоторого другого слова, присутствующего в анализируемом ответе*.

Случай 2. *Орфографические ошибки (из допустимых)* — слово из ответа испытуемого и слово правильного ответа есть *формы одного и того же слова в рамках некоторой лексико-синтаксической связи из известных системе*.

Случай 3. *«Лишние» слова* — в анализируемом ответе имеются слова, не нашедшие прообразов в правильном «варианте».

Ответ засчитывается как *неверный*, если «лишнее» слова фигурируют в *известных* системе лексико-синтаксических *связях*.

Что требует отдельного исследования ?

- В предложенной концепции СЯУ все связи между главным и зависимым словом предполагались одинаково значимыми. Для их применения в задачах оценки знаний по отраслям *схожесть СЯУ* необходимо *переформулировать с позиций нечёткой логики*.
- Для описания функций принадлежности нечётким множествам необходим *системный анализ структуры профессиональных знаний* в конкретной области.
- *Базис импликаций* формального контекста ситуации языкового употребления может послужить основой *разработки стратегий и правил синтаксического анализа*.
- Концепцию модели линейной структуры предложения можно сделать более гибкой, введя *вероятности совместной встречаемости слов* относительно текстов заданной предметной области и жанра.