УДК 004.93

СМЫСЛОВЫЕ ЭТАЛОНЫ И ПЕРЕДАЧА ЗНАНИЙ В ЗАДАЧЕ ИХ ОЦЕНКИ НА ОСНОВЕ ТЕСТОВ ОТКРЫТОГО ТИПА

© Д. В. Михайлов, Г. М. Емельянов

Новгородский государственный университет имени Ярослава Мудрого

e-mail: *Dmitry.Mikhaylov@novsu.ru*

Semantic etalons and knowledge transfer in the problem of their estimation based on the open type tests.

Mikhaylov, D. V. and Yemelyanov, G. M.

Abstract. The article is devoted to the problem of knowledge transfer between experts and learners in machine learning and knowledge control systems that store information in the form of Natural Language (NL) text units. The purpose of this study is to minimize losses of useful information when forming a knowledge-based system that works with a textual description of the subject area test facts. The solution of this problem is suggested within the framework of the theory of Formal Concept Analysis (FCA) based on the concepts of Situations of Language Use (SLU) as a unit of formal description of the semantics. In the article, in particular, coordination of knowledge generated by experts as well as search for the most efficient transfer method of information between the two groups of NL carriers (experts and trainees) are considered to be very important tasks. In accordance to the model proposed be the authors of this article, the use of SLU etalons as units of the thesaurus and concordance of the etalons allows to reduce the size of the text data. The authors describe a system that performs a search of the SE-form closest to a user response, which defines SLU of the correct answer. Next comes the analysis of the word discrepancies, searching consistencies among mismatched responses' parts being compared as a part of correct answer's etalon and evaluation according to found synonyms. For use of such assessments in evaluating the expert knowledge from different industries, it was necessary to reformulate the definition of SLU similarity using fuzzy logic. System analysis of the professional knowledge structure in a particular area is used for a description of the membership functions.

Введение и постановка задачи

На сегодняшний день интеллектуализация автоматизированного контроля знаний есть одна из приоритетных задач развития информационных технологий в образовании. Немаловажную роль при этом играют открытые тесты (тестовые задания открытой формы), которые предполагают ответ обучаемого в виде одного или нескольких предложений на естественном языке. При этом для интерпретации результатов теста открытой формы необходимо учитывать различные эквивалентные по смыслу формы описания одного и того же факта действительности разными экспертами на одном и том же ЕЯ. Ставится задача поиска наиболее рационального

плана передачи смысла между двумя группами носителей ЕЯ: эксперты и обучаемые, а также согласования знаний, формируемых экспертами. Сам же смысл в итоге должен быть отражён в максимально компактном объёме текстовых данных. Именно на основе этих данных оценивается близость ответа испытуемого правильному ответу, который формулируется экспертом. Настоящая работа посвящена решению указанной задачи на основе концепции смыслового эталона СЯУ, предложенной авторами.

1. Ситуация языкового употребления как единица знаний

Пусть Ts есть множество равных по смыслу (Семантически Эквивалентных [3], СЭ) ЕЯ-фраз, задающих различные формы описания некоторого факта предметной области теста и определяющих СЯУ. Представим СЯУ тройкой

$$K = (G, M, I) \tag{1}$$

именуемой в теории АФП [1] формальным контекстом (ФК). Множество его объектов G составляют основы слов, синтаксически подчинённых другим словам из СЭ-фраз в составе Ts. Множество признаков M включает подмножества, обозначаемые далее посредством соответствующего нижнего индекса:

– указаний на основу синтаксически главного слова (индекс 1);
– указаний на флексию главного слова (индекс 2);
– связей «основа-флексия» для синтаксически главного слова (индекс 3);
– сочетаний флексий зависимого и главного слова (индекс 4);
– указаний на флексию зависимого слова (индекс 5),

При этом пара (A, B) есть Формальное Понятие (ФП, [1]) с объёмом A и содержанием B, если $A \subseteq G$, $B \subseteq M$ и $\exists(A', B')$: $A = \{m \in M \mid \forall g \in A : g\ Im\}$, $B = \{g \in G \mid \forall m \in B : g\ Im\}$, причём $A = B'$ и $B = A'$. Отношение $I \subseteq G \times M$ формируется анализом буквенной структуры фраз $Ts_i \in Ts$ путём отбора тех из них, которые отвечают требованию компактного выражения смысла.

В задачах классификации гипотеза компактности есть предположение о том, что схожие объекты гораздо чаще лежат в одном классе, чем в различных [2]. Если представить смысл множества фраз $\{Ts_i : Ts_i \in Ts\}$ как набор функций, которые связывают обозначаемые словами понятия, то каждая такая функция:

– определена на множестве буквенных цепочек, образующих основы слов фраз $Ts_i \in Ts$;
– имеет множество значений, однозначно определяемое некоторым $I' \subset I$,

а требование компактности выражения смысла означает отбор фраз $Ts_i \in Ts$ минимальной длины с наибольшим числом слов, наиболее употребимых в различных фразах из Ts (с учётом синонимов). Именно такие фразы должны составить основу формирования модели (1) как единицы предметных знаний, представляемых текстами заданного ЕЯ. Относительно этих единиц в конечном итоге и оценивается близость ответа испытуемого правильному ответу.

2. Смысловой эталон и его формирование

Рассмотрим $Ts_i \in Ts$ с точки зрения составляющих её символов. У каждой Ts_i выделяется неизменная часть, общая для всех $Ts_i \in Ts$, и флективная часть. Обозначим далее множество индексов для неизменных частей (основ) слов фраз из Ts как J. Последовательность таких индексов для некоторой $Ts_i \in Ts$ назовём Моделью её Линейной Структуры (МЛС), $Ls(Ts_i)$.

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

Лемма 1. *Пара индексов $\{j_1, j_2\} \subset J$ соответствует словам-синонимам, если $\exists\{Ls(Ts_1), Ls(Ts_2)\} \subseteq LS : Ls(Ts_1) = J_1 \bullet \{j_1\} \bullet J_2$ и $Ls(Ts_2) = J_1 \bullet \{j_2\} \bullet J_2$, где $J_1 \subset J$, $J_2 \subset J$, а "$\bullet$" — операция типа конкатенации над множеством J.*

Пусть PJ — множество пар, отвечающих условию Леммы 1. Заменим индексы, вошедшие в пары из PJ, на некоторые $j \in (N \setminus J)$ во всех моделях из LS. Обозначим далее преобразованное множество LS как LS'.

Утверждение 1. *Пусть $\{J_1, J_2\}$ — пара последовательностей индексов в $Ls(Ts_i)$, где $J_1 = \{j_1^1, \ldots, j_2^1\}, J_2 = \{j_1^2, \ldots, j_2^2\}$, а каждой из пар (j_1^1, j_2^1) и (j_1^2, j_2^2) отвечает синтаксическая связь. Тогда смысловой эталон СЯУ определяют те $Ts_i \in Ts$, в моделях линейных структур которых*

$$(J_1 \subset J_2) \vee (J_2 \subset J_1) \vee (|J_1 \cap J_2| = 1) \vee (J_1 \cap J_2 = \varnothing) = true, \tag{2}$$

а сумма длин всех последовательностей указанного вида для всех синтаксических связей на Ts_i должна быть минимальной.

Утверждение 2. *Пусть $freq(w_j)$ — частота появления слова w_j (независимо от его формы) во всех $Ts_i \in Ts$. При этом основу эталона будут составлять фразы с максимумом слов, вошедших в особый кластер $Clust$:*

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

$$(|freq(w_j) - freq(w_k)| < |freq(w_j) - freq(w_l)|) \wedge \\ \wedge(|freq(w_j) - freq(w_k)| < |freq(w_k) - freq(w_l)|) = true.$$

Замечание. При формировании множества $Clust$ учитываются возможные синонимы анализируемых слов (согласно Лемме 1), поэтому для любого w_j значение $freq(w_j)$ оценивают относительно множества LS'.

Пусть $J_{C_l} \subset J$ — множество индексов слов, вошедших в Clust. Рассмотрим множество

$$LC = \bigcup_i LS_i : LS_i \subset LS, \quad \exists Ts_i, \quad Ts_j \in Ts :$$
$$Ls(Ts_i) \in LS_i, |Ls(Ts_i) \cap J_{Cl}| \to max,$$
$$((Ls(Ts_j) \in LS_i) \wedge (Ts_j \neq Ts_i) \to |Ls(Ts_i) \cap J_{Cl}| \subset Ls(Ts_j).$$

Как следует из Утверждения 2, смысловой эталон определяют те фразы, модели линейных структур которых принадлежат LC.

Пусть $frq((j,k), LS)$ есть частота появления пары (j,k) в моделях из множества LS с учётом того, что $(j,k) \Leftrightarrow (k,j)$. Для построения признакового множества формального контекста вида (1) эталона СЯУ требуется найти индексные пары, отвечающие условию (2), и каждой паре нужно задать направление соответствующей синтаксической связи.

Алгоритм 1. Формирование связей для эталона СЯУ.
Вход: LS;
Выход: $R_J = \{((j,k), Dir) : Dir \in \{\leftarrow, \rightarrow\}\}$;
Начало

1. $R_J := \varnothing$;
2. сформировать LC на основе LS;
3. **для всех** $Ls(Ts_i) \in LC$
4. $\quad P_i := \{(j,k) : j,k \in Ls(Ts_i), j \neq k\}$;
5. $P := *_i P_i$ с учётом $(j,k) \Leftrightarrow (k,j)$;
6. $P' := \{(j,k) \in P : frq((j,k), LC) > 1\}$;
7. **для всех** $(j,k) \in P'$
8. $\quad$ **если** найдено $Dir(j,k)$ **то**
9. $\quad\quad R_J := R_J \cup \{(j,k), Dir\}$;

Конец {Алгоритм 1}.

Для каждой пары (j,k), выделенной на *Шаге 6 Алгоритма 1*, поиск $Dir(j,k)$ идёт в три этапа. На первом проверяется, является ли связь, соответствующая паре, ложной.

Определение 1. Пусть $\{j,k,l\} \subset J$, а $St(j)$, $St(k)$ и $St(l)$ есть основы слов, отвечающие индексам j, k и l. Связь, ассоциируемая с парой (j,k), идентифицируется как ложная относительно рассматриваемой СЯУ при одновременном выполнении двух условий:

1. $\exists Ts_i \in Ts : j,k,l \in Ls(Ts_i)$.
2. В рассматриваемой предметной области существует СЯУ, где связь между $St(j)$ и $St(k)$ идентифицирована как ложная, но существует связь либо между $St(j)$ и $St(l)$, либо между $St(k)$ и $St(l)$.

Замечание. Начальные знания системы об истинных и ложных связях формируются в режиме интервью с экспертом. При этом совокупным знаниям по отдельной СЯУ соответствует булев вектор

$$(d_1, \ldots, d_k, \bar{d}_{k+1}, \ldots, \bar{d}_n),$$

где компоненты $d_1, \ldots, d_k$ отождествляются с истинными, а $\bar{d}_{k+1}, \ldots, \bar{d}_n$ — с ложными связями.

Пару (j,k), доказать ассоциацию с ложной связью для которой не удалось, проверяют на возможность отождествления с ранее выделенными связями.

Пусть $w(j) \in Ts_i : w(j) = St(j) \bullet Fl(j)$, где символьная цепочка $Fl(j)$ представляет флективную часть слова $w(j)$, а символом "$\bullet$" обозначается операция конкатенации. Аналогично пусть $w(k) \in Ts_i$ и при этом $w(k) = St(k) \bullet Fl(k)$. Обозначим множество ранее выделенных связей как Lnk. Каждый элемент в Lnk представляется четвёркой

$$(Id, St_1, St_2, FCm),$$

где Id — идентификационный номер СЯУ; St_1 — основа главного, St_2 — зависимого слова; FCm — список пар вида «флексия главного слова — флексия зависимого».

Считается, что паре (j,k) соответствует связь $((j,k), \rightarrow)$ в рамках заданной СЯУ, если для некоторой СЯУ с идентификационным номером Id существует $(Id, St_1, St_2, FCm) \in Lnk : St(j) = St_1$, $St(k) = St_2$, а $(Fl(j), Fl(k)) \in FCm$.

В случае, когда $St(j) = St_2$, $St(k) = St_1$, а список FCm содержит пару $(Fl(k), Fl(j))$, паре (j,k) будет отвечать связь $((j,k), \leftarrow)$.

Как и на этапе формирования начальных знаний, пару (j,k), для которой не нашлось ассоциации ни с одной из ранее выделенных связей (ложных или истинных), проверяют на наличие связи, опрашивая эксперта.

На основе найденного множества R_J далее идёт отбор фраз $Ts_i \in Ts$ для построения множества признаков формального контекста (1) эталона СЯУ.

Первым шагом из состава каждого $LS_i \subset LC$ исключаются те МЛС, которые включают индексы, не вошедшие ни в одну из связей в составе R_J. Введём обозначение LC^* для преобразованного таким образом множества LC, аналогично LC_i^* — для $\forall\ LS_i \subset LC$.

По каждому $LS_i^* \subset LC^*$ отбирается Ts_i:

$$Ls(Ts_i) \in LS_i^*, |Ts_i| \to min. \tag{3}$$

Совокупность фраз $Ts_i \in Ts$ отвечающих условию (3), обозначим как Ts^*.

Заключительный шаг формирования ФК вида (1) эталона СЯУ состоит в построении признакового множества M и объектно-признаковых связей в рамках отношения $I \subseteq G \times M$ на основе найденных R_J и Ts^*.

В целях более точного выделения объектов и признаков эталона введём процедуру согласования знаний относительно разных СЯУ заданной предметной области. Пусть модель (1) есть единица тезауруса, представляемого тройкой

$$Kth = (Gth, Mth, Ith), \tag{4}$$

где Gth состоит из символьных пометок отдельных СЯУ, Mth включает признаки ФК вида (1) каждой $gth \in Gth$. Кроме того, в составе Mth выделяются:

- множество указаний на объекты формальных контекстов вида (1), генерируемых для отдельных $gth \in Gth$ (обозначим далее это множество как M_6);
- множество сочетаний основы и флексии для зависимого слова (M_7);
- множество сочетаний основ зависимого и главного слова (M_8).

Модель (4) позволяет определить процедуру согласования единиц знаний с по-мощью следующего правила.

Правило 1. *Пусть St_j есть основа, Fl_j — флексия слова w, найденные относительно СЯУ S_j. Предположим, что $w = St_1 \bullet Fl_1$ для СЯУ S_1, $w = St_2 \bullet Fl_2$ для СЯУ S_2, причём $S_1 = S_2 \bullet suf$, где suf содержит минимум один символ. Тогда относительно S_1 основа St_1 может быть заменена основой St_2, а флексия Fl_1 — флексией $Fl_3 = suf \bullet Fl_2$ только в том случае, если встречаемость флексий Fl_3 и Fl_2 в отношениях из $Ith \subseteq Gth \times Mth$ не уменьшается при выполнении указанных замен.*

Качественно процесс формирования эталонов СЯУ характеризуется динамикой изменения показателей сжатия информации в тезаурусе, представляемом решёткой для ФК (4) — множеством всех его ФП вместе с отношением порядка. Наиболее естественными показателями такого рода здесь могут быть коэффициенты сжатия по основам и флексиям, аналогичные введённым в [4].

Коэффициент сжатия по основам относительно модели (4) определяется как

$$ksth = \frac{\sum\limits_{i=1}^{nbsth} ksth_i}{nbsth}, \tag{5}$$

где согласно обозначениям для подмножеств множества признаков ФК (1)

$$nbsth = |Mth_1|;\ ksth_i = \frac{\sum\limits_{k=1}^{nmfth}\sum\limits_{j=1}^{ndm_{ki}} nfms_{ijk}}{nbsth_i};\ nmfth = |Mth_2|;$$

$$nfms_{ijk} = \big|\{mth_i \in Mth_3 : Ith(gth_j, mth_i) = true, gth \in Gth,$$

$$\exists m_{bf} \in Mth_2 : m_{bf} = p_{bf} \bullet f_i, mth_i = b_i \bullet ":" \bullet f_i,$$

$$Ith(gth_j, m_{bf}) = true,$$

$$\exists m_{bs} \in Mth_1 : m_{bs} = p_{bs} \bullet b_i, \quad Ith(gth_j, m_{bs}) = true,$$
$$\exists mth_k \in M_6 : mth_k = p_b \bullet b_k, \quad Ith(gth_j, mth_k) = true,$$
$$\exists mth \in M_8 : mth = b_k \bullet b_i, \quad Ith(gth_j, mth) = true\}\big|;$$
$$ndm_{ki} = \big|\{gth_j \in Gth : Ith(gth_j, mth) = true,\ mth \in M_8,\ mth = b_k \bullet b_i\}\big|;$$
$$nmfth = |Mth_2|\,;$$

p_{bf}, p_{bs} и p_b соответствуют символьным константам «главное–флексия:», «главное–основа:» и «основа:», соответственно.

По аналогии с коэффициентом (5) коэффициент сжатия информации по флексиям относительно формального контекста (4) равен

$$ksth = \frac{\sum\limits_{i=1}^{nbsth} ksth_i}{nbsth}, \tag{6}$$

где согласно ранее принятым обозначениям

$$nfsth = |Mth_5|; \quad kfth_i = \frac{\sum\limits_{j=1}^{nfsth_i}\sum\limits_{k=1}^{nmfth} nafth_{ijk}}{nfsth_i};$$

$$nfsth_i = \big|\{gth \in Gth:\ Ith(gth, mth) = true,\ mth \in Mth_5,\ mth = p_{fl} \bullet f_i\}\big|;$$
$$nafth_{ijk} = \big|\{mth \in Mth_4:\ Ith(gth_j, mth) = true,$$
$$\exists m_{bf} \in Mth_2:\ m_{bf} = p_{bf} \bullet f_k,\ mth = f_i \bullet ":" f_k\}\big|;$$

p_{fl} есть обозначение символьной константы «флексия:».

Графики на рис. 1 иллюстрируют динамику изменения значений оценок (5) и (6) при последовательном добавлении в тезаурус СЯУ из табл. 1 (без выполнения процедуры согласования знаний согласно Правилу 1).

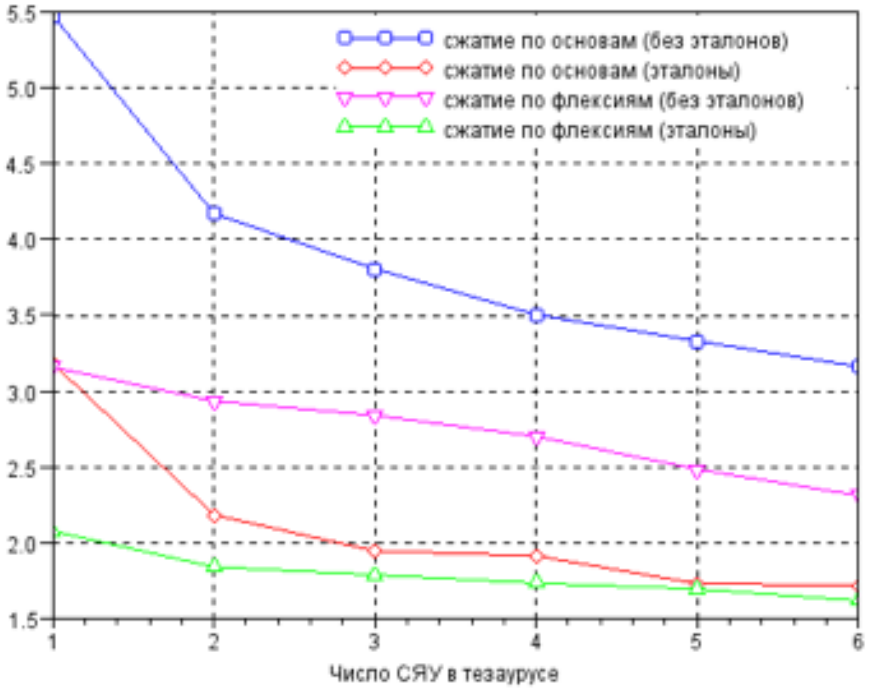

Рис. 1. Сжатие информации относительно формального контекста тезауруса.

Для сравнения в табл. 2 приводятся значения числа СЭ-фраз, задающих СЯУ (N_1), ЕЯ-фраз, определяющих эталон (N_2), исходного числа объектов (N_3) и признаков СЯУ (N_4), числа объектов (N_5) и признаков эталона (N_6).

Диаграмма на рис. 2 иллюстрирует дополнительный рост специфичности формальных понятий в решётке тезауруса при выполнении процедуры согласования знаний в соответствии с Правилом 1 для СЯУ из табл. 1. Индикатором роста специфичности формальных понятий является постепенное уменьшение значений коэффициентов (5) и (6) при добавлении новых СЯУ в тезаурус. При этом специфичность

i	Фраза максимальной длины из определяющих СЯУ
1	Нежелательное переобучение является причиной заниженности сред-ней величины ошибки алгоритма на обучающей выборке.
2	Тренировочная выборка, на ней проявляется эффект заниженных значений средней ошибки, причиной же является переусложненная модель.
3	Контрольная выборка, принятие деревом решения на ней будет с большей вероятностью ошибки именно по причине переподгонки.
4	Оценка частоты ошибок на выборке, взятой в качестве контрольной, может для алгоритма оказаться заниженной по причине переподгонки.
5	Заниженность оценки ошибки распознавания зависит от выбора правила принятия решений.
6	Число закономерностей алгоритмической композиции влияет на частоту ошибок логического классификационного алгоритма на контрольной выборке.

Таблица 1. Ситуации языкового употребления.

i	1	2	3	4	5	6
N_1	56	28	29	30	6	10
N_2	8	9	7	9	1	2
N_3	18	17	15	13	12	14
N_4	177	186	173	162	94	81
N_5	9	12	12	11	8	12
N_6	82	90	80	69	35	53

Таблица 2. Эталоны для СЯУ из табл. 1.

формального понятия численно оценивается кратчайшим расстоянием от вершинного ФП в решётке до рассматриваемого ФП.

Применение эталона СЯУ в качестве единицы тезауруса, задаваемого моделью (4), позволяет сократить его размер не менее чем на 40–50%. Согласование эталонов как единиц знаний по Правилу 1 даёт дополнительное сокращение размеров такого тезауруса в среднем на 1,5%.

Оценка схожести СЯУ в системе контроля знаний

Предложенный метод выделения эталона на СЯУ реализован в рамках демоверсии системы контроля знаний, представленной (вместе с исходными текстами на Visual Prolog 5.2) в подразделе «Участник:Dmitry.Mikhaylov» раздела «Страницы

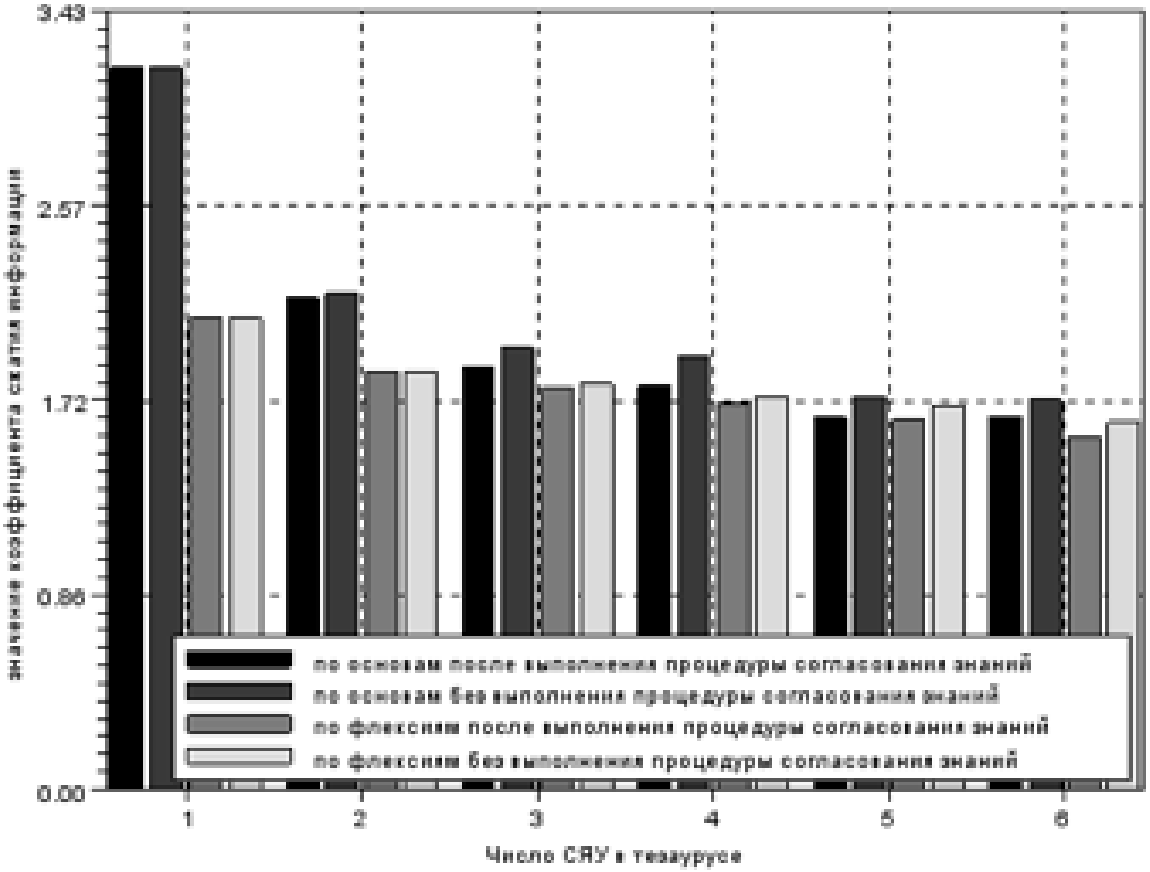

Рис. 2. Сжатие информации тезауруса (эталоны выделены).

участников» ресурса [2]. При этом согласование знаний, формируемых разными экспертами по одной и той же предметной области, может быть наглядно проиллюстрировано вычислением оценок близости ответа испытуемого правильному ответу. Суть данных оценок была рассмотрена нами в [7].

Пусть СЯУ S_1 соответствует правильному ответу, который формулируется экспертом, а СЯУ S_2 — ответу испытуемого. Введём обозначения формальных контекстов вида (1): для $S_1 - Ke$, а для $S_2 - Kx$, где $Ke = (Ge, Me, Ie)$ и $Kx = (Gx, Mx, Ix)$, $Ie \subseteq Ge \times Me$ и $Ix \subseteq Gx \times Mx$, соответственно. Результат объединения M_6, M_7, M_8, Me_4, Mx_4, Me_5 и Mx_5 обозначим как M_U (здесь используются принятые нами ранее обозначения подмножеств в составе признаковых множеств формальных контекстов (1) и (4)).

Определение 2. Будем считать, что S_1 и S_2 связаны отношением близости, если каждому объекту $gx \in Gx$ соответствует такой объект $ge \in Ge$, что выполняется одно из следующих условий:

1. $gx = ge$ и любой признак $me \in Me$ объекта ge относится и к gx.
2. $gx = ge$, при этом условие (1) не выполняется, но существует $gth \in Gth$, обладающий признаком $mth_1 \in M_6 : mth_1 = p_b \bullet ge$ при обязательном выполнении следующих условий:

$$(\exists me_{fl} \in Me_5 : me_{fl} = p_{fl} \bullet fe) \rightarrow (\exists mth_{17} \in M_7 : mth_{17} = ge \bullet ":" \bullet fe),$$
$$\text{при этом } (Ie(ge, me_{fl}) \wedge Ix(ge, me_{fl})) \rightarrow Ith(gth, mth_{17});$$

$$(\exists me_{bs} \in Me_1 : me_{bs} = p_{bs} \bullet be) \rightarrow (\exists mth_{18} \in M_8 : mth_{18} = ge \bullet ":" \bullet be),$$
$$\text{при этом } Ie(ge, me_{bs}) \rightarrow Ith(gth, mth_{18});$$

$$(\exists mx_{bs} \in Mx_1 : mx_{bs} = p_{bs} \bullet bx) \rightarrow (\exists mth_{28} \in M_8 : mth_{28} = ge \bullet'' :'' \bullet bx),$$
$$\text{при этом } Ix(ge, mx_{bs}) \rightarrow Ith(gth, mth_{28});$$

Кроме того, для $\forall mth \in (Mth \setminus M_U)$ истинно:

$$Ith(gth, mth) \rightarrow (Ie(ge, mth) \wedge Ix(ge, mth)). \quad (7)$$

3. $gx \neq ge$, но существует объект $gth \in Gth$, обладающий признаками $mth_1 \in M_6 : mth_1 = p_b \bullet ge$ и $mth_2 \in M_6 : mth_2 = p_b \bullet gx$, при этом для любого признака $mth \in (Mth \setminus M_U)$ справедливо:

$$Ith(gth, mth) \rightarrow (Ie(ge, mth) \wedge Ix(ge, mth)). \quad (8)$$

4. $gx \neq ge$, но существует объект $gth_1 \in Gth$, обладающий признаком $mth_1 \in M_6 : mth_1 = p_b \bullet ge$, а для $\forall me \in (Me_4 \cup Me_5)$ верно:

$$(Ith(gth_1, mth_1) \wedge Ie(ge, me)) \rightarrow Ith(gth_1, me).$$

При этом существуют признаки $mth_2 \in M_6 : mth_2 = p_b \bullet gxg$ и $mx \in (Mx_1 \cup Mx_2 \cup Mx_3)$, для которых верно:

$$(Ith(gth_1, mth_2) \wedge Ix(gx, mx)) \rightarrow Ith(gth_1, mx),$$

где $gxg \neq gx$, а пара (gxg, ge) отвечает условию (3) при генерации ФК вида (1) для объекта gth_1. В то же время существует объект $gth_2 \in Gth$ относительно

которого пара (gx, gxg) также будет отвечать условию (3) настоящего определения. Генерируемый при этом формальный контекст вида (1) для обозначим как Kxg, $Kxg = (Gxg, Mxg, Ixg)$.

Близость ситуаций S_1 и S_2 численно оценивается как

$$spc(S_1, S_2) = \frac{\sum_{k=1}^{n} spc_k}{n}, \tag{9}$$

где $n = |Gx|$, а spc_k есть значение близости объектов в паре (gx_k, ge). В зависимости от выполнимости условий Определения 2 значение spc_k либо равно 1,0, если для (gx_k, ge) выполнено условие (1), либо вычисляется по формуле:

$$-\log_2\left(1 - \frac{D_c}{path_C}\right) \times \frac{|BLCS|}{|B1 \setminus BLCS| + |B2 \setminus BLCS| + |BLCS|}, \tag{10}$$

если для пары (gx_k, ge) выполнено условие (2), (3) либо (4).

Во втором случае имеем гипотетическую решётку ФП (обозначим её как $\Re xe$), в которой объемы объектных ФП (формальных понятий с одним объектом в составе объема) есть $\{gx_k\}$ и $\{ge\}$ (при выполнении условия (2) или (3)) либо $\{gx_k\}$, $\{ge\}$ и $\{gxg\}$ (при выполнении условия (4)). Значение D_c равно числу сравнимых формальных понятий, составляющих цепочку с вершинным ФП решётки $\Re xe$ в качестве максимального ФП и наименьшим общим суперпонятием (НОСП) для объектных формальных понятий решётки $\Re xe$ — в качестве минимального ФП. Множество $BLCS$ есть содержание (множество признаков всех объектов) этого НОСП, а число $path_C$ равно минимальному числу ФП в цепочке, которой принадлежит вершинное ФП, наименьшее ФП решётки $\Re xe$ и формальное понятие с содержанием $BLCS$.

В случае выполнения любого из условий (2), (3) или (4) значение $D_c = 2$ (доказательство очевидно).

При выполнении условия (2) либо (3) $path_C = 4$, а в $BLCS$ войдут признаки $mth \in (Mth \setminus M_U)$, для каждого из которых справедливо либо соотношение (7) (при выполнении условия (2)), либо соотношение (8) (при выполнении условия (3)). Множества B_1 и B_2 в этом случае определяются следующим образом:

$$B_1 = \{me : me \in (Me_1 \cup Me_2 \cup Me_3),\ Ie(ge, me) = true\},$$

$$B_2 = \{mx : mx \in (Mx_1 \cup Mx_2 \cup Mx_3),\ Ix(gx, mx) = true\}.$$

Доказательство выполнимости условия (4) обычно происходит в несколько итераций. При этом в ходе каждой последующей итерации число признаков, не являющихся общими для gx_k и gxg, всегда меньше, чем в предыдущей. На-чальное значение

$path_C$, равное 4, в ходе каждой итерации возрастает на 1, а

$$B_1 = \{mxg : mxg \in (Mxg_1 \cup Mxg_2 \cup Mxg_3),\ Ixg(gxg, mxg) = true\},$$

$$B_2 = \{mx : mx \in (Mxg_1 \cup Mxg_2 \cup Mxg_3),\ Ixg(gx_k, mx) = true\},$$

где $(Mxg_1 \cup Mxg_2 \cup Mxg_3) \subset Mxg$ в соответствии с показанным выше разделением множества признаков ФК вида (1), а $BLCS = B_1 \cap B_2$.

В реализованной системе (рис. 3) для ответа испытуемого производится поиск наиболее близкой (по буквенному составу) из всех СЭ-форм, определяющих СЯУ правильного ответа. Далее идёт анализ словесных несовпадений, поиск соответствий для несовпадающих частей сравниваемых предложений уже в составе эталона правильного ответа и вычисление оценок (10) с учётом найденных синонимов. Указанные оценки вычисляются для случаев неполного ответа, орфографических ошибок, а также лишних слов, не фигурирующих в лексико-синтаксических связях из представленных в базе знаний системы.

Случай 1. Неполный ответ — для всех слов и словосочетаний из ответа испытуемого нашлись прообразы в наиболее близком «правильном» варианте, но для части слов правильного ответа не нашлось прообразов в ответе испытуемого. Ненулевое значение оценки (10) будет только для тех упущенных слов, которые в «правильном» варианте являются синтаксически зависимыми по отношению к некоторым другим словам из анализируемого ответа. Здесь имеет место обобщение оценки (10) на случай, когда для одного из сравниваемых объектов не определены признаки из множеств Mx_5, Mx_4, M_6, M_7, M_8. Таким объектом является основа слова, упущенного в ответе испытуемого. Значение оценки (10) для упущенного слова равно $-\log_2\left(1 - \dfrac{2}{4}\right) \times \dfrac{3}{(8-3)+(8-3)+3} \approx 0.23$.

Случай 2. Орфографические ошибки (из допустимых) — слово из ответа испытуемого и слово из варианта правильного ответа есть различные формы одного и того же слова, допустимыми в рамках одной лексико-синтаксической связи (не обязательно в рамках рассматриваемой СЯУ). В этом случае оценка (10) для анализируемой пары слов вычисляется аналогично общему случаю.

Случай 3. «Лишние» слова — ситуация, когда все слова из наиболее близкого «правильного» варианта нашли прообраз в ответе испытуемого, но в анализируемом ответе есть слова, которые не нашли прообразов в «правильном» (в том числе и на уровне словосочетаний). В этом случае ответ испытуемого не будет засчитан как неверный только тогда, когда «лишние» слова не фигурируют в базе знаний системы ни в одной лексико-синтаксической связи. При этом значение оценки (10) для каждого «лишнего» слова принимается равным нулю.

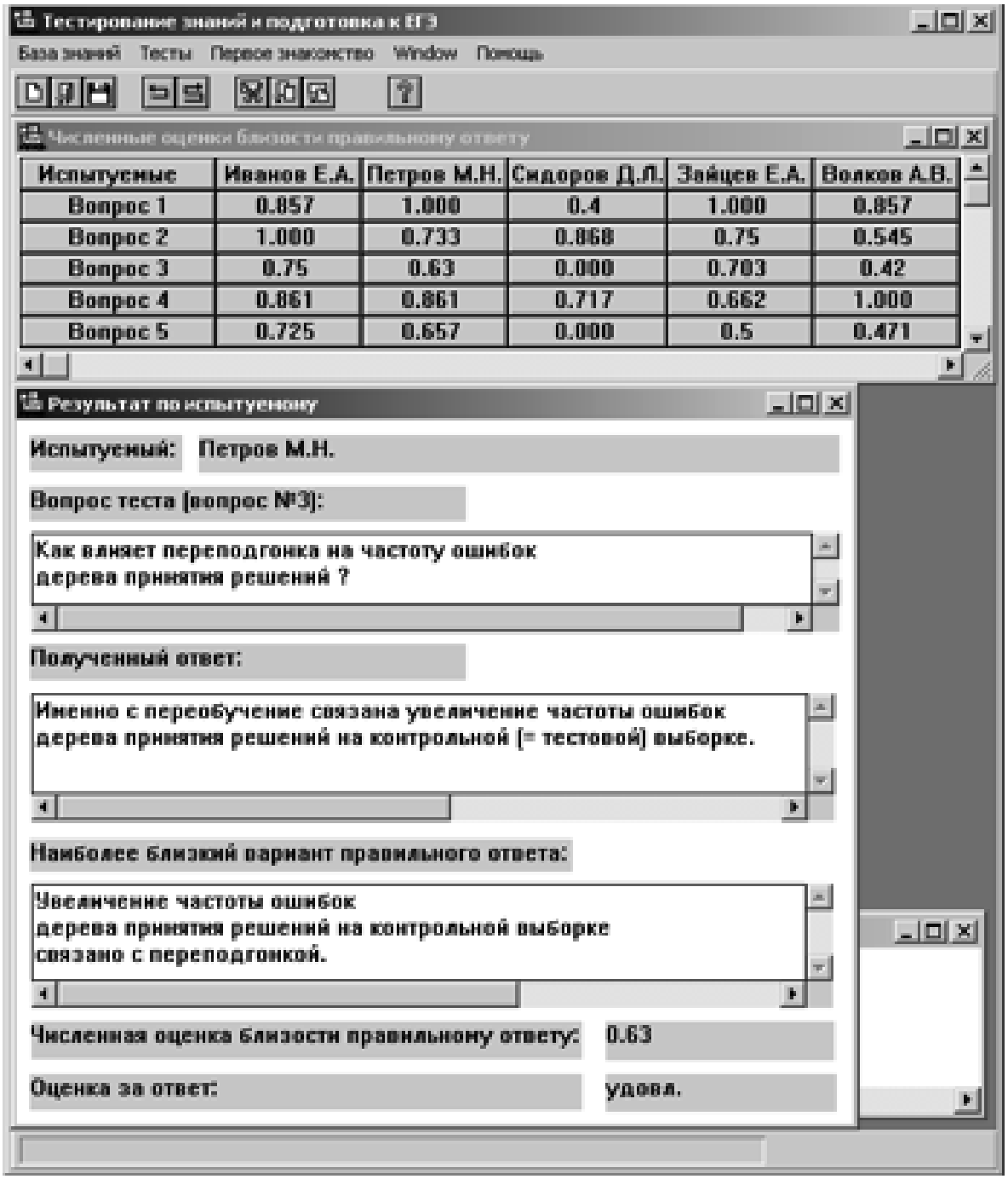

Рис. 3. Интерфейс системы и пример интерпретации ответа.

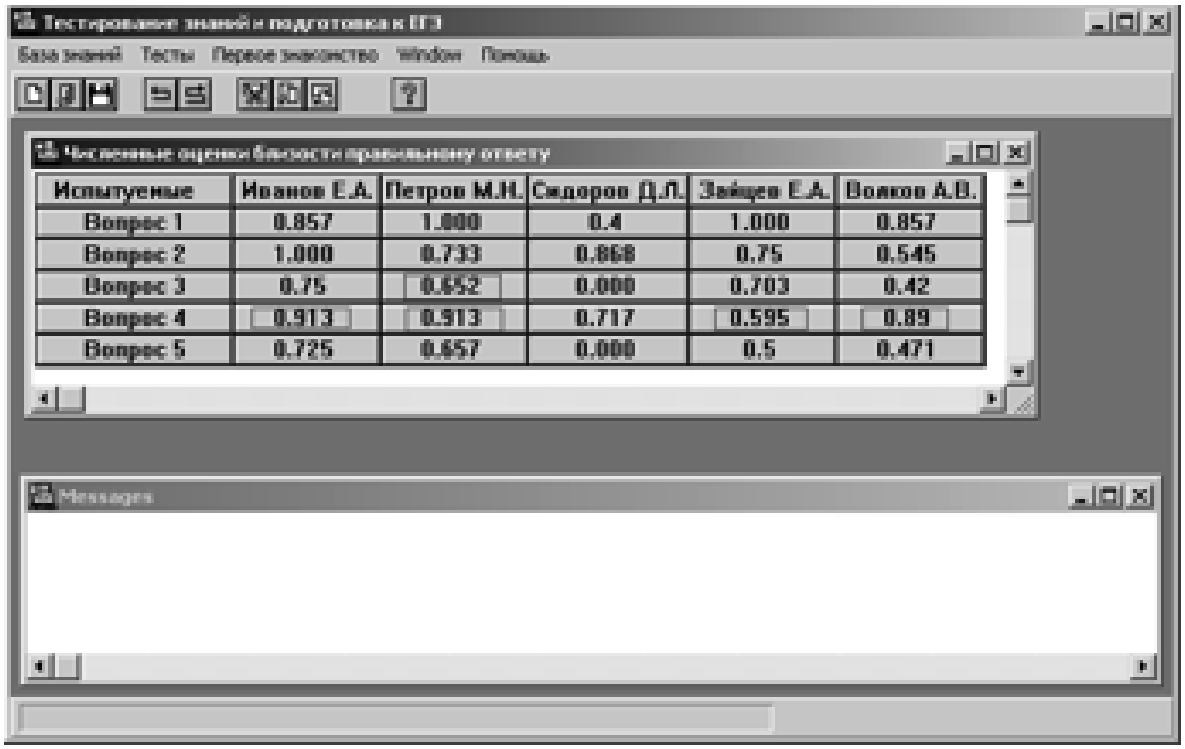

Рис. 4. Результаты из примера на рис. 3 после согласования знаний по разным СЯУ.

Применение *Правила 1* к ситуациям языкового употребления, нашедшим отражение в тезаурусе, последовательно иллюстрируют рис. 3 и 4. Каждая из уточнённых оценок на рис. 4 обведена прямоугольником. Незначительное снижение оценок близости правильному ответу на *Вопрос 4* у испытуемых *Зайцева Е. А.* и *Волкова А. В.* обусловлено заменой выделенных ранее нулевых флексий у ряда слов, представленных в тезаурусе.

Заключение

Сокращая размер базы знаний для оценки семантической схожести текстов предметно-ограниченного ЕЯ минимум на 40–50%, разработанный метод формирования эталона СЯУ позволяет найти необходимый и достаточный объём текстовых данных для передачи знаний между учителем и учеником. При этом предложенная концепция СЯУ составляет основу решения задач поиска систем зависимостей совместной встречаемости осмысленных фрагментов слов в кон-тексте связного текста. Сказанное, в частности, немаловажно для сокращения перебора при построении

смыслового контекста в мультиагентном подходе [6]. В данной работе все виды связей между главным и зависимым словом предпо-лагались одинаково значимыми для оценки схожести фраз. Для применения та-ких оценок в задачах оценки профессиональных знаний по отраслям [5] опреде-ление схожести СЯУ необходимо переформулировать с позиций нечёткой логи-ки. При этом для описания функций принадлежности необходим системный ана-лиз структуры профессиональных знаний в конкретной области. Работа выполнена при поддержке РФФИ (проект №13-01-00055).

Список литературы

1. Ganter, B. and Wille, R. (1999). Formal concept analysis. Berlin: Springer.
2. MachineLearning.ru. (2014). Available at http://www.machinelearning.ru
3. Mikhailov D. V. and Emelyanov G. M. (2009). Forming and clustering of syntactic relations on the bases of Natural Language's using's situations. *Interactive systems and technologies: the problems of human-computer interaction. Collection of scientific papers. Ulyanovks, ULSTU.* Vol. III, pp. 295–307.
4. Емельянов Г. М., Михайлов Д. В. Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний // Всерос. конф. ММРО-15. — М.: Макс Пресс, 2011. — С. 581–584.

   Yemelyanov, G. M. and Mikhaylov, D. V. (2011). Formal Concept Analysis and compression of text information on the problem of automated control of knowledge. MMRO-2015 Conference. Moscow, Max Press, pp. 581–584.

5. Краснов А. Н., Мошков И.С., Якимов В.Н. Компьютерная система анализа текста таксономического типа применительно к оценке профессиональных знаний [Текст] // Междунар. науч.-практ. конф. «Инновация-2011». — Ташкент: Ташкентский гос. техн. ун-т, 2011. — С. 287–289.

   Krasnov, A., Moshkov, I. and Yakimov, V. (2015). Taxonomic type text analysis computer system for the estimation of professional knowledge. International Conference «Innovation-2011». Tashkent, TSTU, pp. 287–289.

6. Минаков И. А. Интеграция профессиональных знаний, представленных в виде текстов на естественном языке // Вестник СамГТУ, серия «Технические науки», 2007. — № 1 (19). — С. 28–35.

   Minakov, I. (2007). Integration of professional knowledge which presented by the natural language texts. Herald of SamSTU, «Technical Sciences» Series, No. 1, pp. 28–35.

7. Михайлов Д. В., Емельянов Г. М. Семантическая схожесть текстов в задаче автоматизированного контроля знаний // Межд. конф. ИОИ-2010. — М.: Макс Пресс, 2010. — С. 516–519.

   Mikhaylov, D. and Yemelyanov, G. (2010). The semantic similarity of texts in the problem of automated control of knowledge. International Conference IOI-2010, pp. 516–519.

Статья поступила в редакцию 05.12.2014