Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний

Емельянов Г. М., Михайлов Д. В.

Новгородский государственный университет
имени Ярослава Мудрого

Всероссийская конференция
«Математические методы распознавания образов» (ММРО-15),

11–17 сентября 2011 г.

г. Петрозаводск

Цель и задачи исследования

Основная цель

Разработка теоретико-методологических основ организации Тестовых Заданий Открытой Формы (ТЗОФ) в системах контроля знаний.

Основные задачи исследования

- Анализ существующих подходов к моделированию семантики конструкций Естественного Языка (ЕЯ) и определение требований к механизму сравнения смыслов на функциональном уровне.
- Создание и апробация методов автоматизированного накопления знаний о Семантической Эквивалентности (СЭ) в предметно-ограниченном ЕЯ.
- Построение механизма интерпретации ответа обучаемого на тестовое задание открытой формы.
- Разработка архитектуры системы контроля знаний на основе ТЗОФ.

Определение 1

Конструкция ЕЯ — последовательность знаков, используемая для фиксации некоторого числа высказываний этого ЕЯ в памяти ЭВМ.

Определение 2

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

Фиксируемый СЯУ S языковой контекст представляется тройкой:

$$S = \left(O, R, T^S\right), \tag{1}$$

где O — множество объектов-участников S;
R — множество отношений между $o \in O$;
T^S — множество форм языкового описания S.

Представим языковой контекст СЯУ посредством формального контекста:

$$K^S = \left(G^S, M^S, I^S\right), \tag{2}$$

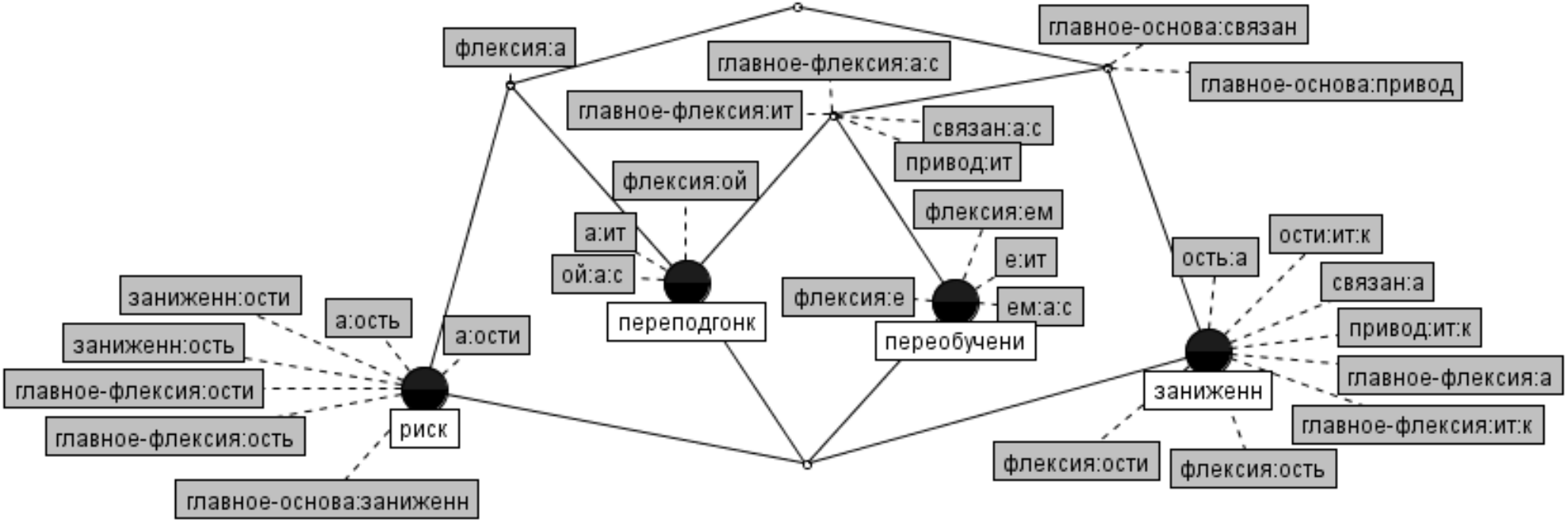
где $\forall g \in G^S$ — основа слова, синтаксически подчинённого другому слову из некоторой $T_i \in T^S$ в составе структуры (1).

Множество признаков M^S включает подмножества, содержащие:

- указания на основу синтаксически главного слова (M_1);
- указания на флексию главного слова (M_2);
- связи «основа–флексия» для синтаксически главного слова (M_3);
- сочетания флексий зависимого и главного слова (M_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 (M_5).

Посредством $I^S \subseteq G^S \times M^S$ выделяются классы отношений из R в (1) по сходству основы главного, флексии зависимого слова, лексической и флективной сочетаемости.

Пример формального контекста СЯУ



Рассмотрим модель тезауруса в виде формального контекста:

$$K^{TH} = \left(G^{TH}, M^{TH}, I^{TH}\right), \tag{3}$$

где G^{TH} состоит из символьных пометок отдельных СЯУ.

M^{TH} содержит признаки формальных контекстов всех $g^{TH} \in G^{TH}$.

Кроме того, в составе M^{TH} выделяются подмножества:

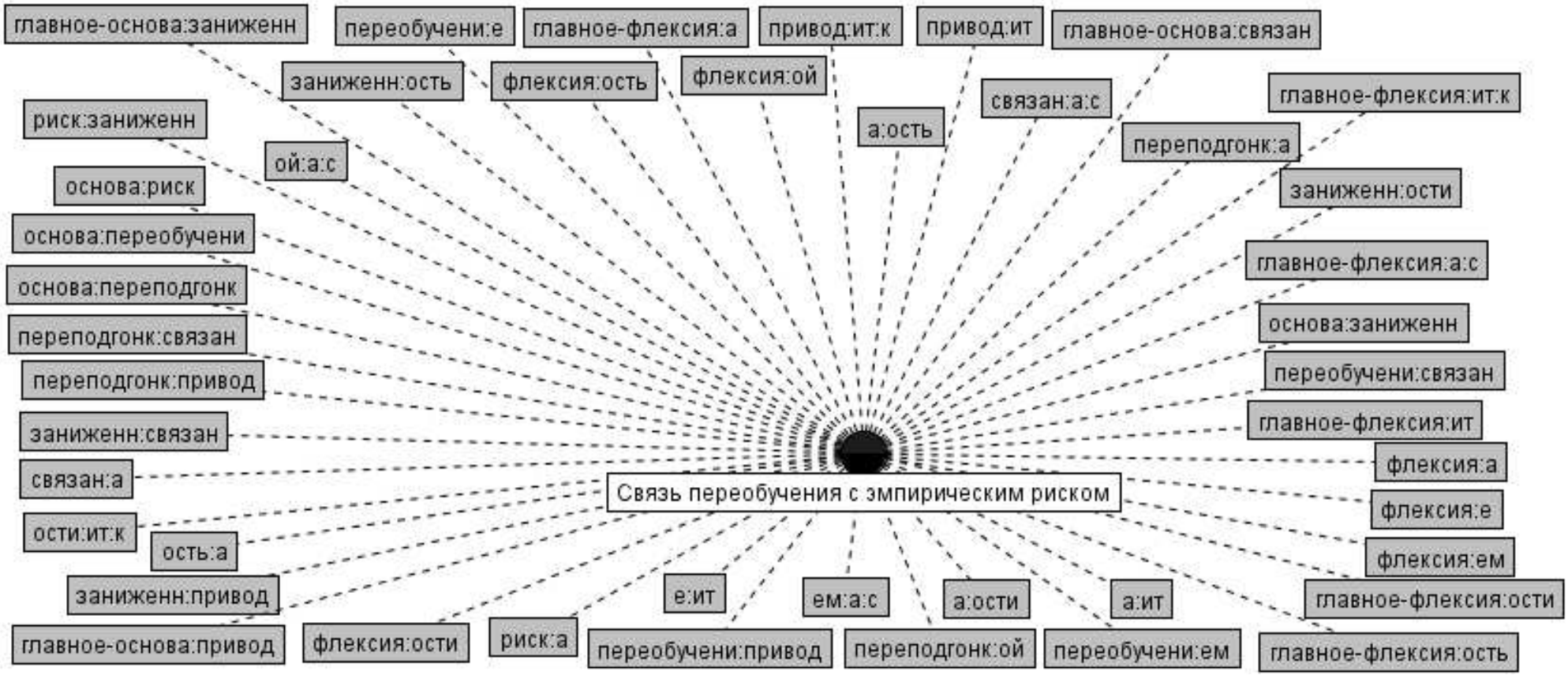
- M_6 — указаний на объекты формальных контекстов вида (2) отдельных $g^{TH} \in G^{TH}$;
- M_7 — множество связей «основа–флексия» для синтаксически зависимого слова;
- M_8 — множество сочетаний основ зависимого и главного слова.

Пусть $K^E = \left(G^E, M^E, I^E\right)$ есть формальный контекст СЯУ S_1 корректного описания некоторого факта, $K^X = \left(G^X, M^X, I^X\right)$ — формальный контекст произвольной СЯУ S_2, а $M^U = M_6 \cup M_7 \cup M_8 \cup M_4^E \cup M_4^X \cup M_5^E \cup M_5^X$.

p_{fl} и p_b есть обозначения для констант «флексия:» и «основа:».

Простейший случай *схожести* S_1 и S_2: для $\forall g^X \in G^X \; \exists\, g^E \in G^E : g^X = g^E$ и любой признак $m^E \in M^E$ объекта g^E относится к g^X.

Пример объекта отдельной СЯУ в формальном контексте тезауруса



Случай 2 соответствия объектов

$g^X = g^E$, условие *простейшего случая* не выполняется, но существует объект $g^{TH} \in G^{TH}$, обладающий признаком $m_1^{TH} \in M_6 : m_1^{TH} = p_b \odot g^E$ при обязательном выполнении следующих условий:

$$\left(\exists\, m_{fl}^E \in M_5^E : m_{fl}^E = p_{fl} \odot f^E\right) \rightarrow \left(\exists\, m_{17}^{TH} \in M_7 : m_{17}^{TH} = g^E \odot «:» \odot f^E\right),$$

$$\text{при этом } \left(I^E\left(g^E, m_{fl}^E\right) \wedge I^X\left(g^E, m_{fl}^E\right)\right) \rightarrow I^{TH}\left(g^{TH}, m_{17}^{TH}\right);$$

$$\left(\exists\, m_{bs}^E \in M_1^E : m_{bs}^E = p_{bs} \odot b^E\right) \rightarrow \left(\exists\, m_{18}^{TH} \in M_8 : m_{18}^{TH} = g^E \odot «:» \odot b^E\right),$$

$$\text{при этом } I^E\left(g^E, m_{bs}^E\right) \rightarrow I^{TH}\left(g^{TH}, m_{18}^{TH}\right);$$

$$\left(\exists\, m_{bs}^X \in M_1^X : m_{bs}^X = p_{bs} \odot b^X\right) \rightarrow \left(\exists\, m_{28}^{TH} \in M_8 : m_{28}^{TH} = g^E \odot «:» \odot b^X\right),$$

$$\text{при этом } I^X\left(g^E, m_{bs}^X\right) \rightarrow I^{TH}\left(g^{TH}, m_{28}^{TH}\right).$$

Кроме того, для $\forall\, m^{TH} \in \left(M^{TH} \setminus M^U\right)$ верно:

$$I^{TH}\left(g^{TH}, m^{TH}\right) \rightarrow \left(I^E\left(g^E, m^{TH}\right) \wedge I^X\left(g^E, m^{TH}\right)\right). \qquad (4)$$

Случай 3 соответствия объектов

$g^X \neq g^E$, но существует объект $g^{TH} \in G^{TH}$, обладающий признаками

$$m_1^{TH} \in M_6 : m_1^{TH} = p_b \odot g^E \text{ и}$$
$$m_2^{TH} \in M_6 : m_2^{TH} = p_b \odot g^X,$$

при этом для любого признака $m^{TH} \in \left(M^{TH} \backslash M^U\right)$ справедливо:

$$I^{TH}\left(g^{TH}, m^{TH}\right) \rightarrow \left(I^E\left(g^E, m^{TH}\right) \wedge I^X\left(g^X, m^{TH}\right)\right). \quad (5)$$

Замечание

Численная оценка схожести СЯУ включает сравнение последовательностей двух и более соподчинённых слов. Случаи схожести здесь анализируются только для главных слов. Последовательности считаются заменяемыми, если возможно их построение по формальному контексту (3) на наборе признаков с префиксом p_{bs} для одной и той же СЯУ.

Случай 4 соответствия объектов

$g^X \neq g^E$, но существует $g_1^{TH} \in G^{TH}$, обладающий признаком $m_1^{TH} \in M_6$: $m_1^{TH} = p_b \odot g^E$, а для $\forall\, m^E \in \left(M_4^E \cup M_5^E\right)$ верно то, что

$$\left(I^{TH}\left(g_1^{TH}, m_1^{TH}\right) \wedge I^E\left(g^E, m^E\right)\right) \rightarrow I^{TH}\left(g_1^{TH}, m^E\right).$$

При этом существуют признаки $m_2^{TH} \in M_6$ и $m^X \in \left(M_1^X \cup M_2^X \cup M_3^X\right)$:

$$\left(I^{TH}\left(g_1^{TH}, m_2^{TH}\right) \wedge I^X\left(g^X, m^X\right)\right) \rightarrow I^{TH}\left(g_1^{TH}, m^X\right),$$

где $m_2^{TH} = p_b \odot g^{X_1}$, $g^{X_1} \neq g^X$, а пара $\left(g^{X_1}, g^E\right)$ соответствует *Случаю 3* при генерации формального контекста для g_1^{TH}.

В то же время существует объект $g_2^{TH} \in G^{TH}$, относительно которого пара $\left(g^X, g^{X_1}\right)$ также будет соответствовать *Случаю 3*.

Генерируемый при этом формальный контекст для g_2^H обозначим далее как K^{X_1}. По аналогии с K^E и K^X, $K^{X_1} = \left(G^{X_1}, M^{X_1}, I^{X_1}\right)$.

Оценка схожести ситуаций языкового употребления S_1 и S_2 относительно их формальных контекстов $K^E = (G^E, M^E, I^E)$ и $K^X = (G^X, M^X, I^X)$, из которых удалена информация РПЗ , вычисляется по формуле:

$$spc(S_1, S_2) = \frac{\sum_{k=1}^{n} spc_k}{n}, \tag{6}$$

где $n = |G^X|$,

spc_k есть численное значение схожести объектов в паре (g_k^X, g^E).

Если (g_k^X, g^E) не относится ни к одному из четырёх случаев соответствия объектов схожих СЯУ, то $spc(S_1, S_2) = 0$.

При взаимно-однозначном соответствии признаков объектов g^E и g^X значение spc_k равно 1.0.

Если пара $\left(g_k^X, g^E\right)$ отвечает одному из *Случаев 2–3* соответствия объектов, то оценка схожести g_k^X и g^E вычисляется по формуле:

$$-\log_2\left(1-\frac{D_c}{path_C}\right)\times\frac{\left|B^{LCS}\right|}{\left|B_1\backslash B^{LCS}\right|+\left|B_2\backslash B^{LCS}\right|+\left|B^{LCS}\right|}, \qquad (7)$$

где $D_c = 2$, число $path_C = 4$.

В множество B^{LCS} войдут признаки $m^{TH} \in \left(M^{TH}\backslash M^U\right)$, для которых справедливо либо соотношение (4) для *Случая 2*,
либо соотношение (5) для *Случая 3*.

При этом

$$B_1 = \left\{ m^E : m^E \in \left(M_1^E \cup M_2^E \cup M_3^E \right), I^E\left(g^E, m^E\right) = \text{true} \right\},$$
$$B_2 = \left\{ m^X : m^X \in \left(M_1^X \cup M_2^X \cup M_3^X \right), I^X\left(g_k^X, m^X\right) = \text{true} \right\}.$$

Если пара $\left(g_k^X, g^E\right)$ отвечает *Случаю 4* соответствия объектов, то оценка схожести g_k^X и g^E вычисляется по формуле:

$$-\log_2\left(1-\frac{D_c}{path_C}\right)\times\frac{\left|B^{LCS}\right|}{\left|B_1\backslash B^{LCS}\right|+\left|B_2\backslash B^{LCS}\right|+\left|B^{LCS}\right|}. \qquad (8)$$

Для рассматриваемого случая имеем:

$$B_1=\left\{\, m^{X_1} : m^{X_1}\in\left(\, M_1^{X_1}\cup M_2^{X_1}\cup M_3^{X_1}\,\right), I^{X_1}\left(g^{X_1}, m^{X_1}\right)=\text{true}\,\right\},$$
$$B_2=\left\{\, m^{X} : m^{X}\in\left(\, M_1^{X_1}\cup M_2^{X_1}\cup M_3^{X_1}\,\right), I^{X_1}\left(g_k^{X}, m^{X}\right)=\text{true}\,\right\},$$

где $D_c=2$, $\left(M_1^{X_1}\cup M_2^{X_1}\cup M_3^{X_1}\right)\subset M^{X_1}$, $B^{LCS}=B_1\cap B_2$.

Соответствие *Случаю 4* обычно проверяется в несколько итераций.

В ходе каждой последующей итерации число признаков, не являющихся общими для g_k^X и g^{X_1}, всегда меньше, чем в предыдущей.

Начальное значение $path_C=4$ и с каждым шагом возрастает на 1.

Пусть T^S — множество СЭ-фраз, определяющих некоторую СЯУ S.

При рассмотрении $\forall T_i \in T^S$ как множества символов справедливо:

$$T_i = T_i^C \cup T_i^F,$$

где T_i^C — общая неизменная часть всех $T_i \in T^S$, T_i^F — изменяемая часть.

Пусть W_{ij} — буквенный состав слова, j — его порядковый номер во фразе.

Тогда

$$W_{ij} = W_{ij}^C \cup W_{ij}^F,$$

где $W_{ij}^C \subset T_i^C$ — неизменная, $W_{ij}^F \subset T_i^F$ — флективная часть.

На основе попарного сравнения W_{ij} различных T_i требуется найти:

- W_{ij}^C и W_{ij}^F каждого W_{ij} при $\left|W_{ij}^C\right| \to \max$;
- синтаксическое отношение R_q, определяющее допустимость сочетания слов с буквенным составом флексий W_{ij}^F и W_{ik}^F, $k \neq j$.

Пусть J — индексное множество для неизменных частей всех слов, употребленных во всех СЭ-фразах множества T^S.

Определение 3

Моделью L линейной структуры фразы $T_i \in T^S$ назовем упорядоченную совокупность индексов $j \in J$ неизменных частей слов, входящих в T_i.

Пусть $h\left(j, L\left(T_i\right)\right)$ — позиция индекса j в модели $L\left(T_i\right)$, где $j \in J$.

Тогда множество связей относительно $L\left(T_i\right)$

$$D : T_i \to \left\{ \left(h\big(j, L(T_i)\big),\ h\big(k, L(T_i)\big) \right) :\ j \neq k \right\}.$$

Определение 4

Связь $d_{qi} = \left(h\big(j, L\left(T_i\right)\big),\ h\big(k, L\left(T_i\right)\big) \right)$ допустима для модели $L\left(T_i\right)$, если существует пара СЭ-фраз $\{T_l, T_m\} \subset T^S$ таких, что и $L\left(T_l\right)$, и $L\left(T_m\right)$ имеют подпоследовательностью либо $\{j, k\}$, либо $\{k, j\}$.

Пусть для любого $T_i \in T^S$ все $d_{qi} \in D(T_i)$ удовлетворяют *определению 4*.

Определение 5

Будем считать, что модель $L(T_i)$ проективна относительно множества синтаксических связей в T^S, если

$$\sum_{q=1}^{|D(T_i)|} \Delta_{qi} \leq \left| L(T_i) \right|, \text{ где}$$

$$\Delta_{qi} = \left| h\big(j, L(T_i)\big) - h\big(k, L(T_i)\big) \right|.$$

Замечание

Сосуществование словоформ в линейном ряду относительно $L(T_i)$ определяется синтагматическими зависимостями, которые выражаются на множестве T_i^F и задаются отношениями из множества R в составе структуры (1). Для построения множества R нужно найти совокупность моделей линейных структур фраз из T^S, отвечающих *определению 5*.

Пусть $\bigcup_i D(T_i)$ есть множество связей, допустимых для $\forall L(T_i)$: $T_i \in T^S$.

Определение 6

Множество пар (j, k), сгруппированных по некоторому общему для них индексу k, есть элемент множества V^J вершин графа синтагм (V^J, I^J). При этом множества E_1 и E_2, входящие в V^J, будут соединены ребром из I^J, если $\exists\, \{j, k, m\} \subset J$: $(j, k) \in E_1$, $(k, m) \in E_2$ и $j \neq m$.

Пусть $G^F = \left\{ f_{ij} : f_{ij} = \odot \left(W_{ij}^F \right) \right\}$, $I^F = \left\{ (f_{ij}, f_{ik}) : s\,(j, k) = \text{true} \right\}$,

где $\odot$ — последовательная конкатенация символов.

Отношение s задается рекурсивно на основе $\left(V^J, I^J\right)$ следующим образом.

1. $s\,(j_1, j_1) = \text{true}$.
2. $s\,(j_1, j_2) = \text{true}$, если выполняется одно из двух условий:
   - $\exists\, E_1 \in V^J$: $(j_1, j_2) \in E_1$, причем $\exists\, j_3 \in J : s\,(j_2, j_3) = \text{true}$;
   - $\exists\, (E_1, E_2) \in I^J : \exists\, j_3 \in J$: $(j_1, j_3) \in E_1$, $(j_3, j_2) \in E_2$, $s\,(j_3, j_2) = \text{true}$.

Отношению I^F соответствует формальный контекст *сочетаемости флексий*:

$$K^F = \left(G^F, M^F, I^F \right), \text{ в котором } M^F = G^F.$$

Пусть $W_{ij} \subset T_i$, где $T_i \in T^S$. Рассмотрим $T_i^{\odot} = \left\{w_{ij} : w_{ij} = \odot\left(W_{ij}\right)\right\}$.

Положим также, что $\exists\, T_i^P \subset T_i$, определяющее последовательность

$$P_i^{\odot} = \left\{u_k : u_k = \odot\left(W_k^P\right), \bigcup_k W_k^P = T_i^P\right\}.$$

Лемма 2

Последовательность $P_i^{\odot}$ содержит слово-предикат, если $\exists\, \{j, 0, k\} \subset L\left(T_i\right)$: $\{w_{ij}, u_1, \ldots, u_p, w_{ik}\} \subset T_i^{\odot}$, где $\{u_1, \ldots, u_p\} = P_i^{\odot}$, а $p = \left|P_i^{\odot}\right|$.

Лемма 3

Слово $u_k \in P_i^{\odot}$ входит в состав Расщепленного Предикатного Значения (РПЗ), если $\exists\, T_j \in T^S : L\left(T_j\right) \neq L\left(T_i\right)$, а $u_k \in P_j^{\odot}$. При этом $\nexists\, T_k \in T^S$, для которого $P_k^{\odot} \subset P_i^{\odot}$, а $L\left(T_k\right) \neq L\left(T_j\right)$ и $L\left(T_k\right) \neq L\left(T_i\right)$.

Пусть $P_i^{\odot\prime}$ — последовательность слов, удовлетворяющих *лемме 3*.

Теорема 2

Для построения формального контекста K^F при наличии РПЗ необходимо и достаточно найти множество $T' \subset T^S$: $T' = \left\{T_i : \left|P_i^{\odot\prime}\right| \to \max\right\}$.

Пусть $\left(V^J, I^J\right)$ — граф синтагм, J — индексное множество, на котором задаются $L\left(T_i\right) : T_i \in T^S$. Рассмотрим

$$I_1^J = \left\{(j,k) \colon \exists\, E \in V^J, (j,k) \in E\right\}.$$

Назовем $\left(V_1^J, I_1^J\right)$, $V_1^J = J$, деревом-прецедентом для T^S.

Пусть $P_i^{\odot\prime}$ — последовательность слов, удовлетворяющих *лемме 3*, а

$T' \subset T^S$ — множество, рассматриваемое *теоремой 2*.

Для $u_k \in \bigcup_i P_i^{\odot\prime} \colon T_i \in T'$ неизменная и флективная части формируются сравнением буквенного состава со всеми $u_j \in \bigcup_l P_l^{\odot} \colon T_l \in \left(T^S \setminus T'\right)$.

При этом необходимо, чтобы $2\left|W_k^C\right| > \left|W_k^F\right| + \left|W_j^F\right|$, где индексы C относятся к составу неизменной, а F — флективной части слова.

Дерево $\left(V_1^J, I_1^J\right)$ преобразуется следующим образом:

- корень изменяется с $k = 0$ на значение k для слова $u_k \in P_i^{\odot\prime}$ с максимальной встречаемостью в разных ЕЯ-фразах из T^S;
- правое поддерево перевешивается на узел j для слова $u_j \in P_i^{\odot\prime}$ наименьшей встречаемости;
- для $\forall\, \{u_l, u_m\} \subset P_i^{\odot\prime}$ дочерним будет узел слова меньшей встречаемости.

Пусть $\left(V^J, I^J\right)$ — граф синтагм, J — индексное множество, на котором задаются $L\left(T_i\right) : T_i \in T^S$, $\left(V_1^J, I_1^J\right)$ — дерево-прецедент для T^S,

$$V_1^J = J,\ I_1^J = \left\{(j,k) : \exists\, E \in V^J, (j,k) \in E\right\},$$

$K^E = \left(G^E, M^E, I^E\right)$ есть искомый формальный контекст эталона.

Если $\exists\ E \in V^J :\ (j,k) \in E$, а дерево $\left(V_1^J, I_1^J\right)$ расширено с учётом *леммы 3* и *теоремы 2*, то для основ b_j и b_k и флексий f_j и f_k элементы множеств G^E, M^E и отношения I^E формируются следующим образом.

Случай 1

Индекс k соответствует родительскому узлу, j — дочернему, линейная структура фразы не содержит предлог между словами с индексами j и k.

При этом в M^E включаются признаки $m_1 = p_{bs} \odot b_k$, $m_2 = p_{bf} \odot f_k$, $m_3 = p_{fl} \odot f_j$ и $m_4 = f_j \odot «:» \odot f_k$, основа b_j включается в множество G^E, $I^E = I^E \cup \{(b_j, m_1),\ (b_j, m_2),\ (b_j, m_3),\ (b_j, m_4)\}$.

Случай 2

Между словами с индексами j и k стоит предлог p_y.

I^E, m_1 и m_3 — аналогично *Случаю 1*, $m_2 = p_{bf} \odot f_k \odot «:» \odot p_y$, $m_4 = f_j \odot «:» \odot f_k \odot «:» \odot p_y$.

Коэффициент сжатия информации по основам относительно модели СЯУ в виде формального контекста (2) равен:

$$k^S = \frac{\sum_{i=1}^{n^{BS}} k_i^S}{n^{BS}}, \tag{9}$$

где в соответствии с принятым разбиением множества M^S

$$k_i^S = \frac{\sum_{j=1}^{n_i^{BS}} \sum_{k=1}^{n^{MF}} n_{ijk}^{AS}}{n_i^{BS}},$$

$$n^{BS} = \left|M_1\right|,$$

$$n^{MF} = \left|M_2\right|,$$

$$n_i^{BS} = \left| \left\{ g \in G^S : I^S(g,m) = \text{true},\ m \in M_1,\ m = p_{bs} \odot b_i \right\} \right|,$$

$$n_{ijk}^{AS} = \left| \left\{ m_k \in M_3 : I^S(g,m_k) = \text{true}, \right.\right.$$
$$\left.\left. \exists\ m_{bf} \in M_2 : m_{bf} = p_{bf} \odot f_k,\ m_k = b_i \odot «:» \odot f_k \right\} \right|.$$

Коэффициент сжатия по флексиям аналогичен коэффициенту по основам:

$$k^F = \frac{\sum_{i=1}^{n^{FS}} k_i^F}{n^{FS}}, \tag{10}$$

где

$$k_i^F = \frac{\sum_{j=1}^{n_i^{FS}} \sum_{k=1}^{n^{MF}} n_{ijk}^{AF}}{n_i^{FS}},$$

$$n^{FS} = \left|M_5\right|,$$

$$n^{MF} = \left|M_2\right|,$$

$$n_i^{FS} = \left|\left\{ g \in G^S : I^S(g, m) = \text{true},\ m \in M_5,\ m = p_{fl} \odot f_i \right\}\right|,$$

$$n_{ijk}^{AF} = \left|\left\{ m \in M_4 : I^S(g_j, m) = \text{true},\ \exists\, m_{bf} \in M_2 : m_{bf} = p_{bf} \odot f_k,\ m = f_i \odot «:» \odot f_k \right\}\right|.$$

Производится по максимуму сжатия информации по основам и флексиям в результирующем формальном контексте.

Утверждение 2

Признак из состава множества признаков формального контекста фразы может быть включён в множество признаков формального контекста эталона, если он входит в признаковую пятёрку $\{m_1, m_2, m_3, m_4, m_5\}$, в которой $m_1 = p_{bs} \odot b$, $m_2 = p_{bf} \odot f_1$, $m_3 = b \odot «:» \odot f_1$, $m_4 = p_{fl} \odot f_2$, $m_5 = f_2 \odot «:» \odot f_1$, а b есть основа некоторого слова. При этом основе b не должен соответствовать объект формального контекста, если есть другой объект этого же контекста, который обладает одновременно признаком m_1 и некоторым другим признаком $m = p_{bs} \odot b_1$, где $b_1 \neq b$, а основе b_1 не соответствует ни одного объекта этого формального контекста при том, что признак m относится более чем к одному объекту.

Замечание

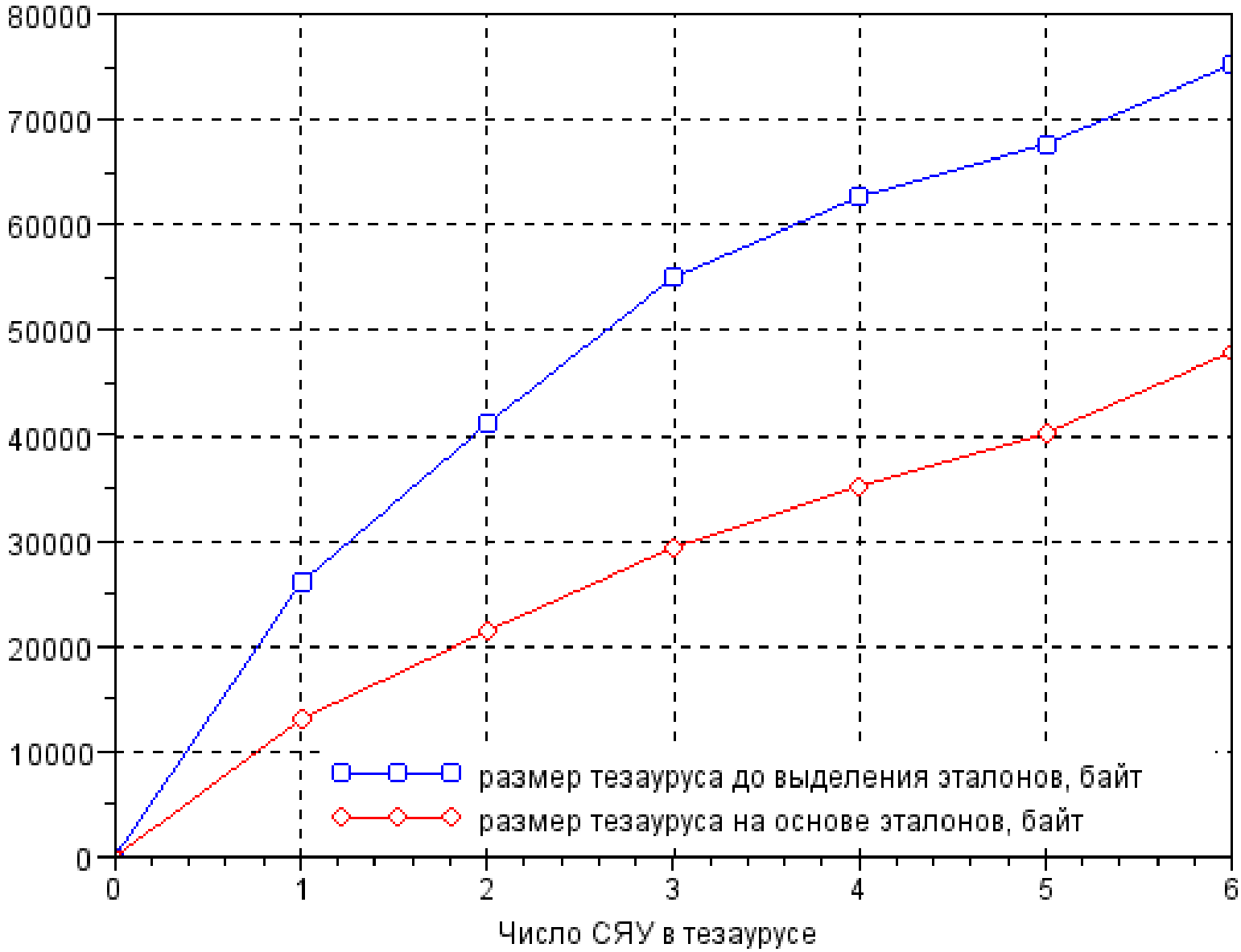
Последовательности из трех и более соподчиненных слов, встречающиеся более чем в 49% исходных СЭ-фраз, выделяются на этапе синтаксического разбора. Для каждой из них строится отдельный формальный контекст, идентичный по структуре формальному контексту эталона.

Ситуации языкового употребления и смысловые эталоны

Порядковый номер СЯУ, i	1	2	3	4	5	6
Число фраз, задающих СЯУ	54	53	26	26	2	3
из них представляют эталон	14	15	5	11	2	3
Исходное число объектов СЯУ	13	15	13	12	8	11
Исходное число признаков СЯУ	160	153	135	102	46	68
Число объектов эталона	9	12	12	12	8	11
Число признаков эталона	75	78	65	71	46	68

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

Соотношение размеров тезауруса для разного числа СЯУ



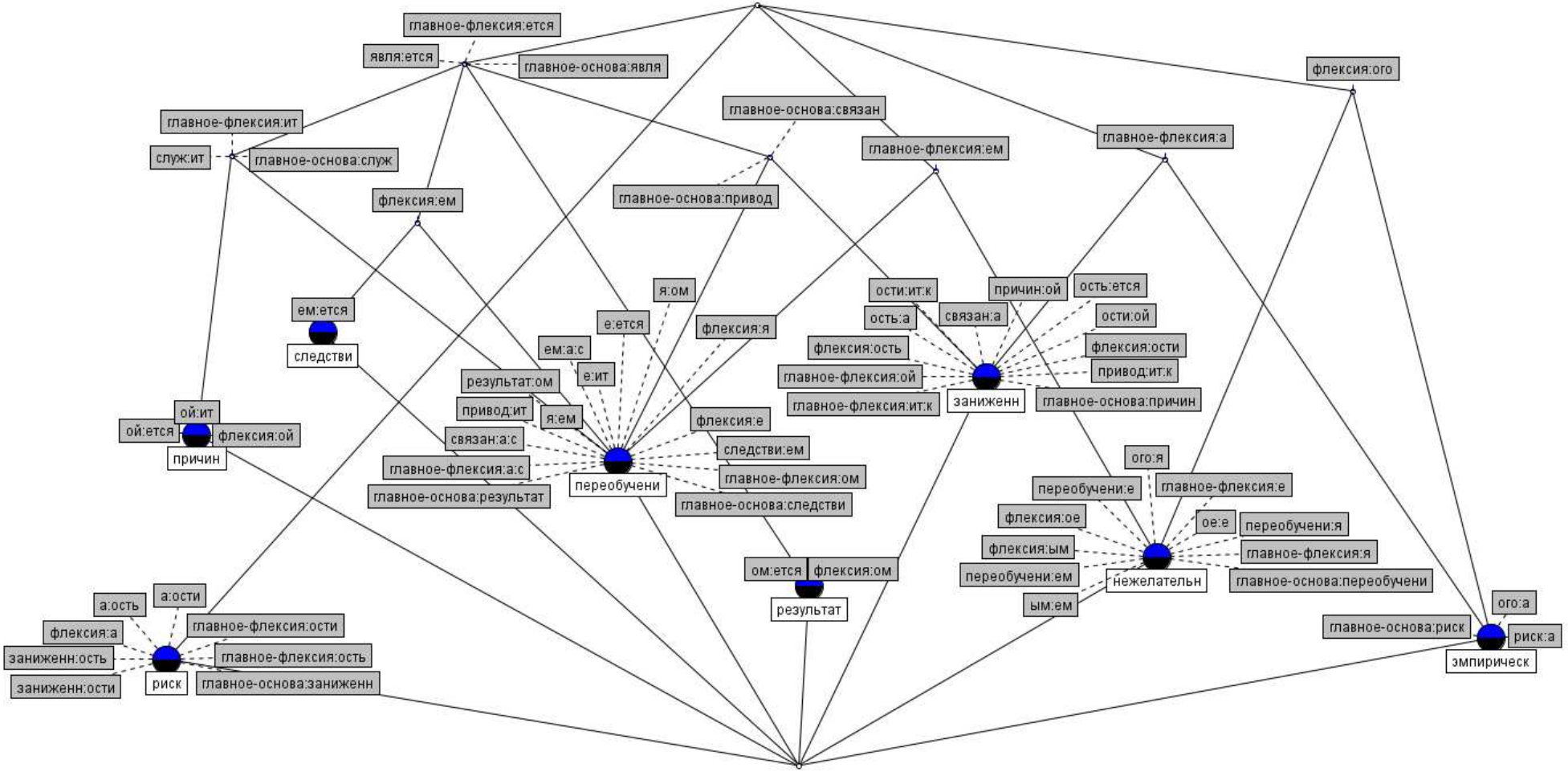
Пример: исходное множество семантически эквивалентных фраз

Синонимичные перифразы

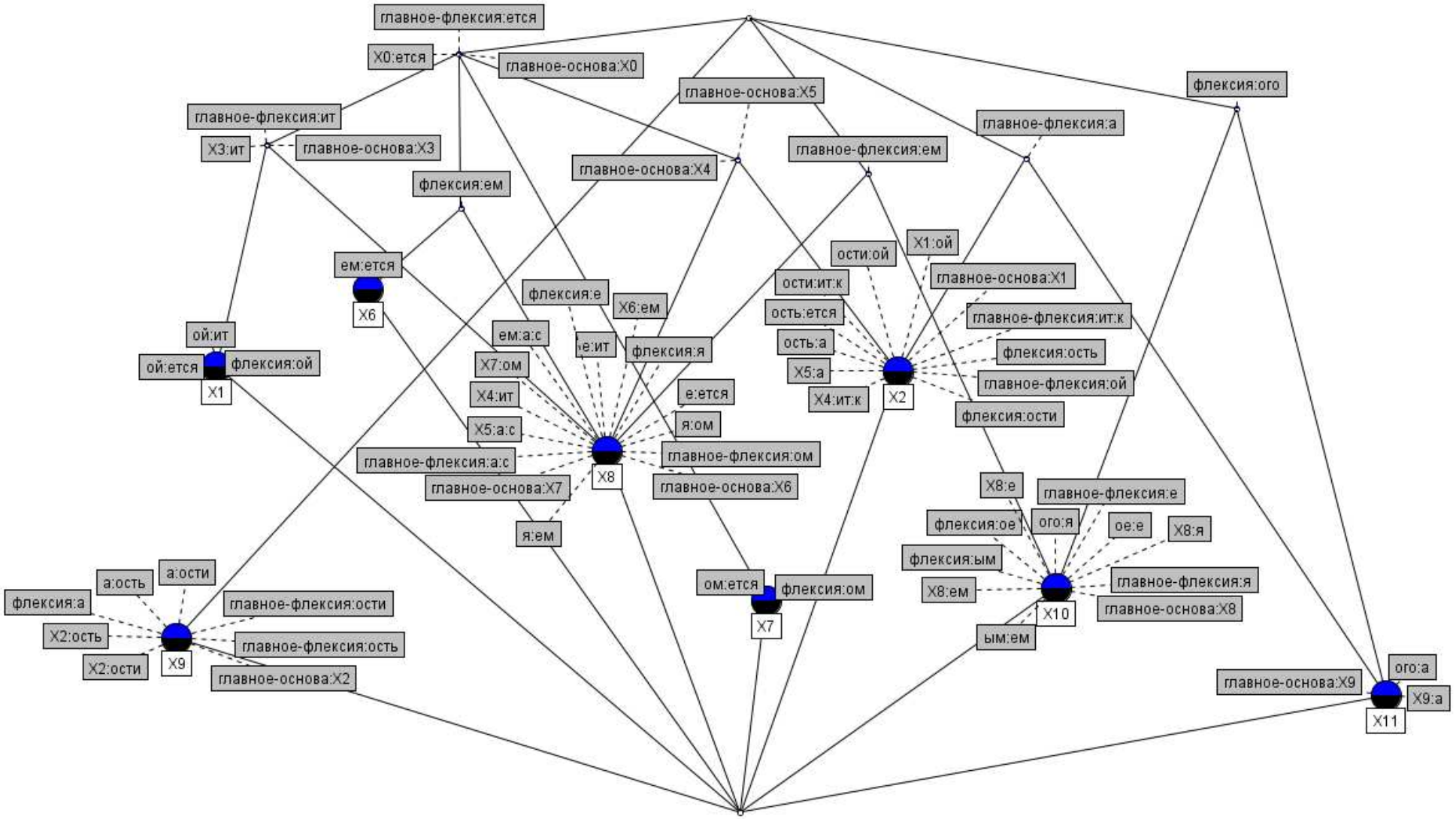
27:89 Insert Indent Modified

"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: формальный контекст смыслового эталона



Шаблон формального контекста эталона



Разметка СЭ-фраз в составе шаблона СЯУ

Разметка СЭ-фраз в составе шаблона СЯУ

11:177 Insert Indent Modified

```
[[wm("X10","ое"),wm("X8","е"),wm("X4","ит"),wm("к"," "),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X10","ое"),wm("X8","е"),wm("X6","ем"),wm("которого"," "),wm("X0","ется"),wm("X2","ость"),wm("X11","ого"),wm("X9",
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X0","ется"),wm("X6","ем"),wm("X10","ого"),wm("X8","я")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X0","ющаяся"),wm("X6","ем"),wm("X10","ого"),wm("X8","я")],
 [wm("X11","ий"),wm("X9"," "),wm("X2","ость"),wm("которого"," "),wm("X0","ется"),wm("X6","ем"),wm("X10","ого"),wm("X8",
 [wm("X11","ий"),wm("X9"," "),wm("X2","ый"),wm("вследствие"," "),wm("X10","ого"),wm("X8","я")],
 [wm("X11","ий"),wm("X9"," "),wm("к"," "),wm("X2","ости"),wm("которого"," "),wm("ведет"," "),wm("X10","ое"),wm("X8","е")],
 [wm("X9"," "),wm("X2","ый"),wm("как"," "),wm("X6","е"),wm("X8","я")],
 [wm("X11","ий"),wm("X9"," "),wm("по"," "),wm("X1","е"),wm("обусловленной"," "),wm("X10","ым"),wm("X8","ем"),wm("може
 [wm("X11","ий"),wm("X9"," "),wm("в"," "),wm("силу"," "),wm("обстоятельств"," "),wm("X5","ных"),wm("с"," "),wm("X10","ым"
                                                                                                    wm("может","
 [wm("X11","ий"),wm("X9"," "),wm("по"," "),wm("X1","е"),wm("вызванной"," "),wm("X10","ым"),wm("X8","ем"),wm("может","
 [wm("X11","ий"),wm("X9"," "),wm("к"," "),wm("X2","ости"),wm("которого"," "),wm("X4","ит"),wm("X10","ое"),wm("X8","е")],
 [wm("X10","ое"),wm("X8","е"),wm("X3","ит"),wm("X1","ой"),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X1","ой"),wm("которой"," "),wm("X0","ется"),wm("X10","ое"),wm("X8",
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X0","ется"),wm("X7","ом"),wm("X10","ого"),wm("X8","я")],
 [wm("X10","ое"),wm("X8","е"),wm("с"," "),wm("которым"," "),wm("X5","а"),wm("X2","ость"),wm("X11","ого"),wm("X9","а")],
 [wm("X11","ий"),wm("X9"," "),wm("с"," "),wm("X8","ем"),wm("X5","а"),wm("его"," "),wm("X2","ость")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X5","а"),wm("с"," "),wm("X8","ем")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X0","ющаяся"),wm("X7","ом"),wm("X10","ого"),wm("X8","я")],
 [wm("X10","ое"),wm("X8","е"),wm("X7","ом"),wm("которого"," "),wm("X0","ется"),wm("X2","ость"),wm("X11","ого"),wm("X9",
 [wm("X10","ое"),wm("X8","е"),wm("X7"," "),wm("которого"," "),wm("есть"," "),wm("X2","ость"),wm("X11","ого"),wm("X9","а")
 [wm("X10","ое"),wm("X8","е"),wm("X4","ящее"),wm("к"," "),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X10","ое"),wm("X8","е"),wm("X3","ащее"),wm("X1","ой"),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("относится"," "),wm("к"," "),wm("X6","ю"),wm("X10","ого"),wm("X8","я")]
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X5","а"),wm("с"," "),wm("X10","ым"),wm("X8","ем")],
 [wm("X10","ое"),wm("X8","е"),wm("X0","ется"),wm("X1","ой"),wm("X2","ости"),wm("X11","ого"),wm("X9","а")],
 [wm("X2","ость"),wm("X11","ого"),wm("X9","а"),wm("X1","ой"),wm("которой",""),wm("X3","ит"),wm("X10","ое"),wm("X8","е")
```

Конкретизации переменных

X_i	основа	X_i	основа	X_i	основа
X_0	явля	X_4	привод	X_8	переобучени
X_1	причин	X_5	связан	X_9	риск
X_2	заниженн	X_6	следстви	X_{10}	нежелательн
X_3	служ	X_7	результат	X_{11}	эмпирическ

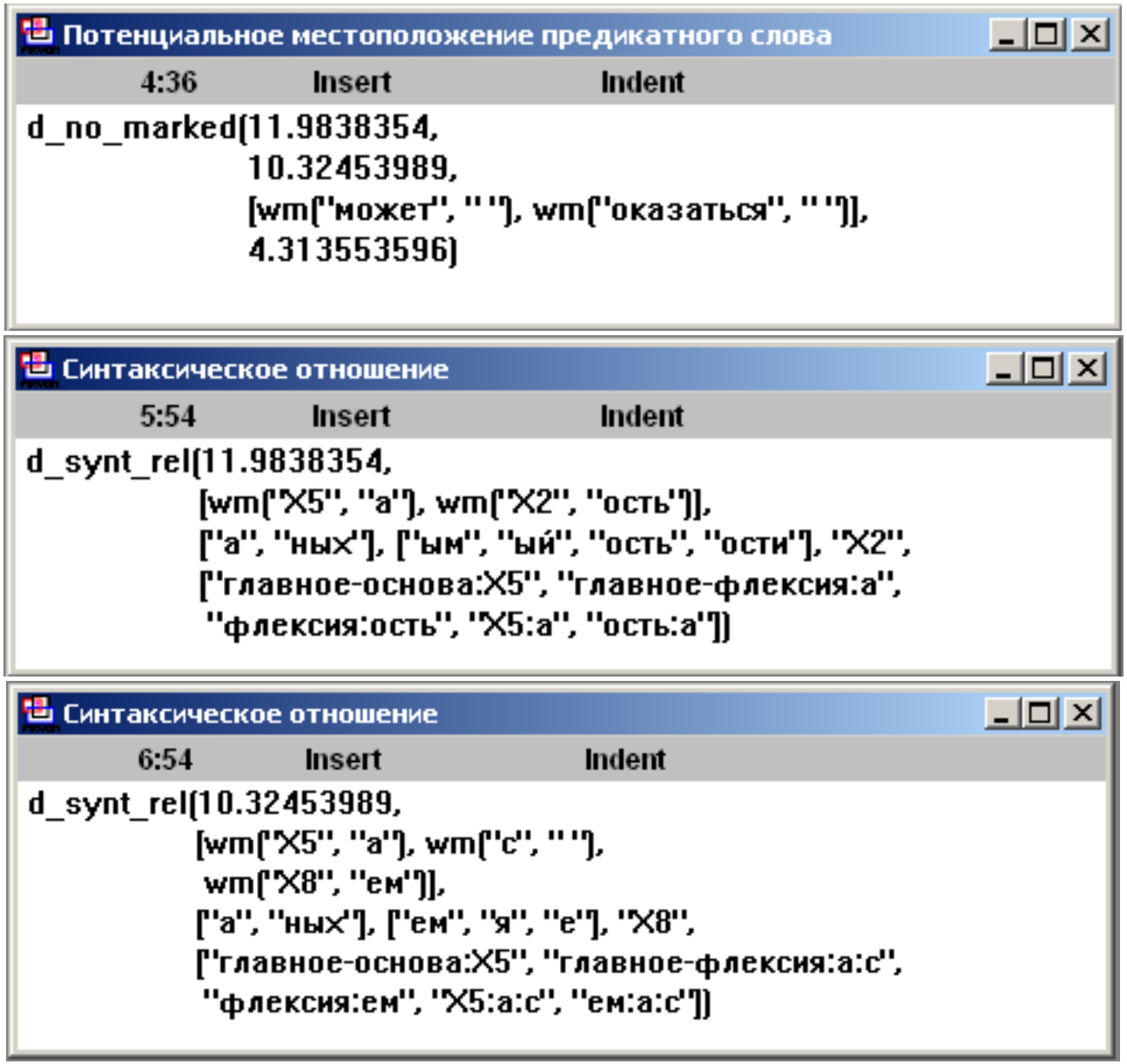
Пусть в формальном контексте эталона все обозначения основ в именах объектов и признаков заменены переменными, для каждой из которых задана конкретизация некоторой основой.

Положим аналогичные замены выполняемыми для каждой из исходных СЭ-фраз с формированием пары «основа–флексия» для каждого слова, множество последовательностей указанных пар обозначим как T^P.

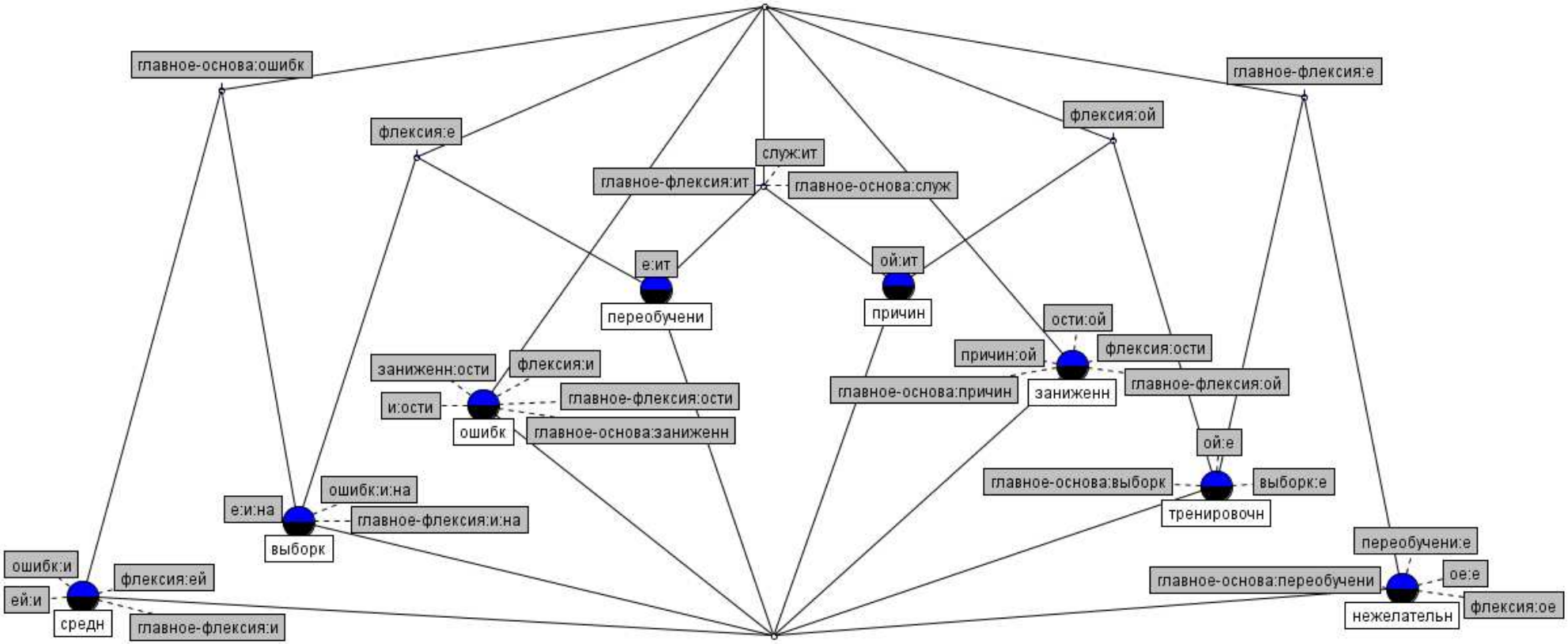
Тогда интерпретация ответа на ТЗОФ в значительном числе случаев есть «наложение» на элементы T^P, формирование списков конкретизаций и сравнение с аналогичными списками для «правильного» ответа.

Процесс происходит за линейное время, пропорциональное $\left|T^P\right|$.

Синтаксические отношения и местоположение предикатного слова



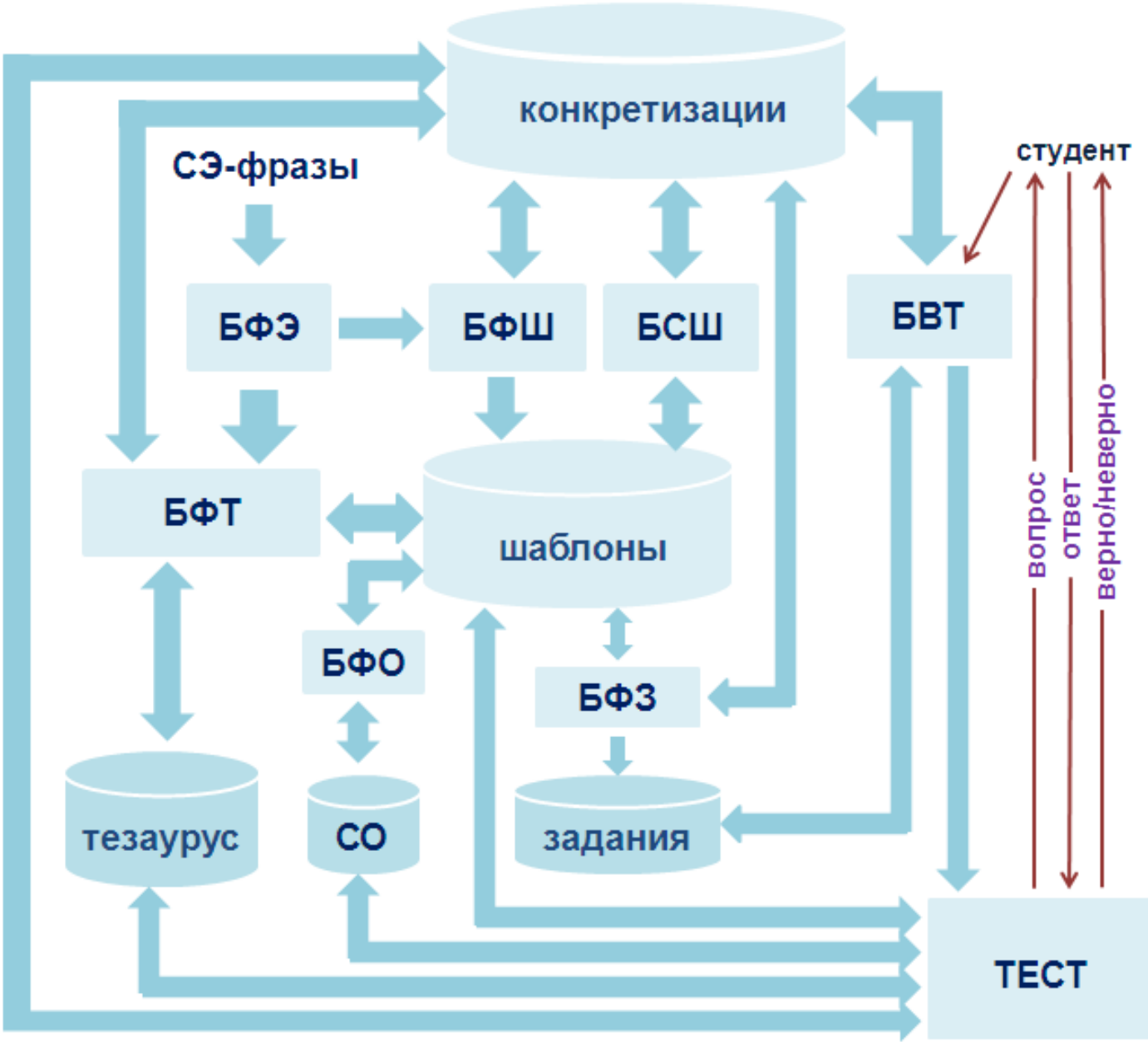
Формальный контекст смыслового эталона по результатам компиляции шаблонов двух ситуаций языкового употребления



Анализируемая фраза: «Нежелательное переобучение служит причиной заниженности средней ошибки на тренировочной выборке».

Использованы синтаксические отношения в рамках фраз: «Переусложнение модели служит причиной заниженности средней ошибки на тренировочной выборке» и «Нежелательное переобучение служит причиной заниженности эмпирического риска».

Архитектура программной системы тестирования знаний



Обозначения:

БФЭ — блок формирования эталонов;
БФШ — блок формирования шаблонов;
БСШ — блок слияния шаблонов;
БФТ — блок формирования тезауруса;
БВТ — блок выбора теста;
ТЕСТ — блок выполнения теста;
БФЗ — блок формирования заданий;
СО — база синтаксических отношений;
БФО — блок формирования отношений.

Выводы

1. Введением смыслового эталона на множестве СЭ-фраз достигается сокращение размера базы знаний для вычисления оценки схожести СЯУ в среднем на 40–50%.
2. Генерация модели смыслового эталона выделением синтагматических зависимостей на множестве СЭ-фраз максимально учитывает особенности предметно-ограниченного подмножества языка.
3. При смысловых ограничениях на перифразирование синтаксический разбор внешней программой на основе наиболее вероятных связей позволяет выделить связи «объект–признак» в рамках эталона с достаточно высокой точностью (менее 2% ошибок).
4. Модель-шаблон СЯУ составляет основу формирования стратегий и правил синтаксического анализа относительно заданного предметно-ограниченного подмножества ЕЯ. Точность построения смыслового эталона повышается задействованием синтаксических отношений, сформированных для разных СЯУ по заданной предметной области.