Формирование единиц представления предметных знаний в задаче их оценки на основе открытых тестов

Емельянов Г. М., Михайлов Д. В., Козлов А. П.

Новгородский государственный университет
имени Ярослава Мудрого

10-я Международная конференция
«Интеллектуализация обработки информации» (ИОИ-2014),

4–11 октября 2014 г.

о. Крит, Греция

Предмет и цель исследований, исследуемая проблема

Предмет исследования

Методы и алгоритмы формирования знаний о синонимии.

Исследуемая проблема

Формирование необходимого и достаточного набора признаков единицы знаний, представляемых текстами предметно-ограниченного естественного языка (ЕЯ) и оцениваемых с применением теста открытой формы.

Основная цель исследований

Разработка и теоретическое обоснование методов и алгоритмов поиска оптимального плана передачи смысла между экспертами и обучаемыми в системе контроля знаний на основе открытых тестов.

Пусть

Ts — множество семантически эквивалентных (СЭ) ЕЯ-фраз, ассоциируемое с ситуацией языкового употребления (СЯУ).

При этом

$$\mathrm{Ts} = \left\{ Ts_i : Ts_i = \bigodot_j w_{ij} \right\},$$

где $\odot$ — операция конкатенации, а w_{ij} представляется последовательностью

$$W_{ij} = Wc_{ij} \odot Wf_{ij},$$

где Wc_{ij} составляют символы неизменной части (основы) слова w_{ij};

Wf_{ij} — изменяемой (флективной) части w_{ij} относительно Ts.

Замечание

Далее отождествим с *основой* и *флексией* слова понятия «префикс» и «суффикс», принятые в информатике.

На множестве суффиксов каждой $Ts_i \in \mathrm{Ts}$ выражаются синтагматические зависимости, которые задаются синтаксическими отношениями и определяют возможность сосуществования словоформ в линейном ряду.

Представим языковой контекст, отвечающий Ts, посредством тройки

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

где $\forall g \in G$ — основа слова, синтаксически подчинённого другому слову из некоторой $Ts_i \in \mathrm{Ts}$.

Множество M включает:

- указания на основы и флексии слов, синтаксически главных по отношению к словам с основами из G;
- связи «основа–флексия» для синтаксически главного слова;
- комбинации флексий зависимых и главных слов.

Задача формирования оптимального плана передачи смысла

Требуется найти $I \subseteq G \times M$, определяющее фразы $Ts_i \in \mathrm{Ts}$ минимальной символьной длины при максимизации числа слов, наиболее употребимых в различных фразах из Ts (с учётом синонимов).

Определение 1

Единицу знаний, представляемую тройкой (1) и формируемую на основе указанных СЭ-фраз, будем отождествлять со смысловым эталоном СЯУ.

Формирование смыслового эталона: основные задачи

1. Выделение неизменяемых частей (основ) и изменяемых (флексий) для слов в составе исходного множества семантически эквивалентных фраз, определяющих ситуацию языкового употребления.
2. Формирование критерия информативности для слов в контексте ситуации языкового употребления.
3. Выделение и классификация связей слов в составе фраз, задающих ситуацию языкового употребления.

Основные проблемы

- Ориентация программ синтаксического анализа текстов на модели словосочетаний и предложений, наиболее вероятные в языке в целом без учёта особенностей предметно-ограниченных ЕЯ-подмножеств.
- Автоматизация формирования знаний о сосуществующих в заданном языковом контексте синтаксических связях с учётом их значимости для передачи требуемого смысла.

Определение 2

Пусть J — множество индексов основ слов, составляющих фразы из Ts.

Последовательность таких индексов для некоторой $Ts_i \in \mathrm{Ts}$ назовём *моделью* её *линейной структуры* (МЛС), $\mathrm{Ls}(Ts_i)$.

Пусть L есть множество моделей линейных структур фраз из Ts на J.

Определение 3

Индексы $j_1, j_2 \in J$ соответствуют словам-синонимам и могут быть заменены одним индексом из $\left(\mathbb{N} \backslash J\right)$, если $\exists \left\{\mathrm{Ls}(Ts_1), \mathrm{Ls}(Ts_2)\right\} \subseteq L$:

$$\mathrm{Ls}(Ts_1) = J_1 \odot \{j_1\} \odot J_2 \text{ и } \mathrm{Ls}(Ts_2) = J_1 \odot \{j_2\} \odot J_2.$$

Введём далее следующие обозначения:

L' — множество моделей линейных структур фраз из задающих СЯУ, преобразованное заменой индексов согласно *Определению 3*;

J' — соответствующее преобразованное индексное множество J;

$\mathrm{N}(j, L')$ — абсолютная частота встречаемости отдельного индекса в моделях линейных структур из L'.

Пусть X — последовательность упорядоченных по убыванию значений $\mathrm{N}(j, L')$ для всех $j \in J'$.

Введём обозначения для используемых далее функций.

Функция	Возвращаемое значение
$\mathrm{first}(X)$	первый элемент последовательности X
$\mathrm{last}(X)$	последний элемент последовательности X
$\mathrm{lrev}(X)$	исходная последовательность X без последнего элемента
$\mathrm{rest}(X)$	исходная последовательность X без первого элемента

Разобьём последовательность X на кластеры с применением алгоритма, содержательно близкого алгоритмам класса FOREL.

Пусть $\mathrm{mc}(H_i)$ — функция, вычисляющая центр масс кластера H_i.

При этом элементы X принадлежат одному кластеру, если

$$\begin{cases} \left|\,\mathrm{mc}(X) - \mathrm{first}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \\ \left|\,\mathrm{mc}(X) - \mathrm{last}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \end{cases}. \tag{3}$$

Функцию, выдающую $\mathrm{true}/\mathrm{false}$ в зависимости от выполнения условия (3), далее обозначим как $\mathrm{good}(X)$.

Алгоритм формирования кластера

Вход: X; // упорядоченная числовая последовательность

Выход: $H_i,\ X_p,\ X_s : X_p \odot H_i \odot X_s = X$; // $\odot$ — операция конкатенации

1: $i := 1$;
2: $H_i := X$;
3: $X_p := \varnothing$;
4: $X_s := \varnothing$;
5: **если** $\mathrm{good}(H_i) = \mathrm{true}$ **или** $|H_i| = 1$ **то**
6: **вернуть** H_i, X_p и X_s;
7: **иначе если** $\left|\mathrm{mc}(H_i) - \mathrm{first}(H_i)\right| > \left|\mathrm{mc}(H_i) - \mathrm{last}(H_i)\right|$ **то**
8: $X_p := \left\{\mathrm{first}(H_i)\right\} \odot X_p$;
9: $H_i := \mathrm{rest}(H_i)$;
10: перейти к шагу 5;
11: **иначе если** $\left|\mathrm{mc}(H_i) - \mathrm{first}(H_i)\right| < \left|\mathrm{mc}(H_i) - \mathrm{last}(H_i)\right|$ **то**
12: $X_s := \left\{\mathrm{last}(H_i)\right\} \odot X_s$;
13: $H_i := \mathrm{lrev}(H_i)$;
14: перейти к шагу 5;
15: **иначе**
16: $X_s := \left\{\mathrm{last}(H_i)\right\} \odot X_s$; // *Для разбивки исходной последовательности*
17: $X_p := \left\{\mathrm{first}(H_i)\right\} \odot X_p$; // *на кластеры данный алгоритм применяется*
18: $Tmp := \mathrm{lrev}(H_i)$; // *рекурсивно к X_p и X_s на его выходе.*
19: $H_i := \mathrm{rest}(Tmp)$; // *Указанный процесс продолжается*
20: перейти к шагу 5; // *до тех пор, пока на очередном шаге X_p и X_s*
21: **конец если** // *не окажутся пустыми.*

Пусть $H_1, \ldots, H_r$ — кластеры, на которые разбита последовательность X, причём для $\forall\, i \neq j$ верно то, что

$$H_i \cap H_j = \varnothing, \text{ а } H_1 \odot H_2 \odot \ldots \odot H_r = X.$$

Обозначим далее множество $\left\{j : \mathrm{N}(j, L') \in H_1\right\}$ как Cl («частые» индексы).

Утверждение 1

Смысловой эталон СЯУ, задаваемой множеством СЭ-фраз Ts, определяют те $Ts_i \in \mathrm{Ts}$, модели линейных структур которых отвечают следующему условию:

$$Cl \cap \mathrm{Ls}'(Ts_i) = Cl, \text{ а } \left|\, \mathrm{Ls}'(Ts_i) \setminus Cl \,\right| \to \min,$$

где $\mathrm{Ls}'(Ts_i) \in L'$, а L' — множество моделей линейных структур фраз, формируемое согласно *Определению 3*.

Данное условие *необходимо, но не достаточно* для отнесения некоторой $Ts_i \in \mathrm{Ts}$ к фразам, определяющим смысловой эталон заданной СЯУ.

Основные проблемы

- *Утверждение 1* затрагивает исключительно лексику отбираемых фраз исходного множества, не принимая во внимание связи слов.
- При отборе «эталонных» фраз не учитывается синонимия, которая охватывает одновременно и синтаксические связи, и лексику (ср. *«Нежелательная переподгонка приводит к заниженности эмпирического риска»* $\Leftrightarrow$ *«Заниженность эмпирического риска является следствием нежелательной переподгонки»*).

Требуется

Выделить связи слов в составе СЭ-фраз исходного множества и оценить значимость найденных связей для передачи единицы знаний в рамках ситуации языкового употребления.

Возможные пути решения

- Изучение статистики встречаемости пар соседних слов в тексте [Яндекс, 2009].
- Вычисление смысловой близости слов на основе их совместной встречаемости в минимальном лексико-синтаксическом контексте, определяемом шаблоном из конечного множества [Панченко А. И., 2012].

Пусть $h\big(j, \mathrm{Ls}(Ts_i)\big)$ — позиция индекса j в модели $\mathrm{Ls}(Ts_i)$. Тогда множество синтагматических связей для $\mathrm{Ls}(Ts_i)$ определяется как

$$D : Ts_i \to \Big\{ \Big(h\big(j, \mathrm{Ls}(Ts_i)\big),\ h\big(k, \mathrm{Ls}(Ts_i)\big) \Big) : j \neq k \Big\}. \tag{4}$$

Замечание

Пара (j, k) содержательно соответствует либо некоторому *словосочетанию* в составе Ts_i, либо *грамматической основе* этой фразы.

Определение 4

Пусть $\mathrm{len}(j, k) = \big| h\big(j, \mathrm{Ls}(Ts_i)\big) - h\big(k, \mathrm{Ls}(Ts_i)\big) \big|$.

Назовём указанную величину *длиной связи*, соответствующей паре (j, k), относительно модели $\mathrm{Ls}(Ts_i)$.

Введём в рассмотрение абсолютные частоты:

$\mathrm{N}\big((j, k), L'\big)$ — встречаемости связи (j, k) в моделях линейных структур из L' независимо от $\mathrm{len}(j, k)$;

$\mathrm{N}\big(\mathrm{len}(j, k), L'\big)$ — встречаемости связи (j, k), имеющей длину $\mathrm{len}(j, k)$, в моделях линейных структур из L'.

Будем оценивать «силу» связи слов, отвечающих индексам j и k, (вне зависимости от взаимного расположения в линейном ряду фразы) посредством следующей *весовой функции:*

$$\mathrm{Wg}\big((j,k),L'\big) = \mathrm{N}\big((j,k),L'\big)\,\frac{\mathrm{N}\big((j,k),L'\big)}{\mathrm{N}(j,L')+\mathrm{N}(k,L')-\mathrm{N}\big((j,k),L'\big)}. \qquad (5)$$

Пусть X^W — упорядоченная по убыванию последовательность значений функции (5) для индексных пар (j,k), выделенных на моделях из L'.

Разобьём X^W на кластеры

$$H_1^W,\ldots,H_q^W : H_1^W \odot H_2^W \odot \ldots \odot H_q^W = X^W$$

с применением алгоритма, представленного на слайде 8.

При этом связи, максимально значимые для формирования оптимального плана передачи смысла заданной СЯУ, будут иметь значения функции (5), вошедшие в кластер H_1^W (далее назовём такие связи «весомыми»).

Обозначим далее множество индексов в составе указанных связей как Cl_1 (по аналогии с Cl из Утверждения 1).

Основные проблемы

- Не предполагая никаких гипотез относительно смысловой связи слов, соответствующих индексам j и k, оценка (5) зависит от частоты встречаемости каждого из них в анализируемых СЭ-фразах.
- «Весомыми» будет связи только между словами, рассматриваемыми *Утверждением 1* в качестве основы отбора «эталонных» фраз.

Возможный путь решения

Выделение лексико-синтаксических шаблонов на основе n-грамм [Bollegala D., 2007].

Недостаток: требуются априори известные возможные значения n (длины выделяемой последовательности от j-го до k-го слова).

Гипотеза

Из связей, не вошедших в «весомые», наименьший разброс длины имеют связи, затрагивающие вершины синтаксических деревьев.

Замечание

При наличии свободного порядка слов во фразе обратное утверждение верно не всегда.

Рассмотрим связи, значения функции (5) которых не вошли в H_1^W.

Разобьём их на кластеры (слайд 8) по величине среднеквадратического отклонения длины связи (СКОДС) относительно L'.

По определению СКОДС для пары (j,k) относительно L' вычисляется по формуле

$$\sigma\big(\mathrm{len}(j,k),L'\big)=\sqrt{\mathsf{E}\big(\mathrm{len}^2(j,k),L'\big)-\mathsf{E}^2\big(\mathrm{len}(j,k),L'\big)},$$

где $\mathsf{E}\big(\mathrm{len}(j,k),L'\big)$ — математическое ожидание длины связи,

$$\mathsf{E}\big(\mathrm{len}(j,k),L'\big)=\sum_i\left(\frac{\mathrm{N}\big(\mathrm{len}_i(j,k),L'\big)}{\mathrm{N}\big((j,k),L'\big)}\mathrm{len}_i(j,k)\right)=$$

$$=\sum_i\Big(p\big(\mathrm{len}_i(j,k),L'\big)\mathrm{len}_i(j,k)\Big).$$

Гипотеза

Индекс $j\in J'$, соответствующий вершине, входит в одну из связей кластера наименьших СКОДС и одновременно в связь из некоторого другого кластера по указанной величине. При этом «индекс вершины» не входит в связи со значениями функции (5) из H_1^W.

Достаточное условие отнесения фразы к определяющим эталон

Пусть Cl_2 — множество кандидатов на роль вершин деревьев фраз из Ts.

Утверждение 2

Смысловой эталон СЯУ, представляемой тройкой (1), определяют те $Ts_i \in \mathrm{Ts}$, для которых помимо условия Утверждения 1 верно то, что

$$\left| \left(\mathrm{Ls}'(Ts_i) \setminus Cl \right) \setminus \left(Cl_1 \cup Cl_2 \right) \right| \to \min$$

при минимальной длине суффикса для $\forall\, w_{ij} : \bigodot\limits_j w_{ij} = Ts_i$.

Обозначим множество фраз $Ts_i \in \mathrm{Ts}$, отобранных согласно условиям Утверждений 1 и 2, как Ts^*. Пусть

$$R_J = \left\{ \left((j,k),\ \mathrm{Dir} \right) : \mathrm{Dir} \in \{\leftarrow, \rightarrow\}, \exists\, Ts_i \in \mathrm{Ts}^* :\ \{j,k\} \subset \mathrm{Ls}'(Ts_i) \right\}, \quad (6)$$

причём если $X^W \neq H_1^W$ и $|\,\mathrm{Ts}^*| > 1$, то либо $\left(\{j,k\} \cap Cl_2 \right) \neq \varnothing$, либо паре (j,k) соответствует связь со значением функции (5) из кластера H_1^W.

При этом связи из R_J задают минимальные семантико-синтаксические текстовые единицы в рамках оптимального плана передачи смысла СЯУ.

Исходные данные для проведения экспериментов

Порядковый номер СЯУ, i	1	2	3	4	5	6
Число СЭ-фраз, задающих СЯУ	56	28	29	30	6	10
Минимальное число слов во фразе	5	8	11	10	10	11
Максимальное число слов во фразе	12	15	16	18	17	14

i	Фразы максимальной длины из определяющих СЯУ
1	*Нежелательная переподгонка является причиной заниженности средней 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	*Оцениваемая частота, с которой алгоритм допускает ошибку на выборке, рассматриваемой как контрольная, может оказаться заниженной по причине переподгонки.*
5	*Распознавание обладает таким свойством, что его ошибка будет иметь заниженную оценку при неудачном выборе правила принятия решений.*
6	*Рост числа базовых классификаторов, который ведет к практически неограниченному увеличению обобщающей способности композиции алгоритмов.*

Выделение основ и флексий для слов в рамках СЯУ: ключевые процедуры и функции алгоритма

- $\mathrm{pref.show}\,(w_{ij})$ возвращает текущее значение префикса слова w_{ij};
- $\mathrm{pref.inc}\,(w_{ij})$ увеличивает длину префикса слова w_{ij} на 1;
- prefs объединяет словоформы в группы (списки) по сходству префикса, сортируя их при этом по убыванию длины;
- $\mathrm{pref.check}\,(\mathrm{Prf})$ 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.

Утверждение 3

Суммарная длина общих префикса и суффикса пары слов здесь должна составлять *минимум треть* длины слова, а разность длин у пары слов с общим префиксом (независимо от суффикса) всегда *меньше половины длины меньшего слова*.

Алгоритм выделения основ и флексий для слов в рамках СЯУ

Вход: Ts;

Выход: $\mathrm{Pw} = \bigcup_{i=1}^{|\mathrm{Ts}|} Pw_i$; // $Pw_i = \left\{ \left(Wc_{ij}, Wf_{ij} \right) : Wc_{ij} \odot Wf_{ij} = W_{ij} \right\}$

1: $\mathrm{Pw} := \varnothing$; // W_{ij} — последовательность символов слова w_{ij}

2: **для всех** W_{ij}: $\bigodot_j w_{ij} = Ts_i$, где $Ts_i \in \mathrm{Ts}$

3: $Wc_{ij} := \{W_{ij}[1]\}$; $Wf_{ij} := \bigodot_{k=2}^{|W_{ij}|} W_{ij}[k]$;

4: **конец для** // инициализации основ и флексий

5: prefs (PrfsTmp);

6: **если** $\mathrm{PrfsTmp} = \varnothing$ **то**

7: **вернуть** Pw и выйти из алгоритма;

8: **иначе**

9: взять очередной Prf из PrfsTmp;

10: **если** pref.check (Prf) = true **то**

11: $\mathrm{Pw} := \mathrm{Pw} \cup \left\{ \left(\mathrm{Prf}, Wf_{ij}(\mathrm{Prf}) \right) \,\middle|\, \mathrm{pref.show}(w_{ij}) = \mathrm{Prf} \right\}$;

12: $\mathrm{PrfsTmp} := \mathrm{PrfsTmp} \setminus \{\mathrm{Prf}\}$;

13: перейти к шагу 6;

14: **иначе**

15: **для всех** $w_{ij} : \mathrm{pref.show}(w_{ij}) = \mathrm{Prf}$

16: pref.inc (w_{ij});

17: **конец для**

18: перейти к шагу 5

19: **конец если**

20: **конец если**

Порядковый номер СЯУ	1	2	3	4	5	6
Число связей со значениями функции (5), вошедшими в кластер H_1^W	4	9	14	11	6	13
Число найденных кластеров по СКОДС	5	6	5	7	1	5
Число фраз, представляющих эталон СЯУ	12	7	8	11	2	1
Общее число связей в рамках эталона СЯУ	26	28	39	43	12	19
в том числе истинных	21	17	23	24	10	14
ложных	5	11	16	19	2	5

Замечание

Для каждой найденной связи её *направление* здесь задаётся экспертом, причём только для связей, определенных им как *истинные*.

Совокупные знания системы по синтагматическим связям в рамках отдельной СЯУ могут быть представлены *булевым вектором*

$$\left(d_1, \ldots, d_k, \overline{d}_{k+1}, \ldots, \overline{d}_n\right), \tag{7}$$

где $d_1, \ldots, d_k$ соответствуют *истинным*, а $\overline{d}_{k+1}, \ldots, \overline{d}_n$ — *ложным* связям.

Программная реализация и результаты экспериментов

Пример: исходное множество семантически эквивалентных фраз

Исходное множество семантически эквивалентных фраз

28:1 Insert Indent

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 заниженны
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.
Нежелательная переподгонка приводит к заниженности эмпирического риска.

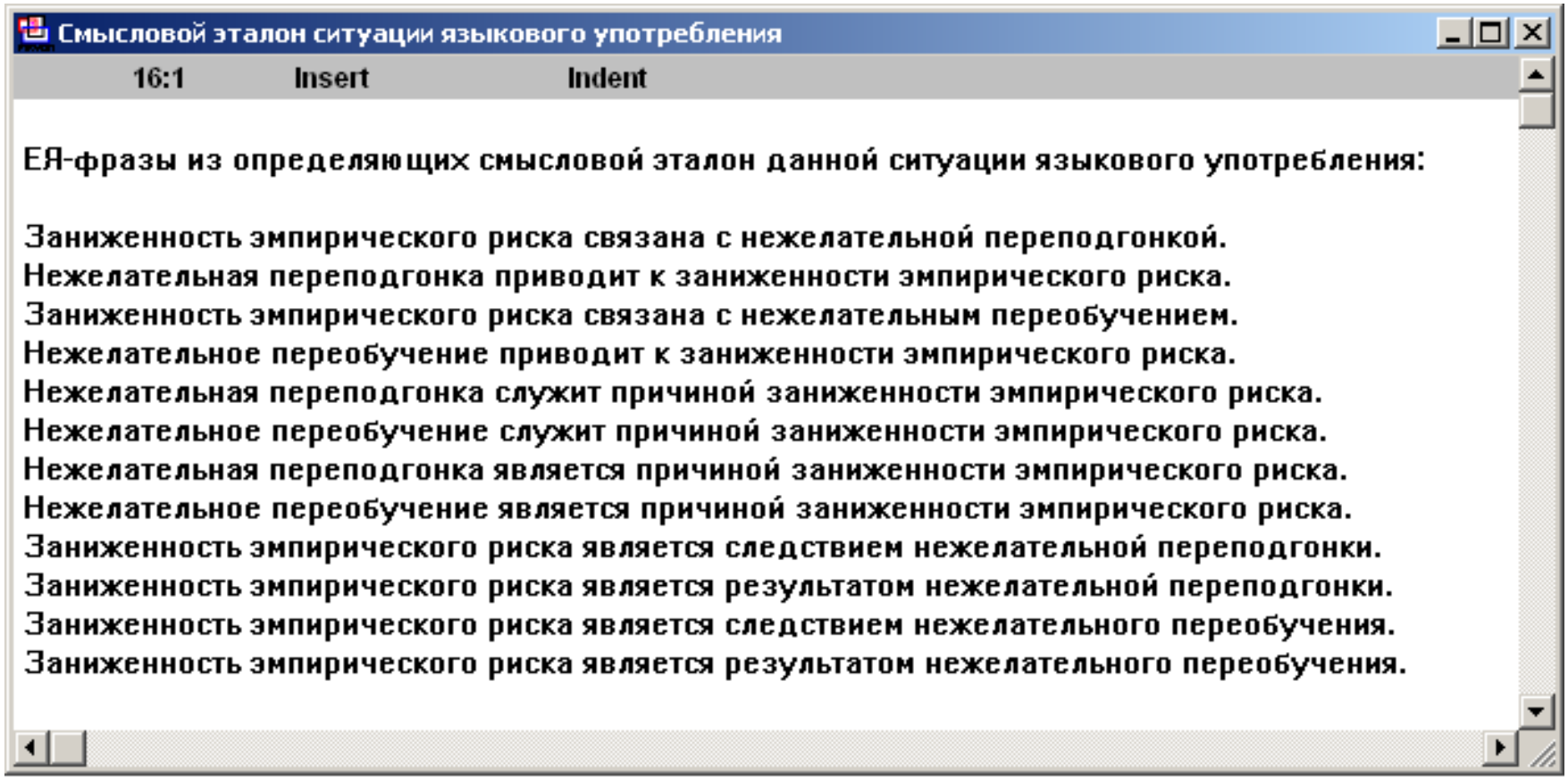
Исходные СЭ-фразы (продолжение) и фраза максимальной длины

Исходное множество семантически эквивалентных фраз

57:1 Insert Indent Modified

Заниженность эмпирического риска является следствием нежелательной переподгонки.
Заниженность эмпирического риска, являющаяся следствием нежелательной переподгонки.
Эмпирический риск, заниженность которого является следствием нежелательной переподгонки.
Эмпирический риск, заниженный вследствие нежелательной переподгонки.
Эмпирический риск, к заниженности которого ведет нежелательная переподгонка.
Риск, заниженный как следствие переподгонки.
Эмпирический риск по причине, обусловленной нежелательной переподгонкой, может оказаться заниженным.
Эмпирический риск в силу обстоятельств, связанных с нежелательной переподгонкой, может оказаться заниженным
Эмпирический риск по причине, вызванной нежелательной переподгонкой, может быть заниженным.
Эмпирический риск, к заниженности которого приводит нежелательная переподгонка.
Нежелательная переподгонка служит причиной заниженности эмпирического риска.
Заниженность эмпирического риска, причиной которой является нежелательная переподгонка.
Заниженность эмпирического риска является результатом нежелательной переподгонки.
Нежелательная переподгонка, с которой связана заниженность эмпирического риска.
Эмпирический риск, с переподгонкой связана его заниженность.
Заниженность эмпирического риска связана с переподгонкой.
Заниженность эмпирического риска, являющаяся результатом нежелательной переподгонки.
Нежелательная переподгонка, результатом которой является заниженность эмпирического риска.
Нежелательная переподгонка, результат которой есть заниженность эмпирического риска.
Нежелательная переподгонка, приводящая к заниженности эмпирического риска.
Нежелательная переподгонка, служащая причиной заниженности эмпирического риска.
Заниженность эмпирического риска относится к следствию нежелательной переподгонки.
Заниженность эмпирического риска связана с нежелательной переподгонкой.
Нежелательная переподгонка является причиной заниженности эмпирического риска.
Заниженность эмпирического риска, причиной которой служит нежелательная переподгонка.
Нежелательная переподгонка является причиной заниженности средней величины ошибки алгоритма на обучающей выборке.
Нежелательная переподгонка, следствием которой является заниженность эмпирического риска.

Результат: ЕЯ-фразы смыслового эталона



Основы слов для «частых» индексов моделей линейных структур:

«нежелательн» *«риск»* *«заниженн»*

«переподгонк/переобучени» *«эмпирическ»*

Кластеризация по значению весовой функции связи

Связи со значениями весовой функции из кластера H_1^W:

Основа для j	Основа для k	$\mathrm{Dir}(j,k)$	$\mathrm{Wg}\big((j,k),L'\big)$
заниженн	риск	$\rightarrow$	16,0556
эмпирическ	риск	$\leftarrow$	15,0588
заниженн	эмпирическ	$\rightarrow$	14,2222
нежелательн	переподгонк, переобучени	$\leftarrow$	12,5000

Связи со значениями весовой функции, не вошедшими в кластер H_1^W:

Основа для j	Основа для k	Dir	Wg	$\sigma\big(\mathrm{len}(j,k),L'\big)$
нежелательн	результат, следстви	$\leftarrow$	1,0000	0,4330
результат, следстви	есть, явля, служ	$\leftarrow$	1,1250	0,4714
есть, явля, служ	причин	$\rightarrow$	1,1250	0,4714
с	нежелательн	$\rightarrow$	0,5294	0,4714
переподгонк, переобучени	результат, следстви	$\leftarrow$	1,3889	0,4899
с	переподгонк, переобучени	$\rightarrow$	1,3889	0,4899
связан	с	$\rightarrow$	5,0000	0,4899
переподгонк, переобучени	привод, ведет	$\leftarrow$	0,2222	0,5000
нежелательн	привод, ведет	$\leftarrow$	0,2667	0,5000
связан	переподгонк, переобучени	$\rightarrow$	1,3889	0,8000
переподгонк, переобучени	есть, явля, служ	$\leftarrow$	2,0000	0,8975
нежелательн	есть, явля, служ	$\leftarrow$	2,4000	0,8975
привод, ведет	к	$\rightarrow$	1,3333	1,0000
есть, явля, служ	заниженн	$\rightarrow$	2,0000	1,2583
заниженн	к	$\leftarrow$	0,5000	1,4142
заниженн	связан	$\leftarrow$	1,3889	1,6733
причин	заниженн	$\rightarrow$	1,3889	2,2450

Кластеризация по среднеквадратическому отклонению длины связи

Кластеры, выделенные по значению среднеквадратического отклонения длины связи:

№ кластера	1	2	3	4	5
Число связей, вошедших в кластер	36	10	8	5	1
Значение СКОДС для связи					
минимальное	0,0000	0,4330	0,7071	1,2000	2,2450
максимальное	0,0000	0,5000	1,0954	1,6733	2,2450

Связи в рамках эталона СЯУ из группируемых по значению СКОДС:

Основа для j	Основа для k	Dir	$\sigma\left(\mathrm{len}(j,k), L'\right)$	№ кластера
нежелательн	результат, следстви	←	0,4330	2
результат, следстви	есть, явля, служ	←	0,4714	2
есть, явля, служ	причин	→	0,4714	2
с	нежелательн	→	0,4714	2
переподгонк, переобучени	результат, следстви	←	0,4899	2
с	переподгонк, переобучени	→	0,4899	2
связан	с	→	0,4899	2
переподгонк, переобучени	привод, ведет	←	0,5000	2
нежелательн	привод, ведет	←	0,5000	2
связан	переподгонк, переобучени	→	0,8000	3
переподгонк, переобучени	есть, явля, служ	←	0,8975	3
нежелательн	есть, явля, служ	←	0,8975	3
привод, ведет	к	→	1,0000	3
есть, явля, служ	заниженн	→	1,2583	4
заниженн	к	←	1,4142	4
заниженн	связан	←	1,6733	4
причин	заниженн	→	2,2450	5

Связи из кластера наименьших значений СКОДС, затрагивающие потенциальные вершины синтаксических деревьев

№ п/п	Основа для j	Основа для k	$\mathrm{E}\big(\mathrm{len}(j,k),L'\big)$	$\sigma\big(\mathrm{len}(j,k),L'\big)$
1	котор	привод, ведет	1,0000	0,0000
2	с	котор	1,0000	0,0000
3	связан	его	1,0000	0,0000
4	котор	связан	1,0000	0,0000
5	обстоятельств	связан	1,0000	0,0000
6	к	результат, следстви	1,0000	0,0000
7	как	результат, следстви	1,0000	0,0000
8	относится	к	1,0000	0,0000
9	причин	котор	1,0000	0,0000
10	причин	вызванной	1,0000	0,0000
11	по	причин	1,0000	0,0000
12	причин	обусловленной	1,0000	0,0000
13	котор	есть, явля, служ	1,0000	0,0000

Выделенные кандидаты на роль вершин синтаксических деревьев:

«привод/ведет» *«связан»* *«с»* *«есть/явля/служ»*
«результат/следстви» *«причин»* *«к»* *«котор»*

Содержательная интерпретация (примеры) для связей из кластера наименьших значений СКОДС, затрагивающих потенциальные вершины

- сочетание сказуемого в составе определительного придаточного с союзным словом, например: *«котор — привод/ведет»* (связь № 1), ср. *«Эмпирический риск, к заниженности которого приводит нежелательное переобучение»*;
- совокупность сочетаний слов и предлога в рамках предложной связи: *«относится — к»* (связь № 8) и *«к — результат/следстви»* (связь № 6), ср. *«Заниженность эмпирического риска относится к следствию нежелательного переобучения»*;
- сочетание слов с целью выразить определённый логический акцент, например: *«связан — его»* (связь № 3), ср. *«Эмпирический риск, с переобучением связана его заниженность»*.

Замечание

К последнему случаю может быть отнесено сочетание определяемого слова и причастия в составе оборота, если он стоит после определяемого слова и его положение не изменяется при перифразировании, например: *«причин — обусловленной»* (связь № 12), *«причин — вызванной»* (связь № 10), ср. *«Эмпирический риск по причине, обусловленной/вызванной нежелательным переобучением/нежелательной переподгонкой, может оказаться заниженным»*.

Найденные синтаксические связи в рамках смыслового эталона

Связи в рамках смыслового эталона:

заниженн —> риск
риск —> эмпирическ
заниженн —> эмпирическ
переподгонк,переобучени —> нежелательн
результат,следстви —> нежелательн
есть,явля,служ —> результат,следстви
есть,явля,служ —> причин
с —> нежелательн
результат,следстви —> переподгонк,переобучени
с —> переподгонк,переобучени
связан —> с
привод,ведет —> переподгонк,переобучени
привод,ведет —> нежелательн
связан —> переподгонк,переобучени
есть,явля,служ —> переподгонк,переобучени
есть,явля,служ —> нежелательн
привод,ведет —> к
есть,явля,служ —> заниженн
к —> заниженн
связан —> заниженн
причин —> заниженн

Ложные связи:

риск — причин
риск — к
есть,явля,служ — риск
риск — связан
риск — с

- Число кластеров, выделенных по значению СКОДС и используемых для определения потенциальных вершин синтаксических деревьев, должно быть минимум два.
- Зависимость возможного числа семантически эквивалентных фраз, определяющих СЯУ и составляющих обучающую выборку, от числа возможных синонимов на лексическом и синтаксическом уровне в рассматриваемом предметно-ограниченном ЕЯ-подмножестве.

Замечание

Типы смысловых связей слов в рамках отдельной фразы из задающих СЯУ изначально не оговариватся и для полноты учёта её смыслового контекста, определяемого тройкой (1), ограниченного набора известных семантических отношений и форм их выражения в текстах, как правило, недостаточно.

Контр-пример: поиск смысловых связей системой «Серелекс»

Лемма	Общее число связей	Английский эквивалент	Общее число связей
эмпирический	0	empiric	4
риск	25	risk	2197
нежелательный	0	undesirable	107
переподгонка, переобучение	0	overfitting	0
заниженность	0	underestimate	8
приводить (к)	0	(to) result (in)	2557
к	406	in	183
связанный (с)	0	relate(d) (to, with)	0
с	1184	to, with	0
причина	145	reason	2728
результат	52	result	2557
следствие	7	result	2557
являться	0	(to) be	0
служить	13	(to) be	0

Использованные коллекции документов:

- заголовки статей Википедии ($2{,}026 \cdot 10^9$ словоформ, 3 368 147 лемм);
- текстовый корпус ukWaC ($0{,}889 \cdot 10^9$ словоформ, 5 469 313 лемм).

Найдены связи: *«risk — result»*, *«risk — reason»* (ложная) и *«риск — с»* (ложная).

№ СЯУ	1	2	3	4	5	6
l_1	56	28	29	30	6	10
n_1	12	15	16	18	17	14
l_2	12	7	8	11	2	1
n_2	7	10	12	13	10	13
vol_1	672	420	464	540	102	140
vol_2	84	70	96	143	20	13

Здесь:

n_1 — максимальное число слов во фразе по СЯУ в целом;

n_2 — во фразе из определяющих эталон;

$\mathrm{vol}_1 = n_1 \cdot l_1$ есть оценка сверху, l_1 — число СЭ-фраз, задающих СЯУ;

$\mathrm{vol}_2 = n_2 \cdot l_2$ есть оценка снизу, l_2 — число СЭ-фраз, определяющих эталон СЯУ.

Выводы

1. *Ключевая особенность* изложенной методики формирования единиц представления экспертных знаний для разработки открытых тестов — *выделение лексоко-синтаксических связей слов* во фразе в рамках СЯУ *без привлечения внешних синтаксических анализаторов*.
2. *Предложенная методика позволяет выделять шаблоны* лексико-синтаксических *связей, необходимых и достаточных* для передачи заданного смысла в предметно-ограниченном ЕЯ-подмножестве.
3. Ассоциация выделяемых связей *с компонентами булева вектора составляет основу* обучения синтаксического анализатора на ситуациях употребления предметно-ограниченного ЕЯ-подмножества.
4. *Экспериментальное подтверждение* соответствия рациональной передаче смысла — *безошибочность разбора* «эталонной» фразы парсером, ориентированным на наиболее вероятные в языке модели предложений при единственности компоненты связности графа разбора.
5. Предложенный метод выделения смысловых эталонов даёт *минимум четырёхкратное сокращение объёма текстовых данных*, необходимых для *передачи* единицы *знаний* посредством ЕЯ без потери полезной составляющей между экспертами и обучаемыми *в открытых тестах*.

Что требует отдельного исследования ?

1. *Согласование данных* об основах и флексиях, выделяемых *по разным СЯУ* относительно фиксированной *предметной области*.

   При этом объём баз знаний, формируемых на основе предложенного метода, может быть *дополнительно сокращён* в среднем *на* 1,5%.

2. *Признаки* направлений связей слов в ситуациях употребления *предметно-ограниченного* подмножества естественного языка:
   - *интерпретация меры* TF-IDF *для оценки важности слова в контексте СЯУ;*
   - *в роли коллекции документов — совокупность СЯУ по заданной предметной области;*
   - *основополагающая гипотеза — зависимое слово имеет больший вес.*

3. *Реконструкция целостного образа СЯУ* (эталон + СЭ-фразы) по текстам тематического корпуса.

   В основе — оценка возможности *совместного появления* связей во фразе, *аналогичная весовой функции* (5) для отдельной связи.

4. *Классификация фраз*, относимых к «эталонным», *по значению суммарной длины связей* слов в их составе в целях компенсации отсутствия ограничения на проективность.