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

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

Основная цель исследований

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

Определение 2

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

Фиксируемый СЯУ S языковой контекст представляется тройкой:

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

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

Замечание

В данной работе элементы множества Ts — *семантически эксивалентные* (СЭ) ЕЯ-фразы.

Пусть Synt — сюръективная функция, определяемая синтаксисом языка.

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

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

Представим СЯУ посредством формального контекста (ФК):

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

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

Множество признаков M включает:

- признаки-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).

Основные подзадачи:

- выделение буквенных инвариантов слов (основ) в составе фраз из Ts;
- формирование критерия информативности слов в контексте СЯУ;
- выделение и классификация синтагматических связей, определяемых отношениями из R в составе модели (1).

Имеем:

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

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}.

Определение 3

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 LS есть множество моделей линейных структур фраз из Ts на J.

Лемма 1

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

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

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

$$D : Ts_i \to \Big\{ \Big(h\big(j, \mathrm{Ls}(Ts_i)\big),\ h\big(k, \mathrm{Ls}(Ts_i)\big) \Big) :\ j \neq k \Big\}. \tag{3}$$

При этом пара (j, k) содержательно соответствует одной *синтагме*.

Замечание

Согласно определению, синтагматические зависимости задаются на множестве флективных частей слов в составе фраз из Ts.

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

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 $\mathrm{Ls}(Ts_i)$.

Пусть

LS' — множество моделей линейных структур фраз из определяющих СЯУ, преобразованное заменой индексов согласно *Лемме 1*;

J' — соответствующее преобразованное индексное множество J.

Введём в рассмотрение абсолютные частоты:

$\mathrm{N}(j, LS')$ — встречаемости отдельного индекса в моделях линейных структур из LS';

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

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

Задача

На основе указанных частот найти набор минимальных текстовых единиц, необходимых и достаточных для формирования оптимального плана передачи смысла ситуации S, представляемой тройкой (1).

Пусть X — последовательность упорядоченных по убыванию значений $\mathrm{N}(j, LS')$ 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

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

Функцию, выдающую $\mathrm{true}/\mathrm{false}$ в зависимости от выполнения условия (4), 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 $H_1, \ldots, H_r$, 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 $\left\{j : \mathrm{N}(j, LS') \in H_1\right\}$ как Cl.

Утверждение 1

Смысловой эталон СЯУ, представляемой тройкой (1), определяют те $Ts_i \in \mathrm{Ts}$, для которых

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

где $\mathrm{Ls}'(Ts_i) \in LS'$.

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

Следующий шаг — *выделение и кластеризация* связей, представляемых множеством (3), с последующим анализом индексов из $\mathrm{Ls}'(Ts_i) \setminus Cl$.

Будем оценивать «силу» связи (независимо от взаимного расположения слов в линейном ряду фразы) посредством следующей *весовой функции:*

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

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

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 9.

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

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

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

Разобьём их на кластеры (слайд 9) по величине среднеквадратического отклонения длины связи (СКОДС) относительно LS'.

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

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

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

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

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

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.

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.

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**

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.*

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.

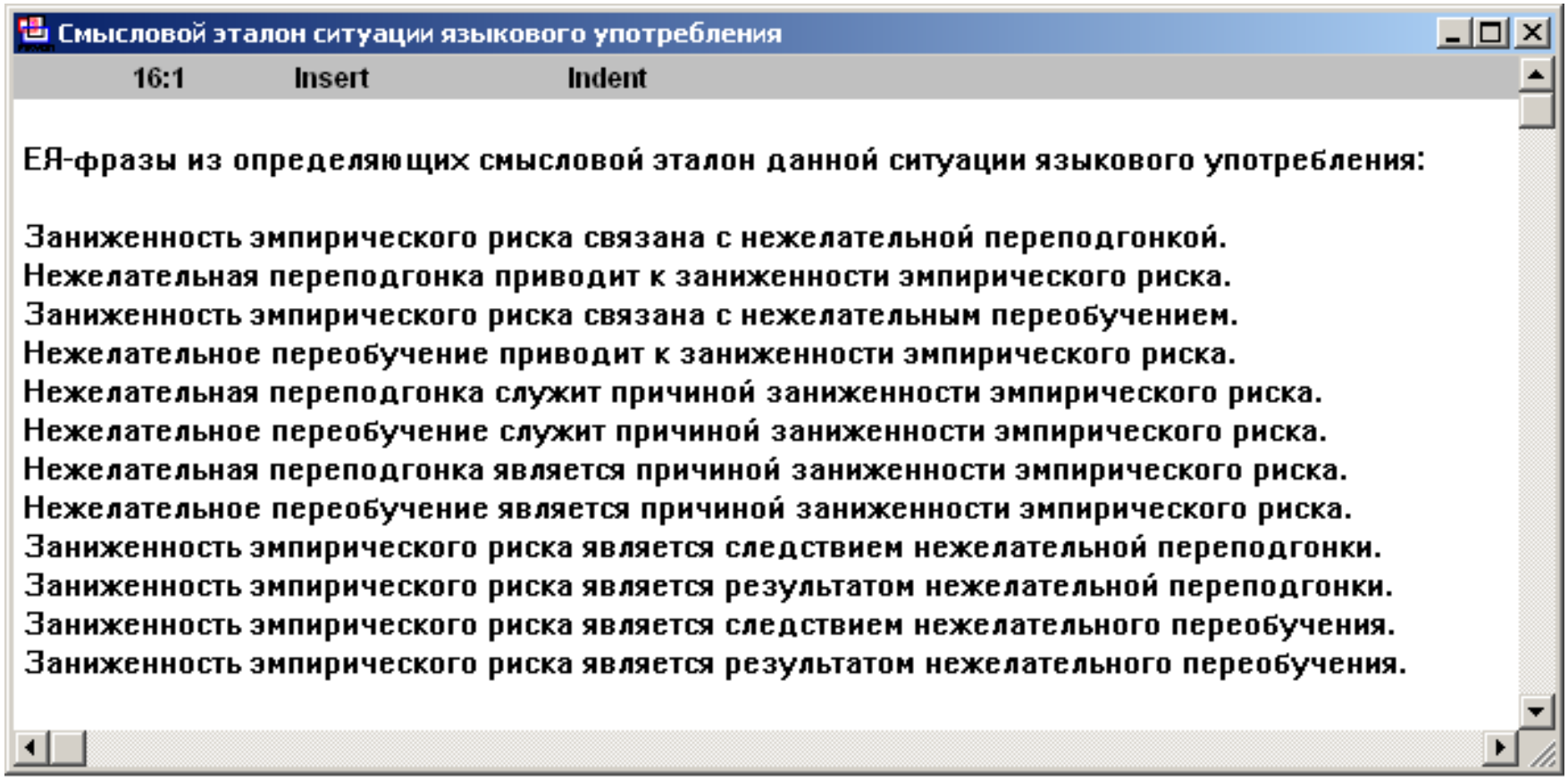
Исходные СЭ-фразы (продолжение) и фраза максимальной длины

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



Выделенные кандидаты на роль вершин синтаксических деревьев:

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

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

Кластеризация по значению весовой функции связи

Связи со значениями весовой функции из кластера H_1^W:

Основа для j	Основа для k	$\mathrm{Dir}(j,k)$	$\mathrm{Wg}\big((j,k),LS'\big)$
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	$\sigma\big(\mathrm{len}(j,k),LS'\big)$
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	$\sigma\left(\mathrm{len}(j,k), LS'\right)$	№ 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

Оценка резервируемого объёма памяти для обработки текстов

№ 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
$\mathrm{vol}(n_1)$	$4{,}790 \cdot 10^8$	$1{,}308 \cdot 10^{12}$	$2{,}092 \cdot 10^{13}$	$6{,}402 \cdot 10^{15}$	$3{,}557 \cdot 10^{14}$	$8{,}718 \cdot 10^{10}$
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;

$\mathrm{vol}(n) = n!$ есть традиционная оценка для фразы длиной максимум в n слов;
vol_1 и vol_2 — оценки с применением предложенного метода выделения эталона СЯУ.

При этом:

$\mathrm{vol}_1 = n_1 \cdot l_1$ 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. 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*.
3. Показанная концепция СЯУ *позволяет* решать задачи *поиска систем зависимостей совместной встречаемости осмысленных фрагментов слов* в контексте связного текста на основе булевых векторов вида (7).
4. Отождествление компонент $d_1, \ldots, d_k$ вектора (7) с направленными лексико-синтаксическими связями *позволяет* формировать *правила принятия решений о наиболее вероятных направлениях* таких связей в заданных языковых контекстах.
5. Для *численной оценки* возможности *совместного появления* связей в контекстах, определяемых векторами указанного вида, может быть использована *оценочная функция*, аналогичная функции (5).

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. *Статистика признаков* словоформ и определение *зависимых* слов *в составе связей* относительно ситуаций 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* на основе вероятностей *совместной встречаемости* лексико-синтаксических *связей слов* в текстах заданного *предметно-ограниченного* ЕЯ-подмножества.