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

Г. М. Емельянов, Д. В. Михайлов, А. П. Козлов
Dmitry.Mikhaylov@novsu.ru
ФГБОУ ВПО «Новгородский государственный университет имени Ярослава Мудрого»

Разработка и анализ результатов открытых тестов требует автоматизации формирования компьютерной модели экспертных знаний, исходно представляемых текстами предметно-ограниченного подмножества естественного языка (ЕЯ). Актуальная при этом задача — выделение необходимого и достаточного набора признаков единицы знаний, оцениваемых с применением теста открытой формы. Для решения указанной задачи в работе предлагается методика выделения структурных единиц, определяющих лексическую сочетаемость и наибоолее характерные синтаксические связи слов в составе множества семантически эквивалентных (СЭ) ЕЯ-описаний фактов предметной области теста. Ранжирование выделяемых связей осуществляется на основе частоты их встречаемости, а также значения среднеквадратического отклонения расстояния между словами в линейном ряду фразы относительно заданного множества СЭ-фраз. Предложенная методика дает минимум четырехкратное сокращение объема текстовой информации, необходимой для оценки правильности ответа испытуемого на вопрос открытого теста.

Ключевые слова: автоматизированный контроль знаний; открытый тест; семантическая эквивалентность; обучаемый парсер; смыслосохраняющее сжатие текста

Formation of the representation of topical knowledge units in the problem of their estimation on the basis of open tests*

G. M. Emelyanov, D. V. Mikhaylov, A. P. Kozlov
Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, Russia

The problem considered is an automated formation of necessary and sufficient feature set of knowledge unit estimated by means of open form test assignments. Such tests assume testee answer in natural language. The most effective open form test implementation implies the known structure of natural-language forms of expression of expert knowledge. For extraction of such forms, it is necessary to analyze equivalent within the meaning descriptions of one and the same fact of topical area in given natural language. The main task here is finding the most rational plan to express the meaning of the expert in the right answer. At that, the meaning eventually must be presented in maximally compact volume of text data. It is relatively to this data that the correctness of testee answer is estimated. The given work represents how it is possible to select this data on the basis of extraction and classification of structural units, defining the lexical-syntactic relations relatively to the set of semantically equivalent phrases in natural language, describing some fact of test topical area. Rating of detected links is carried out on the basis of frequency of their occurrence, as well as values of root-mean-square deviation distance in linear series of the phrase between words as part of the link relatively to the given set of semantically equivalent phrases. The offered by the authors method of finding of these links allows to minimum fourfold reduce the volume of text information necessary for estimation of testee answer correctness to open form test question.

Keywords: computer-aided testing of knowledge; open-form test assignments; semantic equivalence; learnable parser; lossless-in-sense text compression

Введение

Тестовое задание открытой формы в системе автоматизированного контроля знаний предполагает ответ обучаемого (испытуемого) в виде одного или нескольких предложений естественного языка. В отличие от выбора правильного варианта из набора альтернатив, заданий на соответствие, заданий на установление правильной последовательности, тесты открытой формы исключают догадку и позволяют максимально приблизить компьютерный тест к традиционному взаимодействию «Учитель–Ученик».

Тем не менее, как показано в [5], эффективная реализация открытых тестов предполагает известную структуру ЕЯ-форм выражения знаний эксперта. Выделение таких форм требует изучения семантически эквивалентных описаний одного и того же факта заданной предметной области на конкретном естественном языке. Причем сама интерпретация результата теста отнюдь не сводится к простому обнаружению парафраз [14] «ответ испытуемого–правильный ответ». Основная *проблема* здесь — поиск наиболее рационального плана передачи смысла экспертом в «правильном» ответе, сам же смысл в итоге должен быть отражен в максимально компактном объеме текстовых данных. Именно относительно этих данных и оценивается правильность ответа испытуемого. Данная постановка задачи естественно согласуется с пониманием смысла как набора функций [1], которые:

— связывают обозначаемые словами понятия;
— могут быть заданы на множестве символьных цепочек, составляющих основы слов [15].

Поскольку эти функции во многом определяют *минимальные семантико-синтаксические единицы текста* и связи между ними, то решение поставленной *проблемы* есть поиск необходимого и достаточного набора таких единиц, выражающих, как показано в [9], связи:

— неизменяемых частей (основ) синтаксически главного и зависимого слова (лексическая сочетаемость слов);
— изменяемых частей (флексий) главного и зависимого слова в рамках синтаксических отношений.

Именно на основе этих данных и должен производиться отбор ЕЯ-форм выражения знаний эксперта для сопоставления с ответом испытуемого. Немаловажной составляющей здесь является синтаксический разбор, причем как указанных форм, так и анализируемого ответа. Следует отметить, что обучаемый детектор парафраз, предложенный в [14], в качестве исходных данных для обучения нейронной сети также использовал результаты синтаксического разбора исходных фраз парсером, работающим по принципу вероятностной контекстно-свободной грамматики. Для русского языка свободно распространяемым синтаксическим анализатором, дающим наименьший процент ошибок при использовании вероятностных стратегий, до недавнего времени считался «Cognitive Dwarf» (ООО «Когнитивные технологии», `www.cognitive.ru`), используемый в Яндекс с 2010 года.

Актуальной проблемой использования известных программ синтаксического анализа при подготовке открытых тестов и последующей интерпретации ответов испытуемых является ориентация парсера на модели словосочетаний и предложений, наиболее вероятные в языке в целом без учета особенностей предметно-ограниченных ЕЯ-подмножеств, которые и охватываются открытыми тестами. В идеале парсер должен в автоматическом режиме формировать знания о сосуществующих в заданном языковом контексте синтаксических связях на основе закономерностей совместной встречаемости фрагментов символьных последовательностей. При этом исходными данными для обучения парсера бу-

Работа поддержана РФФИ (проект №13-01-00055) и Минобрнауки РФ (базовая часть госзадания).

дут множества СЭ-фраз, описывающих средствами заданного естественного языка факты некоторой ограниченной предметной области.

В работах [13, 15] авторами был предложен принцип выделения и кластеризации рассматриваемых *минимальных семантико-синтаксических текстовых единиц* на множестве эквивалентных по смыслу фраз предметно-ограниченного ЕЯ-подмножества. Новизна решения заключалась в сравнении символьных последовательностей, составляющих эквивалентные по смыслу описания одного и того же объекта (ситуации) на заданном естественном языке, с выделением изменяемых и неизменяемых частей для последующего анализа взаимного расположения фрагментов последовательностей во фразах с разными логическими акцентами относительно одной и той же ситуации. Предложенный принцип выделения смысловых связей слов реализован в рамках демонстрационного варианта системы контроля знаний, представленного в подразделе «Участник: Dmitry.Mikhaylov» информационно-аналитического ресурса `www.machinelearning.ru`.

Вместе с тем, допустимость связей авторами определялась исключительно по наличию минимум двух ЕЯ-фраз исходного синонимического множества, в которых слова пары «главное–зависимое» (независимо от форм) соседствовали в линейном ряду. Сказанное не позволяет оценивать значимость связи для передачи нужного смысла с учетом эффекта свободного порядка слов во фразе [6]. Настоящая работа представляет вариант решения указанной проблемы ранжированием связей на основе частоты их встречаемости, а также значений среднеквадратического отклонения расстояния между словами в линейном ряду фразы относительно заданного множества СЭ-фраз.

Основные понятия и предположения

Будем отождествлять множество Ts, включающее СЭ-фразы предметно-ограниченного естественного языка, с ситуацией употребления последнего для описания факта предметной области теста (далее — ситуацией языкового употребления, СЯУ).

Представим языковой контекст, фиксируемый отдельной СЯУ, посредством тройки

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

в которой $\forall\ g \in G$ есть основа слова, синтаксически подчиненного 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.

Тогда поиск необходимого и достаточного набора минимальных семантико-синтаксических единиц для рациональной передачи заданного смысла означает найти $I \subseteq G \times M$, определяющее фразы минимальной символьной длины при максимизации числа слов, наиболее употребимых в различных фразах из Ts (с учетом возможных синонимов).

Определение 1. *Единицу знаний, представляемую тройкой вида* (1) *и сформированную на основе СЭ-фраз, отвечающих вышеуказанному требованию, будем далее отождествлять со смысловым эталоном СЯУ.*

Применение модели вида (1) в качестве единицы представления знаний о языке и предметной области, а также выделение на основе таких единиц лексико-синтаксических шаблонов с последующей их иерархизацией обсуждалось авторами в [3]. Само же формирование указанной единицы при этом включает следующую совокупность подзадач:

— выделение буквенных инвариантов, составляющих основы слов;
— формирование критерия информативности слов в контексте СЯУ;
— выделение и классификация связей, определяемых отношениями из множества I в составе модели (1).

Для более формальной постановки последней из подзадач введем понятие модели линейной структуры (МЛС) ЕЯ-фразы.

Имеем:

$$\mathrm{Ts} = \Big\{ Ts_i \colon Ts_i = \bigodot_j w_{ij} \Big\},$$

где $\odot$ — операция типа конкатенации, причем w_{ij} представляется последовательностью $W_{ij} = Wc_{ij} \odot Wf_{ij}$, в составе которой Wc_{ij} составляют символы неизменной части (основы) слова w_{ij}, Wf_{ij} — символы его флективной части (флексии).

Далее отождествим с Wc_{ij} и Wf_{ij} принятые в информатике понятия «префикс» и «суффикс» [12]. На множестве суффиксов каждой Ts_i выражаются синтагматические зависимости, которые задаются синтаксическими отношениями и определяют возможность сосуществования словоформ в линейном ряду.

Определение 2. *Пусть J — множество индексов (индексное множество) для основ слов, составляющих фразы из* Ts. *Последовательность таких индексов для некоторой $Ts_i \in \mathrm{Ts}$ назовем моделью ее линейной структуры (МЛС),* $\mathrm{Ls}(Ts_i)$.

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

Определение 3. *Будем говорить, что индексы $j_1, j_2 \in J$ соответствуют словам-синонимам и могут быть заменены одним индексом из $\big(\mathbb{N} \backslash J\big)$, если $\exists \big\{\mathrm{Ls}(Ts_1), \mathrm{Ls}(Ts_2)\big\} \subseteq L$:*

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

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

Далее обозначим множество L, преобразованное заменой индексов в моделях согласно *Определению 3*, как L', а соответствующее преобразованное индексное множество J — как J'.

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

$$D : Ts_i \to \Big\{ \Big(h\big(j, \mathrm{Ls}(Ts_i)\big),\, h\big(k, \mathrm{Ls}(Ts_i)\big) \Big) \,:\, j \neq k \Big\}. \tag{2}$$

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

Определение 4. *Пусть* $\mathrm{len}(j,k) = \big| h\big(j, \mathrm{Ls}(Ts_i)\big) - h\big(k, \mathrm{Ls}(Ts_i)\big) \big|$. *Назовем указанную величину длиной связи, соответствующей паре (j, k), относительно модели* $\mathrm{Ls}(Ts_i)$.

Рассмотрим, в какой мере на основе данных о частоте, с которой связь встречается имеющей различную длину в моделях из L', может быть решена задача поиска необходимого и достаточного набора минимальных семантико-синтаксических единиц в рамках смыслового эталона СЯУ, представляемого моделью (1).

Формирование смыслового эталона

Пусть $\mathrm{N}(j, L')$ — абсолютная частота встречаемости индекса j в моделях линейных структур из множества L', сформированного согласно *Определению 3*, а X — последовательность упорядоченных по убыванию значений указанной частоты для всех $j \in J'$.

Выполним разбивку последовательности X на кластеры с применением *Алгоритма 1*, содержательно близкого алгоритмам класса FOREL [4]. В качестве *центра масс* кластера H_i здесь берется среднее арифметическое всех $x_j \in H_i$. Функцию, вычисляющую центр масс для H_i, далее обозначим как $\mathrm{mc}(H_i)$.

Таблица 1. Вспомогательные функции, используемые Алгоритмом 1

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
$\mathrm{good}(X)$	true либо false в зависимости от выполнения условия (4)

Алгоритм 1 Формирование 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**
   вернуть H_i, X_p и X_s;
6: **иначе если** $\big|\mathrm{mc}(H_i) - \mathrm{first}(H_i)\big| > \big|\mathrm{mc}(H_i) - \mathrm{last}(H_i)\big|$ **то**
7: $X_p := \big\{\mathrm{first}(H_i)\big\} \odot X_p$;
8: $H_i := \mathrm{rest}(H_i)$;
9: перейти к шагу 5;
10: **иначе если** $\big|\mathrm{mc}(H_i) - \mathrm{first}(H_i)\big| < \big|\mathrm{mc}(H_i) - \mathrm{last}(H_i)\big|$ **то**
11: $X_s := \big\{\mathrm{last}(H_i)\big\} \odot X_s$;
12: $H_i := \mathrm{lrev}(H_i)$;
13: перейти к шагу 5;
14: **иначе**
15: $X_s := \big\{\mathrm{last}(H_i)\big\} \odot X_s$;
16: $X_p := \big\{\mathrm{first}(H_i)\big\} \odot X_p$;
17: $Tmp := \mathrm{lrev}(H_i)$;
18: $H_i := \mathrm{rest}(Tmp)$;
19: перейти к шагу 5;

Заметим, что классическая формулировка алгоритма FOREL предполагает минимизацию функционала качества, определяемого как сумма внутрикластерных расстояний

$$\rho_i = \sum_{x_j \in H_i} \big|x_j - \mathrm{mc}(H_i)\big| \tag{3}$$

по всем получившимся кластерам H_i. Одна из основных особенностей использования алгоритмов этого семейства для практических задач — необходимость априорных знаний о

ширине (диаметра) кластера с целью минимизации затрат по пересчету значений функции (3). Для рассматриваемого варианта алгоритма мы ограничиваемся предположением, что элементы одного кластера всегда имеют больше сходств, чем различий. Будем считать, что элементы последовательности X могут быть отнесены к одному кластеру, если

$$\left.\begin{aligned} \big|\,\mathrm{mc}(X) - \mathrm{first}(X)\,\big| &< \frac{\mathrm{mc}(X)}{4}\,; \\ \big|\,\mathrm{mc}(X) - \mathrm{last}(X)\,\big| &< \frac{\mathrm{mc}(X)}{4}\,, \end{aligned}\right\} \tag{4}$$

Алгоритм 1 применяется к последовательностям X_p и X_s на его выходе, каждая из получившихся последовательностей, как и исходная X, будет упорядоченной (доказательство очевидно). Данный процесс продолжается рекурсивно до тех пор, пока на очередном шаге обе последовательности X_p и X_s не окажутся пустыми. В результате исходная последовательность разбивается на подпоследовательности (кластеры) $H_1, \ldots, H_r$, причем для $\forall\ i \neq j$ верно то, что $H_i \cap H_j = \varnothing$, а $H_1 \odot H_2 \odot \ldots \odot H_r = X$.

Пусть $Cl = \big\{j\colon\ \mathrm{N}(j, L') \in H_1\big\}$, а $\mathrm{Ls}'(Ts_i) \in L'$.

Утверждение 1. *Смысловой эталон СЯУ, задаваемой множеством СЭ-фраз* Ts*, определяют те фразы* $Ts_i \in \mathrm{Ts}$*, для которых* $Cl \cap \mathrm{Ls}'(Ts_i) = Cl$*, а* $\big|\ \mathrm{Ls}'(Ts_i) \setminus Cl\ \big| \to \min$.

Данное условие — *необходимое, но не достаточное* для отнесения некоторой фразы Ts_i к множеству фраз, определяющих смысловой эталон заданной СЯУ.

Действительно, *Утверждение 1* затрагивает исключительно лексический состав отбираемых фраз, не принимая во внимание связи слов, определяемые множеством I в составе модели (1). Как следствие, при отборе фраз не учитывается синонимия, затрагивающая одновременно и синтаксические связи, и лексику (ср. *«Нежелательная переподгонка приводит к заниженности эмпирического риска»* $\Leftrightarrow$ *«Заниженность эмпирического риска является следствием нежелательной переподгонки»*).

Предлагаемый вариант решения указанной проблемы основан на частотном анализе индексных пар множества (2). При этом вводятся в рассмотрение абсолютные частоты:

$\mathrm{N}\big((j,k), L'\big)$ — встречаемости связи (j,k) в моделях линейных структур из множества L', сформированного согласно *Определению 3*, независимо от значения длины этой связи, $\mathrm{len}(j,k)$;

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

Будем оценивать «силу» связи слов (вне зависимости от их взаимного расположения в линейном ряду фразы) в контексте СЯУ посредством следующей весовой функции:

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

Суть введения указанной функции — оценить, насколько часто слова с индексами j и k встречаются именно в составе связи (j,k) по отношению ко всем случаям вхождения указанных слов в анализируемые фразы. Такой подход идейно близок методу определения неестественного происхождения текста, основанному на изучении статистики встречаемости пар соседних слов в тексте [2]. Другим методом, близким оценке (5), является вычисление смысловой близости слов на основе их совместной встречаемости в минимальном контексте, соответствующем конкордансу и определяемом лексико-синтаксическим шаблоном из конечного множества [16].

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

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

описанным выше методом с применением Алгоритма 1.

При этом связи, максимально значимые для формирования искомых единиц знаний, будут иметь значения функции (5), вошедшие в кластер H_1^W. Обозначим далее множество индексов, вошедших в указанные связи, как Cl_1 (по аналогии с Cl из Утверждения 1).

Как следует из определения, не предполагая никаких изначальных гипотез относительно смысловой связи слов, соответствующих индексам j и k, оценка (5) тем не менее зависит от частоты встречаемости каждого из них в анализируемых СЭ-фразах. А это означает, что шанс получить оценку из кластера H_1^W будет только у связей между словами, рассматриваемыми *Утверждением 1* в качестве основы отбора «эталонных» фраз. Показанная выше проблема идентификации синтаксических синонимов здесь могла бы быть решена по аналогии с выделением лексико-синтаксических шаблонов на основе n-грамм, как это сделано в работе [11], но при условии априори известных возможных значений n (длины последовательности от j-го до k-го слова включительно). С учетом отсутствия таких данных в настоящей работе было использовано следующее предположение: из тех связей, значения функции (5) для которых не вошли в кластер H_1^W, наименьший разброс длины будет у связей, затрагивающих вершины синтаксических деревьев анализируемых фраз. Отметим, однако, что с учетом возможности свободного порядка слов во фразе обратное утверждение будет не всегда верным.

Сказанное тем не менее позволяет сформулировать эмпирическое правило для выделения кандидатов на роль вышеуказанных вершин на случаи синонимических замен с расщеплением сказуемого (*«приводить» ⇔ «служить причиной»*) и конверсивных замен (*«приводить» ⇔ «являться следствием»*), не охватываемые *Определением 3*, но значимые для смыслового эталона СЯУ.

Для выделения кандидатов в вершины те связи, значения функции (5) которых не вошли в кластер H_1^W, будем группировать описанным ранее методом на основе Алгоритма 1, но по величине среднеквадратического отклонения длины связи (СКОДС) относительно моделей линейных структур из множества L', формируемого согласно *Определению 3*.

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,

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

В качестве основополагающей здесь выдвинута следующая гипотеза: индекс, соответствующий вершине, должен входить в одну из связей кластера наименьших значений СКОДС и одновременно в связь, относящуюся к некоторому другому кластеру из полученных по указанной величине. При этом индекс вершины не входит в связи со значениями функции (5) из кластера H_1^W.

Обозначим множество индексов кандидатов на роль вершин синтаксических деревьев, сформированное по всем исходным СЭ-фразам $Ts_i \in \mathrm{Ts}$, как Cl_2.

Утверждение 2. *Смысловой эталон СЯУ определяют те* $Ts_i \in \mathrm{Ts}$, *для которых помимо выполнения условия Утверждения 1 верно то, что*

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

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

Последовательным отбором фраз, отвечающих условиям Утверждений 1 и 2, решается задача выбора максимально компактного объема текстовых данных из исходного множества Ts для передачи требуемого смысла. Заметим, что данная методика не учитывает проективности фразы в отличие от предложенной авторами в [13], поскольку, как справедливо отмечено в [6], проективность предложения сама по себе не гарантирует сохранение синтаксических групп, что исключает введение искусственного ограничения на проективность при выделении синтагматических связей предлагаемым в работе методом.

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

$$R_J = \Big\{\big((j,k),\ \mathrm{Dir}\big) \colon \mathrm{Dir} \in \{\leftarrow, \rightarrow\}, \exists\, Ts_i \in \mathrm{Ts}^* \colon \{j,k\} \subset \mathrm{Ls}'(Ts_i)\Big\}, \tag{6}$$

причем если $X^W \neq H_1^W$ и $|\,\mathrm{Ts}^*| > 1$, то либо $\{j,k\} \cap Cl_2 \neq \varnothing$, либо паре (j,k) соответствует связь со значением функции (5) из кластера H_1^W. Тогда направленные связи из R_J будут определять искомые минимальные семантико-синтаксические текстовые единицы, задаваемые множеством I в рамках модели (1) для смыслового эталона СЯУ.

Далее проиллюстрируем работу предложенных принципов формирования эталона для СЯУ, описывающей связь между *переобучением* и *эмпирическим риском*.

Экспериментальная апробация

Исходными данными экспериментов был материал ЕЯ-описаний шести фактов предметной области «Математические методы обучения по прецедентам», ранее рассмотренный в [13, 15]. Указанные описания были выполнены независимо друг от друга разными экспертами. В экспериментах число СЭ-фраз, задающих СЯУ, варьировалось в диапазоне от 6 до 56, а число слов во фразе — от 5 до 18, см. табл. 2. Реализация предложенного метода формирования смыслового эталона СЯУ на языке Visual Prolog 5.2 представлена на портале Новгородского государственного университета имени Ярослава Мудрого [8] вместе с исходными текстами программ и результатами машинных экспериментов.

Таблица 2. Исходные данные для проведения экспериментов

Порядковый номер СЯУ	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

Для выделения основ и флексий слов, составляющих исходные СЭ-фразы, была реализована группировка словоформ в рамках СЯУ по общности префикса и (при необходимости) суффикса.

В процессе группировки анализируются абсолютные частоты встречаемости символов на разных позициях относительно начала и конца слова. При этом среди словоформ с некоторым фиксированным префиксом частота встречаемости первого слева символа всегда максимальна. Такая же частота будет и у остальных букв, составляющих общий префикс. Относительно конца слова также производится выделение общего суффикса (в составе основы он здесь не учитывается и включается во флективную часть) — на случай, в частности, наличия возвратных частиц. При этом суммарная длина общего префикса и общего суффикса должна составлять минимум треть длины слова, а разность длин любой пары слов, имеющих общий префикс (как в совокупности с общим суффиксом, так и без такового), всегда меньше половины длины меньшего слова. Здесь, как и при формировании отдельного кластера *Алгоритмом 1*, мы руководствуемся предположением, что у близких друг другу объектов сходства всегда преобладают над различиями. Для слов, имеющих общую основу, это свойство выражается в преобладании сходств по буквенному составу.

Таблица 3. Синтагматические связи и смысловые эталоны для СЯУ из табл. 2

Порядковый номер СЯУ, i	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 смыслового 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

Таблица 4. Фразы максимальной длины из определяющих СЯУ в табл. 2

i	Фраза максимальной длины
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.

В качестве примера на рис. 1 показан фрагмент исходного множества СЭ-фраз, задающих СЯУ №1 (связь переобучения и эмпирического риска) из представленных в табл. 4. Отметим, что данный фрагмент содержит фразы и минимальной (выделена *зеленым* цветом), и максимальной длины (выделена *красным* цветом) из задающих эту СЯУ. В рассматриваемом примере принадлежность фразы к задающим смысловой эталон согласно *Утверждению 1* устанавливалась по одновременному наличию в ее составе слов с основами: *«нежелательн», «переподгонк/переобучени», «риск», «эмпирическ», «заниженн»*.

Как видно из табл. 5, связи между словами с перечисленными основами получили самые высокие значения оценки вида (5) относительно рассматриваемой СЯУ.

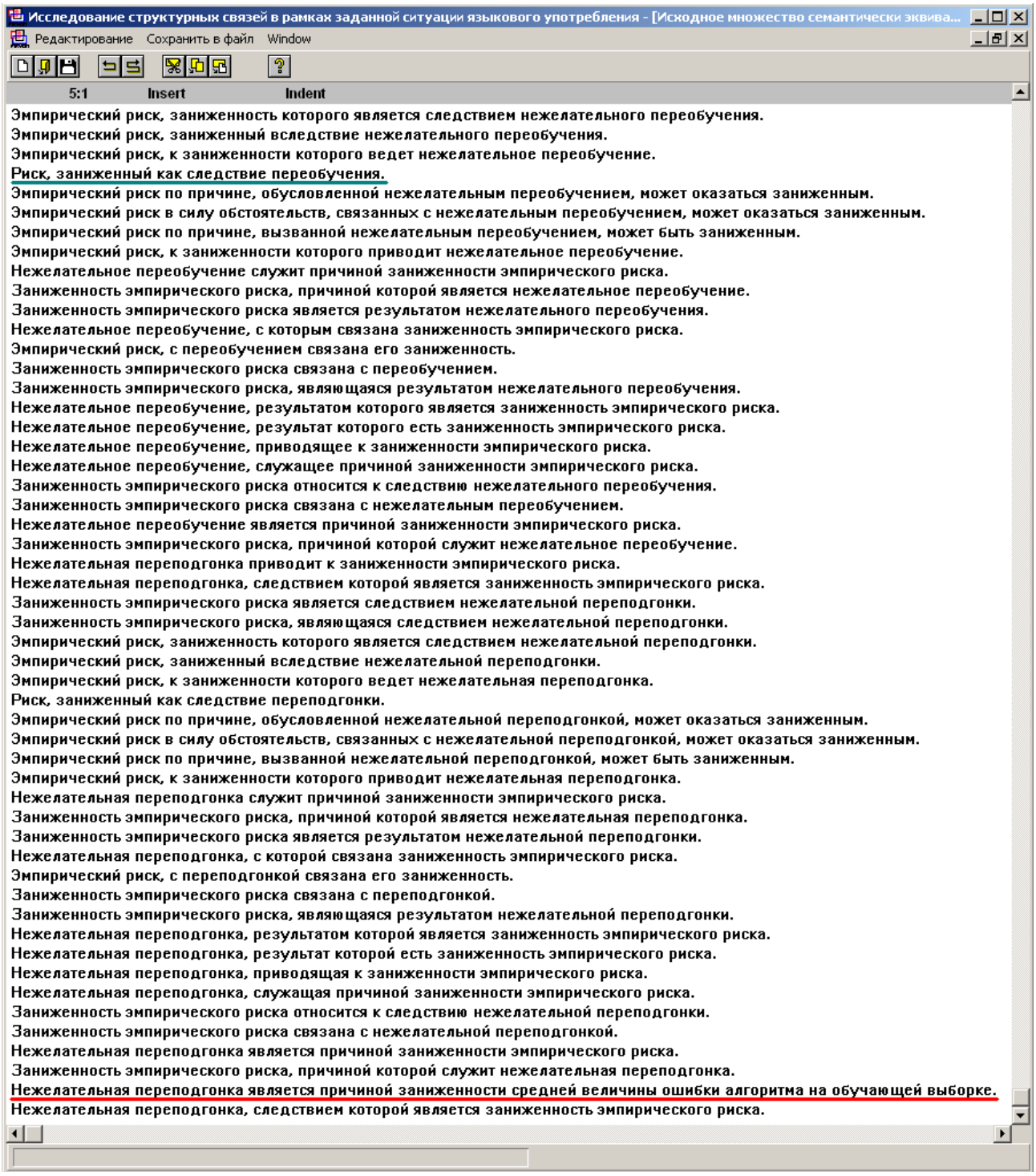

Рис. 1. Фрагмент исходного множества семантически эквивалентных фраз

Результирующие фразы, определяющие смысловой эталон этой СЯУ, и связи в рамках найденного эталона представлены на рис. 2.

Рис. 2. Смысловой эталон и синтаксические связи в его рамках

Данные о кластерах, выделенных по значению среднеквадратического отклонения длины связи относительно моделей линейных структур из множества L', сформированного согласно *Определению 3*, приведены в табл. 6. При этом к основам слов-кандидатов на роль вершин синтаксических деревьев были отнесены: *«привод/ведет»*, *«связан»*, *«результат/следстви»*, *«причин»*, *«котор»*, *«есть/явля/служ»*, а также предлоги *«с»*

Таблица 5. Связи со значениями функции (5) 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

Таблица 6. Кластеры, выделенные по значению СКОДС

№ 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

Таблица 7. Связи со значениями функции (5), не вошедшими в кластер H_1^W

Основа для j	Основа для k	$\mathrm{Dir}(j,k)$	$\mathrm{Wg}\big((j,k),L'\big)$	$\sigma\big(\mathrm{len}(j,k),L'\big)$
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

и *«к»*. Для сравнения в табл. 7 представлены те связи из составивших основу выделения указанных кандидатов, которые определяют отношения из множества I в модели (1) эталона данной СЯУ. Как видно из табл. 6, ни одна из этих связей не попадает в кластер наименьших значений СКОДС. В то же время связи, вошедшие в указанный кластер и затрагивающие потенциальные вершины синтаксических деревьев (см. табл. 8), соответствуют тем фрагментам анализируемых фраз, которые в процессе синонимического перифразирования изменяются в наименьшей степени, а именно:

— сочетанию сказуемого в составе определительного придаточного с союзным словом, например: *«котор — привод/ведет»* (связь № 1), ср. *«Эмпирический риск, к заниженности которого приводит нежелательное переобучение»*;

— совокупности сочетаний слов с предлогом в рамках предложной связи, например: *«относится — к»* (связь № 8) и *«к — результат/следстви»* (связь № 6), ср. *«Заниженность эмпирического риска относится к следствию нежелательного переобучения»*;
— сочетанию слов с целью выразить определенный логический акцент, например: *«связан — его»* (связь № 3), ср. *«Эмпирический риск, с переобучением связана его заниженность»*.

Заметим, что к последнему случаю может быть отнесено также сочетание определяемого слова и причастия в составе причастного оборота, если оборот стоит после определяемого слова и его положение не изменяется в процессе перифразирования, например: *«причин — обусловленной»* (связь № 12), *«причин — вызванной»* (связь № 10), ср. *«Эмпирический риск по причине, обусловленной/вызванной нежелательным переобучением, может оказаться заниженным»* ⇔ *«Эмпирический риск по причине, обусловленной/вызванной нежелательной переподгонкой, может оказаться заниженным»*.

Таблица 8. Связи из кластера наименьших СКОДС, затрагивающие потенциальные вершины

№ п/п	Основа для j	Основа для k	$\mathsf{E}\big(\mathrm{len}(j,k),L'\big)$	$\sigma\big(\mathrm{len}(j,k),L'\big)$
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

С учетом показанных особенностей кластера наименьших СКОДС авторами реализовано еще одно немаловажное требование к «эталонным» фразам помимо задаваемых *Утверждениями 1* и *2* — исключить придаточные предложения там, где требуемый смысл может быть выражен без их участия, например: *«Нежелательное переобучение, следствием которого является заниженность эмпирического риска»*. Формально данное требование можно определить следующим образом: если J_1 и J_2 — множества индексов моделей линейных структур фраз Ts_1 и Ts_2, соответственно, то $J_1 \not\subset J_2$ и $J_2 \not\subset J_1$ при $J_1 \neq J_2$.

Для каждой найденной связи, включаемой в множество (6), ее направление Dir в текущей реализации задается путем опроса эксперта. При этом направление может быть задано только для тех из найденных связей, которые будут определены экспертом как истинные. Сформированные таким образом знания системы об истинных и ложных связях в рамках эталона отдельной СЯУ могут быть представлены булевым вектором

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

где компоненты $d_1,\ldots,d_k$ отождествляются с истинными, а $\overline{d}_{k+1},\ldots,\overline{d}_n$ — с ложными связями. Совокупность знаний в виде векторов (7) по разным СЯУ может быть использована

для изучения закономерностей совместной встречаемости слов в составе лексико-синтаксических связей, представляемых отношениями из множества I в составе модели (1).

Выделение потенциальных вершин деревьев синтаксического подчинения фраз, определяющих эталон, позволяет минимизировать информацию, запрашиваемую у эксперта, при определении направлений для связей, представляемых вектором (7), в контексте отдельной СЯУ. При этом, однако, число кластеров, выделенных по значению СКОДС и используемых для определения кандидатов на роль вершин синтаксических деревьев, должно быть минимум два. В примере для СЯУ № 5 (зависимость оценки ошибки распознавания от выбора решающего правила) из представленных в табл. 4 по значению СКОДС выделен всего один кластер, и, следовательно, кандидаты на роль вершин не могут быть найдены. Таким образом, эффективность применения предложенного метода формирования смыслового эталона напрямую зависит от того, насколько исходные СЭ-фразы, определяющие СЯУ, полно учитывают ее смысловой контекст и возможные формы его выражения во фразах рассматриваемого предметно-ограниченного ЕЯ-подмножества.

С другой стороны, число исходных СЭ-фраз, определяющих СЯУ и составляющих обучающую выборку в предложенном методе формирования смыслового эталона, является конечной величиной и в первую очередь зависит от числа возможных синонимов на лексическом и синтаксическом уровне в рассматриваемом ЕЯ-подмножестве. Типы же смысловых связей слов в рамках отдельной фразы изначально не оговариватся, поэтому для полноты учета смыслового контекста СЯУ, определяемого моделью (1), ограниченного набора известных семантических отношений и форм их выражения в текстах, как правило, недостаточно. Подтверждение тому — результаты поиска смысловых связей с помощью системы «Серелекс» (`http://serelex.cental.be`) между словами, выделенными согласно *Утверждениям 1* и *2* в качестве основы формирования эталона СЯУ № 1 из представленных в табл. 4. В табл. 9 для указанных слов приводятся начальные формы (леммы) вместе с их английскими смысловыми эквивалентами, а также общее число связей, найденное системой «Серелекс» для русского и английского вариантов слов по коллекции текстовых документов, включающей заголовки статей Википедии ($2{,}026 \cdot 10^9$ словоформ, 3 368 147 лемм) и текстовый корпус ukWaC [10] ($0{,}889 \cdot 10^9$ словоформ, 5 469 313 лемм).

Таблица 9. Начальные формы слов, английские эквиваленты и общее число связей

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

Заметим, что из найденных отношений только три (*«risk — result»*, *«risk — reason»* и *«риск — с»*) связывают слова из представленных в табл. 9, причем ни одна из указанных связей не имеет синтаксической природы в контексте рассматриваемой СЯУ, что необходимо для формирования отношения I в составе модели (1). Более того, связь *«риск — с»* по указанной причине в примере на рис. 2 для рассматриваемой СЯУ определена экспертом как нерелевантная (ложная), как и связь *«risk — reason»* (*«риск — причина»*). Таким образом, предложенный в настоящей работе метод формирования смыслового эталона СЯУ, не будучи ориентированным на определенные типы связей между словами исходных СЭ-фраз, позволяет их выделять более точно и на основе меньших обучающих выборок применительно к задаче формирования единиц знаний для открытых тестов.

Следует отметить, что введенная концепция смыслового эталона ситуации языкового употребления позволяет оценить объем текстовой информации, необходимой для передачи единицы знаний посредством естественного языка без потери полезной составляющей с учетом возможных видов синонимии. Предложенный метод формирования эталона СЯУ позволяет дать оценку данного объема сверху как $\mathrm{vol}_1 = n_1 l_1$ и снизу как $\mathrm{vol}_2 = n_2 l_2$, где l_1 — число СЭ-фраз из задающих СЯУ, из которых l_2 определяют эталон, n_1 и n_2 — максимальная длина фразы по СЯУ в целом и из определяющих эталон, соответственно. Соотношение указанных оценок для СЯУ из табл. 4 представлено в табл. 10.

Таблица 10. Оценка объема текстовой информации для передачи единицы знаний

№ 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

Заключение

Предложенная *концепция ситуации языкового употребления* позволяет существенно автоматизировать формирование *единиц* наиболее *оптимального представления* экспертных *знаний* для разработки открытых тестов.

Представленная в работе *методика формирования указанных единиц* позволяет в автоматическом режиме выделять шаблоны лексико-синтаксических связей, необходимых и достаточных для передачи нужного смысла в заданном предметно-ограниченном ЕЯ-подмножестве. Применение векторов вида (7) как основы представления таких знаний дает возможность построения синтаксического анализатора, способного обучаться на текстах предметно-ограниченных подмножеств заданного естественного языка с формированием правил принятия решений о наиболее вероятных связях и их направлениях в тех или иных лексико-синтаксических контекстах.

Следует отметить тем не менее что формируемые при этом правила касаются связей исключительно в рамках эталона СЯУ. Ограничение выделяемых лексико-синтаксических связей указанными рамками в настоящей работе обусловлено использованием именно этих связей для оценки близости ответа испытуемого правильному ответу (сама оценка была ранее рассмотрена авторами в [7]).

Экспериментальным подтверждением того факта, что связи, выделяемые в рамках эталона, отвечают наиболее рациональному плану передачи заданного смысла, в идеале может послужить безошибочность разбора «эталонных» фраз парсером, ориентированным на наиболее вероятные в языке модели словосочетаний и предложений (здесь авторами использовался «Cognitive Dwarf», версия от 24.11.2006 г.). При этом компонента связности формируемого графа разбора фразы должна быть единственна. Требование единственности компоненты связности здесь, однако, выполняется не всегда. Пример: *«Оценка частоты ошибок алгоритма на контрольной выборке, заниженная как следствие переподгонки»* (фраза из определяющих эталон СЯУ №4 в табл. 4). С учетом того, что при отборе «эталонных» фраз предложенным в работе методом рассматриваются все возможные порядки следования слов, выделяемых согласно *Утверждениям 1* и *2* в качестве основы отнесения фразы к определяющим эталон заданной СЯУ, здесь в дальнейшем может потребоваться классификация отбираемых фраз по значению суммарной длины связей слов в их составе.

В предложенном авторами решении допустимость выделяемых связей слов, а также их направление задаются экспертом вручную. Следует отметить, что с учетом особенностей ЕЯ-форм ответов на тестовые задания такие трудозатраты являются вполне оправданными. Привлечение внешних морфологических и синтаксических анализаторов для рассматриваемого круга практических задач потребовало бы существенно больших трудозатрат, в частности, по изучению результатов разбора и их коррекции с учетом особенностей того или иного предметно-ограниченного ЕЯ-подмножества. К тому же, немаловажным негативным фактором здесь бы стал большой объем «лишней» статистики в модели (если анализатор основан на машинном обучении, [6]), что привело бы к чрезмерному расходу памяти при достаточно невысоком быстродействии системы в целом.

Наиболее *слабым местом* предложенного решения является относительно малый объем исходных данных для вычисления исследуемой характеристики связи слов — среднеквадратического отклонения ее длины. Здесь как перспективное направление дальнейших изысканий следует отметить реконструкцию целостного образа СЯУ в виде совокупности определяющих ее СЭ-фраз и смыслового эталона на основе текстов тематического корпуса. Основой такого решения может стать численная оценка возможности совместного появления лексико-синтаксических связей во фразе с использованием оценочной функции, аналогичной функции (5). При этом частоты появления слов вне и в составе связей здесь будут оцениваться уже относительно фраз из текстов заданного корпуса.

Другая немаловажная задача — согласование данных буквенного состава основ и флексий, выделяемых по разным СЯУ относительно фиксированной предметной области. Постановка такой задачи и ее решение в первом приближении было ранее представлено авторами в [13]. Сказанное позволит компенсировать зависимость изменяемости слова от полноты представления ситуации языкового употребления исходным множеством СЭ-фраз, а также дополнительно сократить в среднем на 1,5% объем баз знаний, представляемых тройками вида (1) и формируемых предложенным в настоящей работе методом.

В целом же, как видно из табл. 10, выделение смысловых эталонов ситуаций языкового употребления дает *минимум четырехкратное* сокращение объема текстовой информации, необходимой для передачи единицы знаний посредством естественного языка без потери полезной составляющей.

Авторы выражают признательность рецензентам за ценные замечания и рекомендации по дальнейшему развитию полученных теоретических и прикладных результатов.

Литература

[1] *Краснов А. Н., Мошков И. С., Якимов В. Н.* Компьютерная система анализа текста таксономического типа применительно к оценке профессиональных знаний // *Междунар. науч.-практ. конф. «Инновация–2011»: Сб. науч. статей.* — Ташкент: ТашГТУ, 2011. С. 287–289.

[2] *Huang E.* 2011. Paraphrase detection using recursive autoencoder. Available at: `http://nlp.stanford.edu/courses/cs224n/2011/reports/ehhuang.pdf`.

[3] *Герасимова И. А.* Формальная грамматика и интенсиональная логика. — М.: Институт философии РАН, 2000. 156 с.

[4] *Mikhailov D., Emelyanov G.* Lossless-in-sense textual information's compression based on knowledge base about synonymy // *11-th Conference (International) on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013).* 2013. Vol. 2. Pp. 438–441.

[5] *Осипов Г. С.* Приобретение знаний интеллектуальными системами: основы теории и технологии. — Москва: Наука, 1997. 112 с.

[6] *Emelyanov G. M., Mikhailov D. V.* Sense standards, recognition of textual information and its compression based on knowledge of synonymy // *Pattern Recognition Image Analysis.* 2014. Vol. 24. No. 1. P. 63–72.

[7] *Кудинов М. С.* Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей // *Машинное обучение и анализ данных.* 2013. Т. 1. № 6. С. 714–724.

[8] *Емельянов Г. М., Михайлов Д. В.* Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний // *Всеросс. конф. ММРО-15.* — М.: Макс Пресс, 2011. С. 581–584.

[9] *Charras C., Lecroq T.* 1997. Exact string matching algorithms. Available at: `http://www-igm.univ-mlv.fr/~lecroq/string/index.html`.

[10] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Издательство Института математики, 1999. 270 с.

[11] *Гречников Е. А., Гусев Г. Г., Кустарев А. А., Райгородский А. М.* Поиск неестественных текстов // *Тр. XI Всеросс. научной конференции RCDL'2009.* —Петрозаводск: КарНЦ РАН, 2009. С. 306–308.

[12] *Panchenko A., Morozova O., Naets H.* 2012. A semantic similarity measure based on lexico-syntactic patterns. Available at: `http://www.oegai.at/konvens2012/proceedings/23_panchenko12p/23_panchenko12p.pdf`.

[13] *Bollegala D., Matsuo Yu., Ishizuka M.* 2007. Measuring semantic similarity between words using Web search engines. Available at: `http://www2007.org/papers/paper632.pdf`.

[14] *Михайлов Д. В.* Программа выделения структурных связей в рамках семантически эквивалентных фраз на основе анализа буквенного состава словоформ, 2014. `http://www.novsu.ru/file/1089439`.

[15] *Baroni M., Bernardini S., Ferraresi A., Zanchetta E.* 2008. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. Available at: `http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf`.

[16] *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.

References

[1] *Gerasimova I. A.* 2000. *Formal grammar and intensional logic.* Moscow: Institute of Philosophy of the Russian Academy of Sciences. 156 p.

[2] *Grechnikov E. A., Gusev G. G., Kustarev A. A., Raigorodsky A. M.* 2009. Detection of artificial texts. *RCDL'2009 Proceedings*. Petrozavodsk. 306–308.

[3] *Emelyanov G. M., Mikhaylov D. V.* 2011. Formal concept analysis and compression of textual information in the problem of computer-aided testing of knowledge. *Proceedings of All-Russian Conference MMPR-15*. Moscow. 581–584.

[4] *Zagoruiko N. G.* 1999. *Applied methods of data and knowledge analysis*. Novosibirsk: Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences. 270 p.

[5] *Krasnov A. N., Moshkov I. S., Yakimov V. N.* 2011. Computer system for analysis of text of taxonomic type in application to the assessment of professional knowledge. *Proceedings of Scientific and Practical Conference (International) «Innovation–2011»*. Tashkent 287–289.

[6] *Kudinov M. S.* 2013. Shallow parsing of russian text with conditional random fields. *Machine Learning Data Analysis*. 1(6):714–724.

[7] *Mikhaylov D. V., Emelyanov G. M.* 2010. Semantic affinity of texts in a problem of computer-aided testing of knowledge. *Proceedings of Conference (International) IIP-2010*. Moscow. 516–519.

[8] *Mikhaylov D. V.* 2014. The program for extraction of structural relationship within a semantically equivalent phrases using the alphabetic structure of wordforms. Available at: `http://www.novsu.ru/file/1089439`.

[9] *Osipov G. S.* 1997. *Knowledge acquisition by intellectual systems: Fundamentals of theory and technology*. Moscow: Nauka. 112 p.

[10] *Baroni M., Bernardini S., Ferraresi A., Zanchetta E.* 2008. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. Available at: `http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf`.

[11] *Bollegala D., Matsuo Yu., Ishizuka M.* 2007. Measuring semantic similarity between words using Web search engines. Available at: `http://www2007.org/papers/paper632.pdf`.

[12] *Charras C., Lecroq T.* 1997. Exact string matching algorithms. Available at: `http://www-igm.univ-mlv.fr/~lecroq/string/index.html`.

[13] *Emelyanov G. M., Mikhailov D. V.* 2014. Sense standards, recognition of textual information and its compression based on knowledge of synonymy // *Pattern Recognition Image Analysis*. 1:63–72.

[14] *Huang E.* 2011. Paraphrase detection using recursive autoencoder. Available at: `http://nlp.stanford.edu/courses/cs224n/2011/reports/ehhuang.pdf`.

[15] *Mikhailov D., Emelyanov G.* 2013. Lossless-in-sense textual information's compression based on knowledge base about synonymy // *11th Conference (International) on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013)*. 2:438–441.

[16] *Panchenko A., Morozova O., Naets H.* 2012. A semantic similarity measure based on lexico-syntactic patterns. Available at: `http://www.oegai.at/konvens2012/proceedings/23_panchenko12p/23_panchenko12p.pdf`.