Релевантность множества тематических текстов единице знаний и оценка близости языковых форм её выражения смысловому эталону

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

Новгородский государственный университет
имени Ярослава Мудрого

12-я Международная конференция
«Интеллектуализация обработки информации» (ИОИ-2018),

8–12 октября 2018 г.

г. Гаэта, Италия

Единица знаний

Определяется множеством семантически эквивалентных фраз предметно-ограниченного естественного языка.

Оптимальная передача смысла

Обеспечивается теми фразами из исходного множества эквивалентных по смыслу, которые при минимальной символьной длине имеют максимум слов, наиболее употребимых во всех исходных фразах (с учётом возможных синонимов). Именно такие фразы представляют *смысловой эталон*.

Основные проблемы:

- полнота выделения единиц знаний из текстов тематического корпуса анализом релевантности исходной фразе;
- поиск наиболее рационального языкового варианта описания выделяемого фрагмента знаний, отвечающего *смысловому эталону*.

Смысловой эталон

Определяется *набором* текстовых единиц и их связей, *необходимым и достаточным* для представления единицы знаний.

Основные проблемы:

- неизменность состава исходного множества семантически эквивалентных (СЭ) фраз по ходу построения аннотации;
- *точность* выделения смыслового эталона *существенно зависит от полноты описания языковых форм* выражения единицы знаний;
- требуется *синтаксический разбор* (полный или частичный) *исходных СЭ-фраз* для определения наиболее значимых связей и статистики расстояний между словами в составе связи в рамках отдельных фраз;
- *совместимость* морфологических характеристик слов и формата пометок (тегов) у разных программ морфологического анализа.

Цель исследования

Найти *компромисс* между *точностью выделения связей слов*, наиболее значимых для языкового представления единицы знаний и *числом исходных СЭ-форм* её описания экспертом.

Согласно классическому определению, данная мера есть произведение TF-меры (отношения числа вхождений слова к общему числу слов документа) и инверсии частоты встречаемости в документах корпуса (IDF).

TF-мера оценивает важность слова t_i в пределах отдельного документа d и определяется как

$$\mathrm{tf}\,(t_i, d) = \frac{n_i}{\sum_k n_k}, \tag{1}$$

где n_i — число вхождений слова t_i в документ d,
а в знаменателе — общее число слов в документе.

IDF (inverse document frequency) — обратная частота документа, является единственной для каждого уникального слова в корпусе D и равна

$$\mathrm{idf}\,(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right), \tag{2}$$

где в числителе представлено общее число документов корпуса,
а $|D_i \subset D|$ есть число документов, где t_i встретилось хотя бы раз.

Классификация слов исходной фразы по TF-IDF: базовые предположения

1. Наиболее уникальные слова в документе (с наибольшими значениями TF*IDF) будут относиться к терминам его предметной области.
2. Наличие синонимов у слова-термина ведёт к снижению значения TF относительно документа в случае, когда синонимы встречаются в этом же документе.
3. Термины, преобладающие в корпусе, а также слова общей лексики будут иметь значения IDF, близкие к нулю.
4. Слова-синонимы, уникальные для отдельных документов корпуса, будут иметь более высокие значения IDF.

   Пример — слова общей лексики, задающие конверсивные замены: *«приводить ⇔ являться следствием»*.

Выбор оценки силы связи слов

Дистрибутивно-статистический метод [Москович В. А., 1971] построения тезаурусов — сила связи совместно встречающихся во фразе слов:

$$K_{AB} = \frac{k}{a + b - k}, \tag{3}$$

где a — число фраз текста, которые содержат слово A, b — слово B, k — A и B одновременно.

Пусть

D — исходное текстовое множество (корпус).

X — упорядоченная по убыванию последовательность $\mathrm{tf}(t_i, d) \cdot \mathrm{idf}(t_i, D)$ для всех слов t_i исходной фразы относительно документа $d \in D$.

$H_1, \dots H_r$ — последовательность кластеров, на которые разбивается X алгоритмом, содержательно близким алгоритмам класса FOREL.

Центром масс кластера H_i возьмём среднее арифметическое всех $x_j \in H_i$.

Наибольший интерес для выделения связей представляют слова кластеров:

$H_1(X)$ — *слова-термины* исходной фразы, *наиболее уникальные* для d;

$H_{r/2}(X)$ — *общая лексика*, обеспечивающая *синонимические перифразы*, и *термины-синонимы*.

Определение 1

Будем называть далее слова *связанными в паре* по TF-IDF и вычислять для них оценку (3) только в том случае, если значение указанной меры минимум одного из слов пары принадлежит либо $H_1(X)$, либо $H_{r/2}(X)$.

Пусть $L(d)$ есть последовательность биграмм — пар слов (A, B) исходной фразы, связанных в зависимости от метода выделения связей либо синтаксически, либо по TF-IDF, упорядоченная по убыванию силы связи относительно некоторого документа $d \in D$, $\{(A_1, B_1), (A_2, B_2)\} \subset L(d)$.

Определение 2

Биграммы (A_1, B_1) и (A_2, B_2) войдут в одну n-грамму $T \subseteq L(d)$, если

$$\big((A_1 = A_2) \vee (B_1 = B_2) \vee (A_1 = B_2) \vee (B_1 = A_2)\big) = \text{true}.$$

Значимость n-граммы T для оценки ранга документа d относительно D

$$\mathrm{N}(T, d) = \frac{\sqrt{\sum_{i=1}^{\mathrm{len}(T)} \left[S_i(d)\right]^2}}{\sigma\big(S_i(d)\big) + 1}, \tag{4}$$

где $S_i(d)$ — сила связи слов i-й биграммы относительно d;

$\sigma\big(S_i(d)\big)$ — среднеквадратическое отклонение (СКО) указанной величины;

$\mathrm{len}(T)$ — длина n-граммы T (в биграммах).

Обозначим далее множество n-грамм $\{T\colon\ T \subseteq L(d)\}$ как $\mathbb{T}(d)$.

Ранг документа d относительно исходного текстового множества D:

$$W(d) = \mathrm{N}_{\max}(d) \cdot \log_{10}\Big(\max_{T \in \mathbb{T}(d)} \mathrm{len}\,(T)\Big) \cdot \log_{10}\Big(\big|\mathbb{T}\,(d)\big|\Big), \qquad (5)$$

где $\mathrm{N}_{\max}(d) = \max_{T \in \mathbb{T}(d)} \mathrm{N}(T, d)$,

а n-граммы в $\mathbb{T}\,(d)$ упорядочены по убыванию величины $\mathrm{N}(T, d) \cdot \mathrm{len}\,(T)$.

Пусть

Ts — *группа исходных фраз*, взаимно эквивалентных либо дополняющих друг друга по смыслу и определяющих некоторую *единицу знаний*.

Оценка релевантности

текстового корпуса D единице знаний и ситуации языкового употребления, отождествляемыми с Ts, на основе найденных n-грамм:

$$\mathbb{W}(D) = \frac{1}{\big|D'\big|} \sum_{d \in D'} \left[\frac{\Big|\big\{w \in b\colon\ \exists T \in \mathbb{T}'\,(d)\,,\, b \in T\big\}\Big|}{\Big|\big\{w\colon\ \exists Ts_i \in \mathrm{Ts},\, w \in Ts_i\big\}\Big|} \sum_{T \in \mathbb{T}'(d)} \mathrm{N}(T, d) \right], \qquad (6)$$

где $\mathrm{N}(T, d)$ — оценка значимости n-граммы T согласно (4);

$\mathbb{T}'\,(d)$ — кластер наибольших значений оценки (4) по заданному d;

$D' \subset D$ — кластер наибольших значений оценки (5).

Исходное множество тематических текстов: вариант 1

- 1 статья в журнале «Вестник Российского экономического университета им. Г. В. Плеханова (Вестник РЭУ)»;
- 1 статья в журнале «Философия науки»;
- материалы тезисов четырёх докладов на 4-й Всероссийской конференции студентов, аспирантов и молодых учёных «Искусственный интеллект: философия, методология, инновации» (ИИ ФМИ, 2010 г.);
- материалы тезисов двух секционных и одного пленарного доклада на 7-й Всероссийской конференции ИИ ФМИ, 2013 г.;
- материалы одного пленарного доклада на 8-й Всероссийской конференции ИИ ФМИ, 2014 г.;
- 1 статья в сборнике трудов 9-й Всероссийской конференции ИИ ФМИ, 2015 г.;
- 1 статья в журнале «Таврический вестник информатики и математики (ТВИМ)».

Примечание

Число слов в документах исходного множества здесь варьировалось от 618 до 3765, число фраз — от 38 до 276.

Исходные фразы для формирования единиц экспертных знаний

№	Исходная фраза
1	*Определение модели представления знаний накладывает ограничения на выбор соответствующего механизма логического вывода.*
2	*Под знанием понимается система суждений с принципиальной и единой организацией, основанная на объективной закономерности.*
3	*С точки зрения искусственного интеллекта знание определяется как формализованная информация, на которую ссылаются или используют в процессе логического вывода.*
4	*Факты обычно указывают на хорошо известные обстоятельства в данной предметной области.*
5	*Эвристика основывается на собственном опыте специалиста в данной предметной области, накопленном в результате многолетней практики.*
6	*Метазнания могут касаться свойств, структуры, способов получения и использования знаний при решении практических задач искусственного интеллекта.*
7	*Однородность представления знаний приводит к упрощению механизма управления логическим выводом и упрощению управления знаниями.*
8	*Отличительными чертами логических моделей являются единственность теоретического обоснования и возможность реализации системы формально точных определений и выводов.*
9	*Язык представления знаний на основе фреймовой модели наиболее эффективен для структурного описания сложных понятий и решения задач, в которых в соответствии с ситуацией желательно применять различные способы вывода.*

Исходное множество тематических текстов: вариант 2

- 3 статьи в журнале «Таврический вестник информатики и математики»;
- 2 статьи в сборниках трудов конференций «Интеллектуализация обработки информации» 2010 и 2012 гг.;
- 1 статья в сборнике трудов 15-й Всероссийской конференции «Математические методы распознавания образов» (2011 г.);
- материалы тезисов двух докладов на 13-й Всероссийской конференции «Математические методы распознавания образов» (2007 г.);
- материалы тезисов четырнадцати докладов на 16-й Всероссийской конференции «Математические методы распознавания образов» (2013 г.);
- материалы тезисов двух докладов на конференции «Интеллектуализация обработки информации» 2014 г.;
- материалы одного научного отчёта (Михайлов Д. В., 2003 г.).

Примечание

Число слов в документах исходного множества здесь варьировалось от 218 до 6298, число фраз — от 9 до 587.

Исходное множество тематических текстов: тематика отбираемых работ для варианта 2

- математические методы обучения по прецедентам (К. В. Воронцов, М. Ю. Хачай, Е. В. Дюкова, Н. Г. Загоруйко, Ю. Ю. Дюличева, И. Е. Генрихов, А. А. Ивахненко);
- модели и методы распознавания и прогнозирования (В. В. Моттль, О. С. Середин, А. И. Татарчук, П. А. Турков, М. А. Суворов, А. И. Майсурадзе);
- интеллектуальный анализ экспериментальных данных (С. Д. Двоенко, Н. И. Боровых);
- обработка, анализ, классификация и распознавание изображений (А. Л. Жизняков, К. В. Жукова, И. А. Рейер, Д. М. Мурашов, Н. Г. Федотов, В. Ю. Мартьянов, М. В. Харинов).

Некоторые технические детали

- Для вычисления предлагаемых оценок приведение слов к начальной форме выполнялось с помощью функции $getNormalForms$ в составе *библиотеки русской морфологии*.
- Выделение синтаксических связей реализовано на основе правил, задействованных в работе [Царьков С. В., Естественные и технические науки, 2012, № 6].
- Распознавание границ предложений в тексте по знакам препинания — с помощью обученной модели классификатора, построенного с применением интегрированного пакета *Apache OpenNLP*.
- Обучение распознаванию границ предложений — на основе размеченных данных из *Leipzig Corpora* (газетные тексты на русском языке, 2010 г., всего 10^6 фраз).

Исходные фразы для формирования единиц экспертных знаний

№	Исходная фраза
1	*Переобучение приводит к заниженности эмпирического риска.*
2	*Переподгонка приводит к заниженности эмпирического риска.*
3	*Переподгонка служит причиной заниженности эмпирического риска.*
4	*Заниженность эмпирического риска является результатом нежелательной переподгонки.*
5	*Переусложнение модели приводит к заниженности средней ошибки на тренировочной выборке.*
6	*Переподгонка приводит к увеличению частоты ошибок дерева принятия решений на контрольной выборке.*
7	*Переподгонка приводит к заниженности оценки частоты ошибок алгоритма на контрольной выборке.*
8	*Заниженность оценки ошибки распознавания связана с выбором правила принятия решений.*
9	*Рост числа базовых классификаторов ведёт к практически неограниченному увеличению обобщающей способности композиции алгоритмов.*

Программная реализация и экспериментальные материалы

Исходные фразы в составе групп

№	Группа исходных фраз	
1	*Нежелательная переподгонка является причиной заниженности средней величины ошибки алгоритма на обучающей выборке.*	
	Переобучение приводит к заниженности эмпирического риска.	(2.1)
2	*Определение модели представления знаний накладывает ограничения на выбор соответствующего механизма логического вывода.*	(1.1)
	Однородность представления знаний приводит к упрощению механизма управления логическим выводом и упрощению управления знаниями.	(1.7)
3	*Эвристика основывается на собственном опыте специалиста в данной предметной области, накопленном в результате многолетней практики.*	(1.5)
	Метазнания могут касаться свойств, структуры, способов получения и использования знаний при решении практических задач искусственного интеллекта.	(1.6)

Примечание

Первая цифра в номере справа от фразы обозначает предметную область (1 — Философия и методология инженерии знаний, 2 — Математические методы обучения по прецедентам), вторая — порядковый номер исходной фразы по таблице (слайды 10 и 13).

Оценка релевантности текстового корпуса исходным единицам знаний

№ исх. фразы (группы фраз)[1]		с учётом предлогов/союзов	без учёта предлогов/союзов
		по отдельным фразам	
1	Философия и методология инженерии знаний	0,1443376	0,0861601
2		0,1423988	0,0643456
3		0,3995547	0,5083567
4		0,1513025	0,1650242
5		0,6166341	0,3633269
6		0,1591293	0,1621076
7		0,2127629	0,0326510
8		0,2393714	0,1471097
9		0,5758868	0,3178877
		по группам фраз	
2		0,1259124	0,0666667
3		0,3120782	0,4472640
		по отдельным фразам	
1	Математические методы обучения по прецедентам	0,6517818	0,2905786
2		0,5433360	0,2905786
3		0,2066957	0,2066957
4		0,1962131	0,1962131
5		0,3398426	0,0599116
6		0,2031058	0,2676248
7		0,2507539	0,3768646
8		0,2621604	0,2166871
9		0,1825379	0,1977494
		по группам фраз	
1		0,5516767	0,6094707

[1] Выделение связей слов — без привлечения базы синтаксических правил

Добавление фраз аннотации в число исходных

Релевантность текстового корпуса единице знаний и ситуации языкового употребления, оценка (6)

с учётом предлогов и союзов	без учёта предлогов и союзов
Исходная фраза №9, «Философия и методология инженерии знаний»	
0,5758868	0,3178877
Шаг 1	
«Специфика структурно-фреймовой организации состоит в том, чтобы во фрейме (а он представляет собой достаточно сложную концептуальную конструкцию, записанную средствами программной части вычислительной (информационной) системы) все понятия, относящиеся к охватываемой данным фреймом предметной области, имели внутреннюю интерпретацию, т. е. были наделены смыслом на соответствующем языке представления знаний».	
1,5439084	1,8877527
Шаг 2	
«Каждое выражение, входящее во фрейм, каждый знак в нём, несущий самостоятельную информационную нагрузку, являются интерпретированными, т. е. заключают в себе смысл, заложенный человеком с помощью соответствующей программы или же сконструированный системой».	
0,9197011	3,8257502
Попытка Шага 3	
«Перспективы развития эффективных систем представления знаний на основе естественного языка, т. е. вербальных знаковых систем, понятных человеку без особо сложной выучки, сегодня во многом связаны с построением так называемых фреймов».	
3,5916774	0,2568324

Численная оценка близости фразы смысловому эталону: основные эмпирические соображения

1. Близость фразы эталону следует оценивать по результатам разбиения её слов на классы по значению меры TF-IDF в совокупности с оценкой силы связи сочетаний слов в составе фразы.
2. Оценку силы связи слов следует вычислять относительно не отдельных текстов, а всего рассматриваемого тематического текстового множества (корпуса).
3. Учитывая требование минимизации длины фразы, актуальным здесь является рассмотрение только тех связей, которые имеют синтаксическую природу.
4. Разделение слов исходной фразы на общую лексику и термины по значению меры TF-IDF должно быть выражено как можно в большей степени.
5. Слова в кластерах, сформированных по TF-IDF слов исходной фразы относительно некоторого документа корпуса, должны быть распределены более или менее равномерно.

Пусть $H_1, \dots H_r$ — последовательность кластеров, сформированных по TF-IDF слов исходной фразы относительно документа d в составе корпуса D.

Документы $d \in D$ сортируются по убыванию произведения оценок:

$$val_1 = -\frac{1}{\log_{10}\left[\sqrt{\Sigma^2_{H_1} + \Sigma^2_{H_{r/2}} + \Sigma^2_{H_r}}\right]} \tag{7}$$

и, соответственно,

$$val_2 = 10^{-\sigma\left(\left|H_i,\, i=1,\dots,r\right|\right)}, \tag{8}$$

где $\Sigma^2_{H_1}$, $\Sigma^2_{H_{r/2}}$ и $\Sigma^2_{H_r}$ есть квадраты сумм значений TF-IDF слов, отнесённых, соответственно, к кластерам H_1, $H_{r/2}$ и H_r;

$\sigma\left(\left|H_i,\, i = 1, \dots, r\right|\right)$ — СКО числа элементов в кластере.

Примечание

Помимо H_1 и $H_{r/2}$, содержательный интерес здесь также представляет кластер H_r, которому соответствуют *слова-термины, преобладающие* в корпусе.

Для оценки близости эталону для каждой фразы Ts_i группы исходных Ts берётся (val_1, val_2) по $d \in D$, получившему наибольшее $val_1 \cdot val_2$, далее val_1 и val_2 делятся на свои максимумы по Ts и приводятся к $[0, 1]$.

Нормированные val_1 и val_2 далее обозначим как val'_1 и val'_2.

Сила связи слов и оценка близости фразы эталону

Рассмотрим следующий вариант оценки (3) силы связи слов A и B, K'_{AB}:

- вводится требование наличия синтаксической связи слов A и B;
- значения a, b и k в (3) вычисляются относительно всего корпуса D;
- оценка вычисляется только при принадлежности значения TF-IDF минимум одного из слов (A, B) либо кластеру $H_1(X)$, либо $H_{r/2}(X)$.

Пусть

$R(Ts_i, d)$ — множество связей слов фразы Ts_i, для которых определена оценка K'_{AB} относительно документа d;

$R_1(Ts_i, d)$ — множество связей, отнесённых к кластеру наибольших значений указанной величины;

$K_1(Ts_i, d)$ — сумма значений оценки K'_{AB} для связей из $R_1(Ts_i, d)$.

Тогда оценка близости фразы Ts_i эталону относительно $d \in D$ определяется аналогично ранжированию документов по релевантности исходной фразе как

$$W^R(Ts_i, d) = K_1(Ts_i, d) \frac{\left|R_1(Ts_i, d)\right|}{\left|R(Ts_i, d)\right|}. \qquad (9)$$

Максимальное значение оценки (9) для Ts_i по всем $d \in D$, приводимое к $[0, 1]$ делением на свой максимум по всем $Ts_i \in \mathrm{Ts}$, далее обозначим как val'_3.

«Неэталонные» фразы

Пусть Val' — тройка значений $val'_1 \cdot val'_2$, val'_3 с учётом и val'_3 без учёта предлогов/союзов для отдельной фразы Ts_i из группы исходных Ts.

Введём в рассмотрение СКО, разность $(\max - \min)$ и частное $(\max / \min)$ наибольшего и наименьшего значения в Val', далее — *СКО-оценки*.

Фраза не относится к «эталонным», если:

- по одной из величин $val'_1 \cdot val'_2$ либо val'_3 фраза относится к кластеру наибольших, по другой — к кластеру наименьших значений; при этом одна из СКО-оценок попадает в кластер наибольших значений;
- одновременно по значениям $val'_1 \cdot val'_2$ и val'_3 (как с учётом предлогов и союзов, так и без таковых) фраза относится к кластеру наименьших значений;
- по значениям любой из СКО-оценок фраза относится, но ни по одной из величин $val'_1 \cdot val'_2$ и val'_3 — не относится к кластеру наибольших значений.

В итоге *определять эталон* будут фразы из числа отнесённых к кластерам наибольших значений $val'_1 \cdot val'_2$ и val'_3, при этом *из рассмотрения исключаются* попадающие под одно из трёх вышеперечисленных правил.

№	Фраза из множества эквивалентных по смыслу
1	*Нежелательная переподгонка является причиной заниженности средней величины ошибки алгоритма на обучающей выборке.*
2	*Нежелательное переобучение приводит к заниженности эмпирического риска.*
3	*Заниженность эмпирического риска является следствием нежелательного переобучения.*
4	*Нежелательное переобучение служит причиной заниженности эмпирического риска.*
5	*Заниженность эмпирического риска является результатом нежелательного переобучения.*
6	*Заниженность эмпирического риска связана с переобучением.*
7	*Заниженность эмпирического риска относится к следствию нежелательного переобучения.*
8	*Заниженность эмпирического риска связана с нежелательным переобучением.*
9	*Нежелательное переобучение является причиной заниженности эмпирического риска.*
10	*Нежелательная переподгонка приводит к заниженности эмпирического риска.*
11	*Нежелательная переподгонка, следствием которой является заниженность эмпирического риска.*
12	*Эмпирический риск, к заниженности которого ведёт нежелательная переподгонка.*
13	*Риск, заниженный как следствие переподгонки.*

Исходные данные для оценивания близости смысловому эталону

№ фразы	Максимальные по текстовому корпусу значения оценки (7)	оценки (8)	оценки (9) без учёта предлогов и союзов	оценки (9) с учётом предлогов и союзов
1	0,4671832	0,0802703	0,3294206	0,8005671
2	0,4314354	0,3162278	0,2000000	0,5555556
3	0,4317240	0,2646365	0,1428571	0,1428571
4	0,4314354	0,3162278	0,1666667	0,1666667
5	0,4102928	0,3906175	0,1428571	0,1428571
6	0,4313110	0,2646365	0,3333333	0,2500000
7	0,4318168	0,1525820	0,1428571	0,1666667
8	0,4313110	0,3162278	0,3333333	0,2500000
9	0,4232485	0,2833201	0,1666667	0,1666667
10	0,4314424	0,3162278	0,0444445	0,2000000
11	0,4188872	0,1275184	0,0555556	0,0555556
12	0,4317911	0,1525820	0,2222222	0,1666667
13	0,3896160	0,3162278	0,0416667	0,0416667

Кластеризация фраз из представленных на слайде 21

№ кластера	*Номера фраз на слайде 21, отнесённых к кластеру (в порядке убывания оценки)*
на основе величины $val'_1 \cdot val'_2$	
1	5, 10, 2, 4, 8, 13, 9, 3, 6
2	7, 12, 11
3	1
на основе величины val'_3 *с учётом предлогов и союзов*	
1	1, 2
2	6, 8
3	10, 4, 7, 9, 12, 3, 5
4	11, 13
на основе $\max - \min$ *в* Val'	
1	1
2	5, 6, 8, 13, 10, 4, 12, 9, 3
3	11, 7
4	2

№ кластера	*Номера фраз на слайде 21, отнесённых к кластеру (в порядке убывания оценки)*
на основе СКО для значения элементов в Val'	
1	1
2	5, 8, 6, 13, 10, 4, 12, 9, 3
3	7, 11
4	2
на основе величины val'_3 *без учёта предлогов и союзов*	
1	6, 8, 1
2	12
3	2, 4, 9, 3, 5, 7
4	11, 10, 13
на основе $\max / \min$ *в* Val'	
1	13
2	10, 5, 1
3	11, 4, 3, 6, 8, 12, 9
4	7, 2

Итеративно отбираемые фразы в состав множества исходных для примера на слайде 16

№ фразы	Максимальные по текстовому корпусу значения оценки (7)	оценки (8)	оценки (9) без учёта предлогов и союзов	оценки (9) с учётом предлогов и союзов
1	*«Язык представления знаний на основе фреймовой модели наиболее эффективен для структурного описания сложных понятий и решения задач, в которых в соответствии с ситуацией желательно применять различные способы вывода».*	0,4809750	$1{,}480 \cdot 10^{-4}$	0,4866667	0,1041667
2	*«Специфика структурно-фреймовой организации состоит в том, чтобы во фрейме (а он представляет собой достаточно сложную концептуальную конструкцию, записанную средствами программной части вычислительной (информационной) системы) все понятия, относящиеся к охватываемой данным фреймом предметной области, имели внутреннюю интерпретацию, т. е. были наделены смыслом на соответствующем языке представления знаний».*	0,5740613	$2{,}854 \cdot 10^{-4}$	0,1477941	0,1322368
3	*«Каждое выражение, входящее во фрейм, каждый знак в нём, несущий самостоятельную информационную нагрузку, являются интерпретированными, т. е. заключают в себе смысл, заложенный человеком с помощью соответствующей программы или же сконструированный системой».*	0,5709775	$7{,}693 \cdot 10^{-4}$	2,1250000	0,2222222

Кластеризация фраз из представленных на слайде 24

№ кластера	*Номера фраз на слайде 21, отнесённых к кластеру (в порядке убывания оценки)*
на основе величины $val'_1 \cdot val'_2$	
1	3
2	2
3	1
на основе величины val'_3 *с учётом предлогов и союзов*	
1	3
2	2, 1
на основе $\max - \min$ *в* Val'	
1	2
2	1
3	3

№ кластера	*Номера фраз на слайде 21, отнесённых к кластеру (в порядке убывания оценки)*
на основе СКО для значения элементов в Val'	
1	2, 1
2	3
на основе величины val'_3 *без учёта предлогов и союзов*	
1	3
2	1
3	2
на основе $\max / \min$ *в* Val'	
1	2
2	1
3	3

Сокращение объёма текстовой информации для передачи единицы знаний

Оценивается как $(l_1 \cdot n_1)/(l_2 \cdot n_2)$, где n_1 — число фраз, представляющих единицу знаний, из которых n_2 определяют эталон; l_1 и l_2 — максимальная длина фразы (в словах) из задающих единицу знаний и из определяющих эталон.

Выводы

1. Основной *результат* настоящей работы — *метод* оценки близости фразы естественного языка смысловому эталону относительно представляемой ей единицы знаний.
2. Очевидное *преимущество* предложенного метода — *отсутствие необходимости* описания как можно большего числа СЭ-форм выражения соответствующей единицы знаний в языке.
3. Предложенный метод выделения смысловых эталонов даёт *минимум двукратное* сокращение *объёма текстовых данных*, необходимых *для передачи единицы знаний* посредством заданного естественного языка без потери полезной составляющей.
4. Открытая проблема — *результаты работы* предложенных решений *существенно зависят* от подбора тематического корпуса экспертом. При этом учитывается и уровень сложности отбираемого в корпус текста, и его значимость в решаемой задаче.
5. Представляет интерес *исследование* динамики изменения оценки (3) по документам корпуса для синтаксически связанных слов исходной фразы.