Мера TF-IDF, сила связи слов и формирование единиц представления знаний в открытых тестах

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

Новгородский государственный университет
имени Ярослава Мудрого

Всероссийская конференция с международным участием
«Математические методы распознавания образов» (ММРО-18),

9–13 октября 2017 г.

г. Таганрог, Ростовская обл.

Единица знаний, оцениваемая открытым тестом

Определяется множеством семантически эквивалентных фраз предметно-ограниченного естественного языка (ЕЯ).

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.

Основные проблемы:

- выделение единиц знаний из текстов тематического корпуса;
- отбор текстов в корпус анализом релевантности исходной фразе;
- полнота отражения в исходных фразах выделяемого фактического знания.

Предмет исследования

Методы и алгоритмы формирования знаний на основе текстового корпуса.

Задачи эксперта, требующие автоматизации

1. Поиск эквивалентных по смыслу форм выражения отдельного фрагмента фактического знания в заданном естественном языке.
2. Сопоставление фрагментов собственных знаний эксперта с наиболее близкими фрагментами знаний других экспертов.

Требования к решению

1. Выделение из текста понятий и отношений между ними.
2. Выявление в текстовом корпусе контекстов использования общей лексики, обеспечивающей синонимичные перифразы.

Образ исходной фразы в анализируемом тексте

- Фрагмент анализируемого текста, отвечающий составляющей образа, отождествим с некоторой смысловой связью слов в исходной фразе.
- Сила связи слов каждого такого фрагмента всегда больше силы связи любого слова данного фрагмента и слова, не принадлежащего ему.
- Сочетания общей лексики и терминов, преобладающих в корпусе, в анализируемом тексте можно отнести к составляющим искомого образа только при наличии фрагментов с большей силой связи слов.
- В общем случае не выдвигается требование наличия в тексте строго заданной части составляющих образа исходной фразы.
- Допускаются связи слов из различных фраз в группе исходных, взаимно эквивалентных либо дополняющих друг друга по смыслу и представляющих единый образ.

Основные проблемы:

- ограничение рассмотрения связей слов биграммами и рамками синтаксиса естественного языка;
- точность выделения фрагмента знаний (понятий и их связей) при единственной исходной фразе.

Наиболее близкие идеи:

- синтаксические n-граммы [Grigori Sidorov, 2013];
- поиск синтаксически связанных групп соседних слов с помощью условных случайных полей [Кудинов М. С., 2013].

Базовые предположения:

- пути в деревьях зависимостей либо деревьях составляющих как основу выделения n-грамм следует отсчитывать не от вершины дерева, а от сочетаний слов с наибольшими значениями силы связи;
- внутри связанных фрагментов текста допускается наличие предлогов и союзов.

Выбор оценки силы связи слов

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

$$K_{AB} = \frac{k}{a + b - k}, \tag{1}$$

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

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

$$\mathrm{tf}\,(t_i, d) = \frac{n_i}{\sum_k n_k}, \tag{2}$$

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

$$\mathrm{idf}\,(t_i, D) = \log\left(\frac{|D|}{|D_i|}\right), \tag{3}$$

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 значению 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 $\Leftrightarrow$ являться следствием»*.

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, 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 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:

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

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

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

Аналогично, но по значению функции (4), разбивается $\mathbb{T}\,(d)$ для $\forall\, d \in D'$, $\mathbb{T}'\,(d)$ — кластер наибольших значений оценки (4) по заданному d.

При этом для фразы s документа $d \in D'$ возможны два варианта оценки

$$\mathrm{N}\,(s) = \Big|\big\{w \in b\colon\ \exists\, T \in \mathbb{T}'\,(d)\,,\, b \in T\big\}\Big| \tag{6}$$

и, соответственно,

$$\mathrm{N}\,(s) = \Big|\big\{b\colon\ \exists\, T \in \mathbb{T}'\,(d)\,,\, b \in T\big\}\Big| \tag{7}$$

как основа разбиения на кластеры всего множества $\big\{s\colon\ s \in d \,\big|\, d \in D'\big\}$.

Фразы аннотации

Составляют первый кластер из полученных по значению величины $\mathrm{N}\,(s)$.

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:

$$\mathbb{W}(D) = \frac{1}{\left|D'\right|} \sum_{d \in D'} \left[\frac{\left|\left\{w \in b\colon\ \exists T \in \mathbb{T}'(d)\,,\, b \in T\right\}\right|}{\left|\left\{w\colon\ \exists Ts_i \in \mathrm{Ts},\, w \in Ts_i\right\}\right|} \sum_{T \in \mathbb{T}'(d)} \mathrm{N}(T, d) \right], \quad (8)$$

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).

Основные критерии

1. Исходные фразы формулируются независимо друг от друга разными экспертами.
2. Исходные множества текстов подбираются так, чтобы сравнить образы исходной фразы, выделяемые в текстах:
   - для отдельных исходных фраз и их совокупности с учётом возможных межфразовых связей;
   - оценкой силы связи слов пары и последовательностей таких пар в составе n-грамм;
   - анализом синтаксических зависимостей и привлечением меры TF-IDF при поиске связей слов.
3. Максимально полная и наглядная иллюстрация выявления в текстах контекстов использования как слов-терминов, так и общей лексики, обеспечивающей синонимические перифразы исходной фразы.

Исходное множество 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 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 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 результаты экспериментов

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 14 и 17).

далее к примерам

Отбор релевантных фраз для групп исходных фраз на основе n-грамм[1]

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
с привлечением TF-IDF, оценка (6)								*синтаксические правила, оценка (6)*						
2	1	0	0	1	0	0	1	16	2	0	3	0	0	3
3	1	1	1	1	1	3	2	3	1	1	2	1	1	3
с привлечением TF-IDF, оценка (7)								*синтаксические правила, оценка (7)*						
2	3	1	0	1	1	0	1	4	0	0	2	0	0	1
3	2	1	2	2	1	5	5	2	0	0	1	0	0	1

Здесь:

N — общее число отобранных фраз;

N_1 — число фраз, представляющих выразительные средства языка;

N_2 — число фраз, представляющих синонимы;

N_3 — число фраз, представляющих связи понятий предметной области;

N_1^1 — число представляемых в найденных фразах выразительных средств языка;

N_2^1 — число представляемых в найденных фразах синонимов;

N_3^1 — число представляемых связей для понятий из упомянутых в исходных фразах.

[1] Здесь и далее на *слайдах 20–29* учитываются сочетания в т. ч. с предлогами и союзами

Отбор релевантных фраз для групп исходных фраз на основе n-грамм

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
Отбор релевантных для отдельных фраз групп №2 и №3														
с привлечением TF-IDF, оценка (6)								*синтаксические правила, оценка (6)*						
1.1	2	0	0	2	0	0	1	5	0	0	3	0	0	3
1.5	3	0	1	2	0	1	2	3	0	2	1	0	2	1
1.6	3	0	0	1	0	0	1	3	1	0	2	1	0	2
1.7	3	0	0	1	0	0	2	3	0	0	2	0	0	2
с привлечением TF-IDF, оценка (7)								*синтаксические правила, оценка (7)*						
1.1	1	0	0	1	0	0	1	4	0	0	2	0	0	2
1.5	1	0	1	1	0	1	1	1	0	1	1	0	1	1
1.6	10	1	0	4	1	0	3	1	1	0	1	1	0	1
1.7	3	0	0	1	0	0	2	1	0	0	0	0	0	0

N — общее число отобранных фраз; N_2 — фраз, представляющих синонимы;

N_1 — выразительные средства языка; N_3 — связи понятий;

N_1^1 — число представляемых в найденных фразах выразительных средств языка;

N_2^1 — число представляемых в найденных фразах синонимов;

N_3^1 — число представляемых связей для понятий из упомянутых в исходных фразах.

Отбор релевантных для фраз групп №2 и №3 по числу «наиболее сильных» связей

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
	с привлечением TF-IDF							*на основе синтаксических правил*						
1.1	1	0	0	1	0	0	1	2	0	0	1	0	0	1
1.5	2	2	2	0	2	2	0	4	0	0	2	0	0	5
1.6	6	1	1	1	1	1	1	1	1	1	0	1	1	0
1.7	6	0	0	2	0	0	3	6	0	0	2	0	0	3

Здесь:

N — общее число отобранных фраз;

N_1 — число фраз, представляющих выразительные средства языка;

N_2 — число фраз, представляющих синонимы;

N_3 — число фраз, представляющих связи понятий предметной области;

N_1^1 — число представляемых в найденных фразах выразительных средств языка;

N_2^1 — число представляемых в найденных фразах синонимов;

N_3^1 — число представляемых связей для понятий из упомянутых в исходных фразах.

Отбор релевантных фраз для групп исходных фраз

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
с привлечением TF-IDF, оценка (6)								*синтаксические правила, оценка (6)*						
1	3	1	1	1	1	1	2	1	0	0	1	0	0	2
с привлечением TF-IDF, оценка (7)								*синтаксические правила, оценка (7)*						
1	1	0	1	1	0	1	2	1	0	0	1	0	0	2
Отбор релевантных для отдельных фраз группы №1														
с привлечением TF-IDF, оценка (6)								*синтаксические правила, оценка (6)*						
2.1	1	1	0	0	1	0	0	1	1	0	0	1	0	0
с привлечением TF-IDF, оценка (7)								*синтаксические правила, оценка (7)*						
2.1	1	1	0	0	1	0	0	1	1	0	0	1	0	0
по числу «наиболее сильных» связей из выделенных														
с привлечением TF-IDF								*на основе синтаксических правил*						
2.1	2	0	0	1	0	0	1	1	1	0	0	1	0	0

N — общее число отобранных фраз; N_2 — фраз, представляющих синонимы;

N_1 — выразительные средства языка; N_3 — связи понятий;

N_1^1 — число представляемых в найденных фразах выразительных средств языка;

N_2^1 — число представляемых в найденных фразах синонимов;

N_3^1 — число представляемых связей для понятий из упомянутых в исходных фразах.

Пример: выделение составляющих образа для группы исходных фраз №3

На основе n-грамм, без привлечения базы синтаксических правил, оценка (6):

Отбираемая фраза	Что представляет	Слова наиболее значимых n-грамм
Эвристика может пониматься как: — *научно-прикладная дисциплина, изучающая творческую деятельность;* — *приёмы решения проблемных (творческих, нестандартных, креативных) задач в условиях неопределённости, которые обычно противопоставляются формальным методам решения, опирающимся, например, на точные математические алгоритмы;* — *метод обучения;* — *один из способов создания компьютерных программ — эвристическое программирование.*	**Связь понятий** *эвристика – знание* **с** *приёмами решения задач,* **перифраза** *в результате* $\Longleftrightarrow$ *как результат,* **синонимы** *способ* $\Longleftrightarrow$ *приём, опираться* $\Longleftrightarrow$ *основываться, практический* $\Longleftrightarrow$ *прикладной*	*эвристика, в, задача, на, способ, решение, мочь*

Отбор релевантных фразе (1.6) по числу «наиболее сильных» связей из выделяемых по TF-IDF:

Отбираемая фраза	Что представляет	«Наиболее сильные» связи
Стремление преодолеть узость алгоритмического подхода привело к возникновению эвристического направления в разработке проблем искусственного интеллекта, где эвристика понимается как термин, противостоящий понятию алгоритма, который представляют собой «набор инструкций или четко сформулированных операций, составляющих определенную процедуру».	**Связь понятия** *искусственный интеллект* **из исходной фразы с понятием** *эвристика*	*искусственный – интеллект*

Пример: «Философия и методология инженерии знаний», фраза №3

На основе n-грамм, без привлечения базы синтаксических правил, оценка (6):

Отбираемая фраза	Что представляет
При этом модель знания понималась как формализованная в соответствии с определенными структурными планами информация, сохраняемая в памяти, и которая может быть им использована в ходе решения задач на основании заранее запрограммированных схем и алгоритмов.	**Соотнесение понятия** *знания* из исходной фразы **с понятием** *модели знания*, **перифраза** *определяется как* $\Longleftrightarrow$ *понимается как*

По числу «наиболее сильных» связей из выделяемых по TF-IDF:

Отбираемая фраза	Что представляет
Согласно Дж. фон Нейману, информация имеет двоякую природу: она может трактоваться как программа или алгоритм по работе с данными и как информация об объектах, т. е. те данные, с которыми программа работает.	**Связи для понятия** *информация*
Информация представляет собой закодированное в эксплицитной форме знание, по которому человек способен творчески его воссоздать.	
При этом модель знания понималась как формализованная в соответствии с определенными структурными планами информация, сохраняемая в памяти, и которая может быть им использована в ходе решения задач на основании заранее запрограммированных схем и алгоритмов.	**перифраза** *определяется как* $\Longleftrightarrow$ *понимается как*

Оценка релевантности текстового корпуса исходным единицам знаний

№ исх. фразы (группы фраз)[2]	с учётом предлогов/союзов/междометий	без учёта предлогов/союзов/междометий
	Философия и методология инженерии знаний	
	по отдельным фразам	
1	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	
3	0,3120782	0,4472640
	Математические методы обучения по прецедентам	
	по отдельным фразам	
1	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

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

Сравнение наиболее значимых связей и n-грамм для отбора фраз (без привлечения синтаксических правил, оценка — по числу слов)

№ исх. фразы	Слова, не вошедшие в наиболее значимые связи	n-граммы
	Философия и методология инженерии знаний	
2	*знание*	*и, на, с*
3	*знание, с, или, использовать*	
4		*на*
5	*на, собственный, опыт, область*	
6	*и*	
7	*представление, и*	
8	*реализация, система, и, возможность*	
9	*с, понятие, структурный, соответствие, представление, в, ситуация*	*различный*
	Математические методы обучения по прецедентам	
2	*заниженность*	
3	*заниженность, причина*	
4	*заниженность, являться*	
5	*к, средний*	
6	*приводить, к*	*принятие, решение*
8		*принятие*

В данной иллюстрации сравнение ведётся по тем документам, которые вошли в число *наиболее релевантных* исходной фразе при ранжировании как на основе n-грамм, так и «наиболее сильных» связей.

Сравнение наиболее значимых связей и n-грамм для отбора фраз (без привлечения синтаксических правил, оценка — по числу слов)

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
	максимизацией числа «наиболее сильных» связей по K_{AB}									*анализом n-грамм на найденных связях слов*								
Философия и методология инженерии знаний																		
N	1	2	11	1	2	6	6	6	1	2	4	4	2	3	3	3	2	3
N_1	0	0	1	1	2	1	0	0	0	0	0	1	1	0	0	0	1	0
N_2	0	0	2	0	2	1	0	1	0	0	0	0	1	1	0	0	0	0
N_3	1	0	5	1	0	1	2	3	1	2	2	3	0	2	1	1	2	3
Математические методы обучения по прецедентам																		
N	2	1	15	15	5	1	6	1	1	1	1	2	1	2	1	9	2	6
N_1	0	1	3	2	0	0	0	0	1	1	1	2	1	0	1	1	0	1
N_2	0	1	2	2	1	0	1	0	1	0	1	2	1	1	1	0	0	0
N_3	1	0	7	4	0	0	0	1	0	0	0	1	0	1	1	5	1	4

Здесь:

N — общее число отобранных фраз;

N_1 — число фраз, представляющих выразительные средства языка;

N_2 — число фраз, представляющих синонимы;

N_3 — число фраз, представляющих связи понятий предметной области.

Сравнение наиболее значимых связей и n-грамм для отбора фраз (с привлечением синтаксических правил, оценка — по числу слов)

№ исх. фразы	Слова, не вошедшие в наиболее значимые связи	n-граммы
	Философия и методология инженерии знаний	
1	*знание, выбор*	
2	*основать, организация*	
5	*в, на, специалист, результат, практика, область*	*опыт, накопить*
8	*определение, возможность*	
	Математические методы обучения по прецедентам	
2	*заниженность*	
3	*заниженность*	
4	*заниженность, являться*	
5	*средний*	
6	*приводить, к*	
9	*алгоритм, к*	

В данной иллюстрации сравнение также ведётся по документам, которые вошли в число *наиболее релевантных* исходной фразе при ранжировании как на основе n-грамм, так и «наиболее сильных» связей.

Сравнение наиболее значимых связей и n-грамм для отбора фраз (с привлечением синтаксических правил, оценка — по числу слов)

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
	максимизацией числа «наиболее сильных» связей по K_{AB}									*анализом n-грамм на найденных связях слов*								
Философия и методология инженерии знаний																		
N	2	4	1	3	4	1	6	1	5	5	2	19	7	3	3	3	1	1
N_1	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1	0	1	0
N_2	0	0	0	2	0	1	0	0	0	0	0	3	0	2	0	0	0	0
N_3	1	2	0	1	2	0	2	0	2	3	1	4	4	1	2	2	0	1
Математические методы обучения по прецедентам																		
N	1	1	15	15	5	11	1	1	1	1	1	2	1	2	1	9	4	1
N_1	1	1	3	2	0	0	0	0	1	1	1	2	1	0	0	1	0	0
N_2	0	1	2	2	1	9	0	0	1	0	1	2	1	2	1	0	0	0
N_3	0	0	7	4	0	4	0	1	0	0	0	1	0	2	0	3	0	0

Здесь:

N — общее число отобранных фраз;

N_1 — число фраз, представляющих выразительные средства языка;

N_2 — число фраз, представляющих синонимы;

N_3 — число фраз, представляющих связи понятий предметной области.

Альтернативное решение: поиск фраз на готовом синтаксически размеченном текстовом корпусе

Слова и их сочетания для отбора фраз из *Национального корпуса русского языка*:

№	Слова и сочетания слов
Философия и методология инженерии знаний	
1	*модель – представление – знание, механизм – логический – вывод*
2	*система – суждение, объективный – закономерность*
3	*процесс – логический – вывод*
4	*данный – предметный – область*
5	*эвристика, данный – предметный – область*
6	*метазнания, свойство – знание, структура – знание, способ – получение – знание, способ – использование – знание, задача – искусственный – интеллект*
7	*представление – знание, управление – вывод, механизм – логический – вывод, управление – знание*
8	*теоретический – обоснование – модель, логический – модель, система – вывод, система – определение, точный – вывод*
9	*язык – представление – знание, фреймовый – модель, способ – вывод*

№	Слова и сочетания слов
Математические методы обучения по прецедентам	
1	*переобучение, эмпирический – риск*
2	*эмпирический – риск*
3	*эмпирический – риск*
4	*эмпирический – риск*
5	*ошибка – средний*
6	*частота – ошибка, контрольный – выборка*
7	*оценка – частота, контрольный – выборка*
8	*ошибка – распознавание, правило – принятие – решение*
9	*базовый – классификатор*

Пример для Национального корпуса русского языка (продолжение)

№	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
	Философия и методология инженерии знаний									*Математические методы обучения по прецедентам*								
N	13	73	2	15	83	33	79	224	20	56	1	1	1	24	17	21	5	2
N_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N_3	2	5	0	1	5	3	3	2	2	0	0	0	0	0	0	0	1	0
N_3^1	2	6	0	2	4	3	3	2	2	0	0	0	0	0	0	0	1	0

Здесь:

N — общее число отобранных фраз;

N_1 — число фраз, представляющих выразительные средства языка;

N_2 — число фраз, представляющих синонимы;

N_3 — число фраз, представляющих связи понятий предметной области;

N_3^1 — число представляемых связей для понятий из упомянутых в исходных фразах.

Выводы

1. Основной *результат* настоящей работы — *метод* формирования тематического корпуса текстов, релевантных по описываемым фрагментам знаний совокупности исходных фраз, с выделением составляющих её образа в виде слов и их сочетаний.
2. По сравнению с поиском совокупностей указанных составляющих на готовом синтаксически размеченном корпусе, охватывающем весь заданный ЕЯ, предложенный метод *позволяет* в среднем в **17** раз сократить выход фраз, не релевантных исходным ни по описываемому фрагменту знания, ни по языковым формам его выражения.
3. Реализованный *вариант контекстно-зависимого аннотирования* ориентирован в первую очередь *на поиск форм выражения связей понятий* предметной области, в текстах которой доля общей лексики сравнима с долей терминов.
4. Открытая проблема — *скорость и точность морфологического анализа*. Здесь представляет интерес реализация предложенного в работе метода на языке Python с привлечением библиотеки *NLTK* и морфологического анализатора *Pymorphy* как альтернатива реализованному решению на базе *библиотеки русской морфологии*.