ВЫДЕЛЕНИЕ ЗНАНИЙ И ЯЗЫКОВЫХ ФОРМ ИХ ВЫРАЖЕНИЯ НА МНОЖЕСТВЕ ТЕМАТИЧЕСКИХ ТЕКСТОВ АНАЛИЗОМ СВЯЗЕЙ СЛОВ В СОСТАВЕ N-ГРАММ

Д.В. Михайлов [1], А.П. Козлов [1], Г.М. Емельянов [1]

[1] Новгородский государственный университет имени Ярослава Мудрого, Новгород, Россия

Аннотация

Статья посвящена взаимосвязанным проблемам выделения единиц знаний из множества (корпуса) тематических текстов анализом релевантности исходной фразе и полноты отражения в исходных фразах выделяемого фактического знания. Данные проблемы актуальны для построения систем обработки, анализа, оценивания и понимания информации. Конечной практической целью здесь является поиск наиболее рационального варианта передачи смысла средствами заданного естественного языка для последующей фиксации фрагментов знаний в тезаурусе и онтологии предметной области. При этом релевантность текста по описываемому фрагменту знания (включая формы выражения в языке) определяется совместным использованием оценки силы связи встречающихся в его фразах сочетаний слов исходной фразы и разбиением этих слов на классы по значению меры TF-IDF относительно текстов корпуса. В настоящей работе рассматривается расширение связей слов от традиционных биграмм до трёх и более элементов для выделения составляющих образа исходной фразы в виде сочетаний связанных по смыслу слов (с привлечением базы известных синтаксических отношений и без использования таковой). С целью более полного описания выделяемого в текстах корпуса фрагмента экспертного знания вводятся в рассмотрение совокупности исходных фраз, взаимно эквивалентных либо дополняющих друг друга по смыслу и представляющих единый образ. По сравнению с поиском составляющих рассматриваемого образа на готовом синтаксически размеченном текстовом корпусе 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.

Ключевые слова: распознавание образов, интеллектуальный анализ данных, теория информации, тест открытой формы, языковое представление экспертных знаний, контекстно-зависимое аннотирование, поисковое ранжирование документов.

Цитирование: **Михайлов, Д.В.** Выделение знаний и языковых форм их выражения на множестве тематических текстов анализом связей слов в составе *n*-грамм / Д.В. Михайлов, А.П. Козлов, Г.М. Емельянов // Компьютерная оптика. – 2017. – Т. 41, № 3. – С. 461-471 – DOI: 10.18287/2412-6179-2017-41-3-461-471.

Введение

Эффективность методов и алгоритмов распознавания образов и интеллектуального анализа данных во многом определяется спецификой решаемой задачи. Немаловажную роль при этом играет разработка способов и средств описания самих задач, в частности, если задача включает иерархию взаимосвязанных подзадач. Как уже отмечалось в [1], естественным источником знаний при описании задач здесь будут публикации отечественных и зарубежных научных школ по соответствующей проблематике. Актуальная при этом проблема – поиск наиболее рационального варианта передачи смысла в единице знаний, определяемой множеством семантически эквивалентных (СЭ) фраз предметно-ограниченного естественного языка (ЕЯ). При этом в круг задач эксперта, требующих автоматизации, входит:

– поиск СЭ-форм выражения отдельного фрагмента фактического знания в заданном ЕЯ;
– сопоставление знаний эксперта с наиболее близкими фрагментами знаний других экспертов.

В работе [2] нами было предложено решение задачи поиска в текстовом множестве фраз, максимально близких исходной по описываемому фрагменту знания и формам его выражения в языке. При этом релевантность текста, в котором осуществляется поиск фраз, определялась суммарной численной оценкой силы связи встречающихся в его фразах сочетаний слов исходной фразы. Следует отметить, что данный метод ограничивал рассмотрение связей слов биграммами и рамками синтаксиса ЕЯ, что критично для случаев, когда доля общей лексики сравнима с долей слов-терминов (например, в текстах по близким искусственному интеллекту разделам философии науки и техники). Кроме того, в целях более точного описания представляемого фрагмента знаний (понятий и их связей) в ряде случаев число исходных фраз здесь целесообразно увеличивать до двух и более. В настоящей работе делается попытка решения данного круга проблем уходом от жёсткой ориентации на синтаксис ЕЯ с одновременным расширением выделяемых во фразах биграмм до трёх и более элементов.

1. N-граммы, мера TF-IDF и оценка силы связи слов

«Классические» N-граммы (по К. Шеннону – *L*-граммы, [3]) как последовательности из *n* элементов нашли широкое применение в математических исследованиях, биологии, а также информационном поиске. Наиболее близкими рассматриваемой в настоящей работе проблематике являются синтаксические *n*-граммы [4], которые определяются не линейной структурой текста, а путями в деревьях синтаксических зависимостей либо деревьях составляющих. Отметим, что при использовании силы связи слов исходной фразы отно-

сительно текста как основы оценки его релевантности последней такие пути следует отсчитывать не от вершины дерева, а от сочетаний слов с наибольшими значениями силы связи. В отличие от поиска синтаксически связанных групп соседних слов с помощью условных случайных полей [5], наличие внутри связанных фрагментов текста предлогов и союзов здесь не является критичным, что немаловажно для поиска в текстах языковых выразительных средств конструирования перифраз исходной фразы. В качестве оценки силы связи слов (A, B) возьмём оценку, показавшую наилучшие результаты в [2] и определяемую как

$$K_{AB} = k/(a+b-k), \quad (1)$$

где a – число фраз текста, которые содержат слово A, b – слово B, k – A и B одновременно. Из оценок силы связи слов в дистрибутивно-статистическом методе построения тезаурусов [6] данная оценка, содержательно близкая коэффициенту Танимото [7], наиболее наглядна, но в то же время учитывает встречаемость каждого слова в отдельности. За основу выделения самих связей в настоящей работе наряду с синтаксическими зависимостями из [2] в качестве альтернативы берётся разбиение слов исходной фразы по значению статистической меры TF-IDF [1].

Пусть X – упорядоченная по убыванию последовательность значений указанной меры для слов исходной фразы относительно заданного текста в составе некоторого множества (корпуса). Сама мера есть произведение отношения числа вхождений слова к общему числу слов документа и обратной частоты документа $idf(t_i, D) = \log(|D|/|D_i|)$, $|D_i \subset D|$ – число документов корпуса D, в которых слово t_i встретилось хотя бы один раз. Разобьём X на кластеры $H_1,\ldots,H_r$ с применением алгоритма, содержательно близкого алгоритмам класса FOREL [8]. За центр масс кластера H_i здесь, как и в работах [1, 2], мы возьмём среднее арифметическое всех $x_j \in H_i$.

Алгоритм 1. Формирование кластера.
Вход: X; // упорядоченная по убыванию
// числовая последовательность
Выход: H_i, X_p, X_s; // X_p, X_s – префикс/суффикс X
// относительно H_i, $X_p \bullet H_i \bullet X_s = X$
Начало
1: $H_i := X$;
2: $X_p := \varnothing$;
3: $X_s := \varnothing$;
4: **если** $\mathrm{good}(H_i) = true$ **или** $\mathrm{diam}(H_i) = 1$ **то**
вернуть H_i, X_p и X_s;
5: **иначе если** $|\mathrm{mc}(H_i) - \mathrm{first}(H_i)| > |\mathrm{mc}(H_i) - \mathrm{last}(H_i)|$ **то**
6: $X_p := \{\mathrm{first}(H_i)\} \bullet X_p$;
7: $H_i := \mathrm{rest}(H_i)$;
8: перейти к шагу 4;
9: **иначе если** $|\mathrm{mc}(H_i) - \mathrm{first}(H_i)| < |\mathrm{mc}(H_i) - \mathrm{last}(H_i)|$ **то**
10: $X_s := \{\mathrm{last}(H_i)\} \bullet X_s$;
11: $H_i := \mathrm{lrev}(H_i)$;
12: перейти к шагу 4;
13: **иначе**
14: $X_s := \{\mathrm{last}(H_i)\} \bullet X_s$;
15: $X_p := \{\mathrm{first}(H_i)\} \bullet X_p$;
16: $Tmp := \mathrm{lrev}(H_i)$;
17: $H_i := \mathrm{rest}(Tmp)$;
18: перейти к шагу 4;
Конец *{Алгоритм 1}*.

Здесь и далее "$\bullet$" – операция конкатенации.

Табл. 1. Вспомогательные функции Алгоритма 1

Функция	**Возвращаемое значение**
first(X), last(X)	первый/последний элемент последовательности X
lrev(X), rest(X)	исходная последовательность X без последнего/первого элемента
good(X)	**true** либо **false** в зависимости от выполнения условия (2)
mc(X)	центр масс X
diam(H_i)	диаметр кластера H_i

Так же, как и в [1, 2], элементы последовательности X будут отнесены к одному кластеру, если

$$\begin{cases} |\mathrm{mc}(X) - \mathrm{first}(X)| < \mathrm{mc}(X)/4 \\ |\mathrm{mc}(X) - \mathrm{last}(X)| < \mathrm{mc}(X)/4 \end{cases}. \quad (2)$$

Алгоритм 1 применяется к последовательностям X_p и X_s на его выходе. Данный процесс продолжается рекурсивно до тех пор, пока на очередном шаге X_p и X_s не будут пустыми. В итоге исходная последовательность X разбивается на подпоследовательности-кластеры $H_1,\ldots,H_r$, при этом для $\forall i \neq j \;\; H_i \cap H_j = \varnothing$, а $H_1 \bullet H_2 \bullet \ldots \bullet H_r = X$. Для выделения связей здесь важны слова кластера H_1 (термины из исходной фразы, наиболее уникальные для анализируемого текстового документа), а также «серединного» кластера $H_{r/2}$, куда войдёт общая лексика, обеспечивающая синонимические перифразы, и термины-синонимы. При этом оценка (1) для пары слов исходной фразы вычисляется только в том случае, если значение TF-IDF минимум одного из слов пары принадлежит либо H_1, либо $H_{r/2}$. Назовём далее такие слова связанными в паре по TF-IDF.

Пусть d – некоторый документ в составе исходного текстового множества D, а $L(d)$ есть последовательность пар слов (A, B) исходной фразы, где внутри каждой пары слова связаны в зависимости от метода выделения связей либо синтаксически, либо по TF-IDF, причём относительно d биграммы в $L(d)$ упорядочены по убыванию оценки (1).

Определение 1. Биграммы (A_1, B_1) и (A_2, B_2) войдут в одну n-грамму $T \subseteq L(d)$, если

$$((A_1 = A_2) \vee (B_1 = B_2) \vee (A_1 = B_2) \vee (B_1 = A_2)) = true\,.$$

После выделения очередной n-граммы T последовательность $L(d)$ просматривается с конца и из неё удаляются элементы, присутствующие исключительно в T, что определяется по наличию общих элементов у биграмм из T и из $L(d) \backslash T$. Данный процесс продолжается до тех пор, пока на очередном шаге $L(d)$ не окажется пустой. В худшем случае (максимальная длина n-граммы) здесь имеет место число сравнений, оцениваемое сверху как $|L(d)|^2 + |L(d)|$. Максимальная длина n-граммы при этом ограничена значением $|L(d)|$

и имеет место в случае выделения связей по TF-IDF при числе кластеров по указанной величине для слов исходной фразы, меньшем трёх. Оценка качества кластеризации слов по TF-IDF, предложенная в [1], в данной работе не применяется из соображений максимально полного учёта возможных сочетаний слов вне зависимости от значения меры TF-IDF каждого из них. Отметим, что в случае использования вышеупомянутой оценки качества кластеризации слов в совокупности с соответствующим ранжированием документов (требующим, однако, дополнительных вычислительных затрат) максимальная длина *n*-граммы сокращается приблизительно на одну треть.

Значимость *n*-граммы T длиной $\mathrm{len}(T)$ для оценивания ранга документа d относительно D можно определить из геометрических соображений как

$$\mathrm{N}(T,d)=\sqrt{\sum_{i=1}^{\mathrm{len}(T)}\left[\mathrm{S_i}(d)\right]^2}\Big/\left(\sigma(S_i(d))+1\right), \qquad (3)$$

где $S_i(d)$ – значение оценки (1) для i-й биграммы относительно d; $\sigma(S_i(d))$ – её среднеквадратическое отклонение (СКО); $\mathrm{len}(T)$ здесь и далее определяется числом биграмм в составе T. Добавление единицы в знаменателе формулы (3) имеет целью предотвратить деление на ноль в случае нулевого СКО.

Содержательно оценка (3) подразумевает максимизацию суммы силы связи слов при минимуме её СКО по всем связям слов в составе *n*-граммы. При этом связи не обязательно охватывают слова исключительно внутри одной фразы: допускаются связи слов из различных фраз в группе исходных, взаимно эквивалентных либо дополняющих друг друга по смыслу и представляющих единый образ.

2. Поисковое ранжирование документов и отбор фраз в аннотацию

Найденные *n*-граммы $\{T:T\subseteq L(d)\}=:\mathbf{T}(d)$ сортируются по убыванию величины $\mathrm{N}(T,d)\cdot\mathrm{len}(T)$. Функция ранжирования документов по релевантности исходной фразе (группе фраз) здесь может быть определена из соображений максимальной полноты представляемого в *n*-граммах образа исходной фразы как

$$W(d)=\mathrm{N}_{\max}(d)\cdot\log_{10}\left(\max_{T\in\mathbf{T}(d)}\mathrm{len}(T)\right)\cdot\log_{10}\left(\left|\mathbf{T}(d)\right|\right), \qquad (4)$$

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

Сортировкой документов $d\in D$ по убыванию значений функции (4) с последующим разделением на кластеры *Алгоритмом 1* отбираются документы с наибольшими значениями данной оценки (принадлежащими первому кластеру в составе формируемой последовательности). Множество указанных документов обозначим далее как D' (следуя нотации работ [1, 2]).

Аналогично, но по значению оценки (3), разбивается множество $\mathbf{T}(d)$ для $\forall d\in D'$, $\mathbf{T}'(d)$ – кластер наибольших её значений по заданному документу d. При этом возможны два альтернативных варианта оценки близости фразы s документа $d\in D'$ исходной фразе (группе исходных фраз): либо по числу слов

$$\mathrm{N}(s)=\left|\left\{w\in b:\exists T\in\mathbf{T}'(d),b\in T\right\}\right|, \qquad (5)$$

либо по числу биграмм

$$\mathrm{N}(s)=\left|\left\{b:\exists T\in\mathbf{T}'(d),b\in T\right\}\right|, \qquad (6)$$

в составе наиболее значимых *n*-грамм.

Идейно оценки (3) и (4) близки методу выделения многословных терминов C-Value [9] в плане предположения о достаточно большом числе биграмм в наиболее значимых *n*-граммах. Вместе с тем предложенный в настоящей работе метод не требует вычисления частот вхождения анализируемой *n*-граммы в состав других *n*-грамм. Действительно, отбираемая в соответствии с оценкой (3) *n*-грамма уже имеет максимально возможную длину, а требование принадлежности входящих в неё слов к терминам здесь не принципиально.

Назовём далее поиск фраз в документах $d\in D'$ на основе оценок (5) и (6) построением аннотации. Как и в работе [2], здесь имеет место разновидность контекстно-зависимого аннотирования [10], где одна аннотация строится сразу для нескольких документов.

Само построение аннотации ведётся по тому же алгоритму, который был использован авторами в [2].

Пусть $S=\{s:s\in d,\ d\in D'\}$, а X^W и X^{N} – последовательности значений оценок $W(d)$ и $\mathrm{N}(s)$ для документов $d\in D$ и фраз $s\in S$ соответственно. Будем применять обозначение X^{N} также и к последовательности значений оценки (3) для *n*-грамм. Кластеры, формируемые на последовательностях X^W и X^{N}, а также на D, S и $\mathbf{T}(d)$, обозначим далее как $H_1^W\ldots H_{r(D)}^W$, $H_1^{\mathrm{N}}\ldots H_{r(S)}^{\mathrm{N}}$ и $H_1^{\mathrm{N}}\bullet H_2^{\mathrm{N}}\bullet\ldots\bullet H_{r(\mathbf{T}(d))}^{\mathrm{N}}$ и, соответственно, $H_1^D\ldots H_{r(D)}^D$, $H_1^S\ldots H_{r(S)}^S$ и $H_1^T\bullet H_2^T\bullet\ldots\bullet H_{r(\mathbf{T}(d))}^T$. В совокупности с ранжированием текстов построение аннотации здесь формально можно представить следующим образом.

Алгоритм 2. Построение аннотации.

Вход: D; // исходное текстовое множество

Выход: H_1^S; // результирующее множество фраз

Начало

1: $X^W:=\varnothing$;
2: **для всех** $d\in D$
3: сформировать $\mathbf{T}(d)$;
4: упорядочить $\{T:T\in\mathbf{T}(d)\}$
по убыванию величины $\mathrm{N}(T,d)\cdot\mathrm{len}(T)$;
5: вычислить $W(d)$ согласно формуле (4);
6: $X^W:=X^W\cup\{W(d)\}$;
7: отсортировать X^W и D
по убыванию значения функции W;
8: сформировать
$H_1^W\bullet H_2^W\bullet\ldots\bullet H_{r(D)}^W$ **и** $H_1^D\bullet H_2^D\bullet\ldots\bullet H_{r(D)}^D$
с применением *Алгоритма 1*;
9: $D':=H_1^D$;
10: **для всех** $d\in D'$
11: $X^{\mathrm{N}}:=\varnothing$;
12: **для всех** $T\in\mathbf{T}(d)$

13: $X^{\mathrm{N}} := X^{\mathrm{N}} \cup \{\mathrm{N}(T,d)\}$;
14: отсортировать X^{N} и $\mathbf{T}(d)$
по убыванию значения функции N;
15: сформировать
$H_1^{\mathrm{N}} \bullet H_2^{\mathrm{N}} \bullet \ldots \bullet H_{r(\mathbf{T}(d))}^{\mathrm{N}}$ **и** $H_1^{T} \bullet H_2^{T} \bullet \ldots \bullet H_{r(\mathbf{T}(d))}^{T}$
с применением *Алгоритма 1*;
16: $\mathbf{T}'(d) := H_1^{T}$;
17: $X^{\mathrm{N}} := \varnothing$;
18: сформировать S;
19: **для всех** $s \in S$
20: вычислить $\mathrm{N}(s)$;
21: $X^{\mathrm{N}} := X^{\mathrm{N}} \cup \{\mathrm{N}(s)\}$;
22: отсортировать X^{N} и S
по убыванию значения функции N;
23: сформировать
$H_1^{\mathrm{N}} \bullet H_2^{\mathrm{N}} \bullet \ldots \bullet H_{r(S)}^{\mathrm{N}}$ **и** $H_1^{S} \bullet H_2^{S} \bullet \ldots \bullet H_{r(S)}^{S}$
с применением *Алгоритма 1*;
24: вернуть H_1^{S};
Конец *{Алгоритм 2}*.

Как видно из представленного псевдокода, введение в рассмотрение n-грамм на множестве связей слов даёт меньшее, чем в случае использования биграмм [2], число вызовов *Алгоритма 1*. Действительно, по сравнению с предложенным в [2] методом при вычислении функции (4) кластеризации n-грамм не производится, выполняется лишь их сортировка (*Шаг 4 Алгоритма 2*). Кроме того, число n-грамм в общем случае по определению меньше, чем связей, а их кластеризация производится не по всем документам множества D на входе *Алгоритма 2*, а лишь для максимально релевантных, составляющих множество D'.

Для снижения вычислительных затрат по выделению самих n-грамм целесообразно (если поиск ведётся на одном и том же текстовом множестве) запоминать значения найденных для них оценок. Кроме того, учитывая используемое в [9] предположение о возможности вхождения n-грамм друг в друга, представляет интерес их иерархизация по составу биграмм от «наиболее сильных» по оценке (1).

3. Экспериментальные исследования

Исходные текстовые множества для апробации предложенного метода подбирались таким образом, чтобы сравнить образы, выделяемые в тексте:

- 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.

Как и в [1, 2], основным критерием здесь была максимально полная и наглядная иллюстрация выявления в текстах контекстов использования слов-терминов и общей лексики. Сами же исходные текстовые множества полностью совпадали с задействованными в экспериментах из работы [2]. Для сравнения: число слов в документах первого множества варьировалось от 618 до 3765, число фраз – от 38 до 276. Аналогичные показатели для второго варианта исходного множества текстов – 218 и 6298 и соответственно 9 и 587.

В экспериментах по формированию единиц экспертных знаний участвовали две группы исходных фраз – каждая группа для своего варианта исходного текстового множества (табл. 2 и 3).

Табл. 2. Исходные фразы, предметная область «Философия и методология инженерии 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.*

Реализация предложенных методов на языке Java и результаты экспериментов представлены на портале НовГУ по адресу: http://www.novsu.ru/file/1258899 в ZIP-архиве (редакция от 21.01.2017). Архив включает каталог с исполняемым jar-файлом, подкаталогами текстовых корпусов и обученной модели классификатора для выделения границ предложений (подробнее – там же, файл *readme.doc*). В этом же каталоге находятся примеры аннотации (*issue.txt*), найденных морфологических характеристик (*issue_ch_lemmas.txt*) и связей для слов из исходных фраз (*issue_rels.txt*).

В качестве примера рассмотрим выделение составляющих образов для представленных в табл. 4 групп исходных фраз из табл. 2 и 3. Каждая группа составлена экспертом из эквивалентных либо дополняющих друг друга по смыслу фраз. Первая включает фразу №1 из табл. 3 (вместе с синонимической перифразой), для которой были получены вполне удовлетворительные результаты как на основе ме-

ры TF-IDF в [1], так и на основе синтаксических связей в рамках биграмм в [2]. Две другие группы включают фразы из табл. 2, причём удовлетворительными по данным экспериментам оказались лишь отдельные результаты.

Табл. 3. Исходные фразы, предметная область «Математические методы обучения по прецедентам»

№	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.*

Вместе с тем фразы внутри этих групп взаимно дополняют друг друга по смыслу, что немаловажно для предположения о соответствии им единого образа. Для сравнения по группам из табл. 4 в табл. 5 приведено общее число отобранных фраз (N), в том числе представляющих (с точки зрения эксперта) выразительные ЕЯ-средства (N_1), синонимы (N_2) и связи для понятий из упомянутых в исходных фразах (N_3).

Табл. 4. 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	*Определение модели представления знаний накладывает ограничения на выбор соответствующего механизма логического вывода. Однородность представления знаний приводит к упрощению механизма управления логическим выводом и упрощению управления знаниями.*
3	*Эвристика основывается на собственном опыте специалиста в данной предметной области, накопленном в результате многолетней практики. Метазнания могут касаться свойств, структуры, способов получения и использования знаний при решении практических задач искусственного интеллекта.*

В целях более полной оценки результативности поиска здесь также показано число представляемых найденными фразами выразительных средств языка (N_1^1), синонимов (N_2^1) и понятийных связей (N_3^1), где $\forall N_i^1$ есть сумма значений соответствующего показателя по отобранным фразам из относимых к i-му «подмножеству» значения N, $i = 1,\ldots,3$.

Табл. 5. Отбор релевантных фраз для групп из табл. 4 на основе n-грамм

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
без привлечения базы синтаксических правил, оценка (5)							
1	3	1	1	1	1	1	2
2	1	0	0	1	0	0	1
3	1	1	1	1	1	3	2
без привлечения базы синтаксических правил, оценка (6)							
1	1	0	1	1	0	1	2
2	3	1	0	1	1	0	1
3	2	1	2	2	1	5	5
с привлечением базы синтаксических правил, оценка (5)							
1	1	0	0	1	0	0	2
2	16	2	0	3	0	0	3
3	3	1	1	2	1	1	3
с привлечением базы синтаксических правил, оценка (6)							
1	1	0	0	1	0	0	2
2	4	0	0	2	0	0	1
3	2	0	0	1	0	0	1

Аналогичные данные приводятся далее и по экспериментам с отдельными фразами из табл. 2 и 3. Сравнение при этом также ведётся с результатами построения аннотации предложенным в [2] методом по числу связей слов, относимых *Алгоритмом 1* к «наиболее сильным» согласно оценке (1) при максимуме суммы её значений по всем связям слов исходной фразы (далее – по «наиболее сильным» связям, табл. 8 и 9).

Как видно из табл. 6–9 (строки для фраз из состава представленных в табл. 4 групп выделены более тёмным фоном), введение в рассмотрение совокупности исходных фраз, взаимно эквивалентных либо дополняющих друг друга по смыслу (с точки зрения носителя ЕЯ), совместно с n-граммами позволяет в ряде случаев более точно описывать выделяемый в текстах образ в виде сочетаний связанных по смыслу слов. Сказанное наглядно иллюстрируется сравнением соотношений N_i^1/N_i и N для всех $i = 1,\ldots,3$ в табл. 6–9 и табл. 5 по группам и отдельным фразам в их составе.

Хорошим примером может послужить результат по группе фраз №3 из табл. 4, где на основе n-грамм, выделяемых без привлечения синтаксических правил, применением оценки (5) была выбрана фраза:

«Эвристика может пониматься как:

– *научно-прикладная дисциплина, изучающая творческую деятельность;*
– *приемы решения проблемных (творческих, нестандартных, креативных) задач в условиях неопределенности, которые обычно противопоставляются формальным методам решения, опирающимся, например, на точные математические алгоритмы;*
– *метод обучения;*

– *один из способов создания компьютерных программ – эвристическое программирование».*

Табл. 6. Отбор релевантных для отдельных представленных в табл. 2 фраз на основе n-грамм

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
без привлечения базы синтаксических правил, оценка (5)							
1	2	0	0	2	0	0	1
2	4	0	0	2	0	0	2
3	4	1	0	3	1	0	3
4	2	1	1	0	1	1	0
5	3	0	1	2	0	1	2
6	3	0	0	1	0	0	1
7	3	0	0	1	0	0	2
8	2	1	0	2	1	0	2
9	3	0	0	3	0	0	5
без привлечения базы синтаксических правил, оценка (6)							
1	1	0	0	1	0	0	1
2	1	0	0	1	0	0	1
3	2	1	0	1	2	0	1
4	1	0	1	0	0	1	0
5	1	0	1	1	0	1	1
6	10	1	0	4	1	0	3
7	3	0	0	1	0	0	2
8	6	1	0	2	1	0	2
9	2	0	0	2	0	0	4
с привлечением базы синтаксических правил, оценка (5)							
1	5	0	0	3	0	0	3
2	2	0	0	1	0	0	1
3	19	0	3	4	0	2	4
4	7	1	0	4	1	0	4
5	3	0	2	1	0	2	1
6	3	1	0	2	1	0	2
7	3	0	0	2	0	0	2
8	1	1	0	0	1	0	0
9	1	0	0	1	0	0	1
с привлечением базы синтаксических правил, оценка (6)							
1	4	0	0	2	0	0	2
2	1	0	0	0	0	0	0
3	2	0	0	0	0	0	0
4	7	1	1	3	1	1	3
5	1	0	1	1	0	1	1
6	1	1	0	1	1	0	1
7	1	0	0	0	0	0	0
8	1	1	0	0	1	0	0
9	1	0	0	1	0	0	1

Данная фраза отобрана из тезисов доклада Е.И. Кузичкиной на 7-й Всероссийской конференции студентов, аспирантов и молодых учёных «Искусственный интеллект: философия, методология, инновации» (2013 г.) и была единственной вошедшей здесь в аннотацию, при этом из слов наиболее значимых *n*-грамм во фразе присутствуют *эвристика*, *в*, *задача*, *на*, *способ*, *решение*, *мочь*. Одновременное наличие этих слов в отбираемой фразе позволяет соотнести понятия *эвристика* и *знание*, упоминаемые в исходных фразах, с *приёмами решения задач*, а также реализовать вариант языковых выразительных средств *«в результате ⇔ как результат»* плюс синонимические замены *«способ ⇔ приём»*, *«опираться ⇔ основываться»* и *«практический ⇔ прикладной»*.

Табл. 7. Отбор релевантных для отдельных представленных в табл. 3 фраз на основе n-грамм

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
без привлечения базы синтаксических правил, оценка (5)							
1	1	1	0	0	1	0	0
2	1	1	1	0	1	1	0
3	2	2	2	1	1	1	2
4	1	1	1	0	1	1	0
5	2	0	1	1	0	1	1
6	1	1	1	1	1	1	1
7	9	1	0	5	1	0	5
8	2	0	0	1	0	0	1
9	6	1	0	4	1	0	4
без привлечения базы синтаксических правил, оценка (6)							
1	1	1	0	0	1	0	0
2	1	1	1	0	1	1	0
3	2	2	2	1	1	1	2
4	1	1	1	0	1	1	0
5	2	0	1	1	0	1	1
6	1	1	1	1	1	1	1
7	3	1	0	2	1	0	2
8	2	0	0	1	0	0	1
9	6	1	0	4	1	0	4
с привлечением базы синтаксических правил, оценка (5)							
1	1	1	0	0	1	0	0
2	1	1	1	0	1	1	0
3	2	2	2	1	1	1	2
4	1	1	1	0	1	1	0
5	2	0	2	2	0	1	2
6	1	0	1	0	0	1	0
7	9	1	0	3	1	0	3
8	4	0	0	0	0	0	0
9	1	0	0	0	0	0	0
с привлечением базы синтаксических правил, оценка (6)							
1	1	1	0	0	1	0	0
2	1	1	1	0	1	1	0
3	2	2	2	1	1	1	2
4	1	1	1	0	1	1	0
5	3	0	2	2	0	1	2
6	3	1	3	1	1	2	1
7	5	1	0	3	1	0	3
8	3	0	0	0	0	0	0
9	1	0	0	0	0	0	0

Отметим, что определение *эвристики*, альтернативное первой фразе группы №3 из табл. 4, было и среди фраз, наиболее релевантных исходной №6 в табл. 2 по «наиболее сильным» связям (по TF-IDF): *«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»*. Из «наиболее сильных» пар слов, которые служили основой отбора фраз, здесь содержится только *«искусственный интеллект»*, что заметно снизило точность выделения составляющих образа исходной фразы. Фактически найденная фра-

за лишь соотносит понятие *искусственный интеллект* из исходной фразы с понятием *эвристика*.

Преимущества поиска составляющих образа исходной фразы на основе *n*-грамм совместно с выделением связей слов по TF-IDF наиболее наглядно иллюстрируются экспериментами с фразой №3 из табл. 2, для которой контекстно-зависимым аннотированием по «наиболее сильным» связям слов в [2] удовлетворительного решения найдено не было.

Табл. 8. Отбор релевантных для представленных в табл. 2 фраз по «наиболее сильным» связям

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
без привлечения базы синтаксических правил							
1	1	0	0	1	0	0	1
2	2	0	0	0	0	0	0
3	11	1	2	5	1	2	7
4	1	1	0	1	1	0	1
5	2	2	2	0	2	2	0
6	6	1	1	1	1	1	1
7	6	0	0	2	0	0	3
8	6	0	1	3	0	1	3
9	1	0	0	1	0	0	1
с привлечением базы синтаксических правил							
1	2	0	0	1	0	0	1
2	4	1	0	2	1	0	2
3	1	0	0	0	0	0	0
4	3	1	2	1	1	2	1
5	4	0	0	2	0	0	5
6	1	1	1	0	1	1	0
7	6	0	0	2	0	0	3
8	1	0	0	0	0	0	0
9	5	0	0	2	0	0	2

Табл. 9. Отбор релевантных для представленных в табл. 3 фраз по «наиболее сильным» связям

№	N	N_1	N_2	N_3	N_1^1	N_2^1	N_3^1
без привлечения базы синтаксических правил							
1	2	0	0	1	0	0	1
2	1	1	1	0	1	1	0
3	15	3	2	7	2	2	7
4	15	2	2	4	1	1	4
5	5	0	1	0	0	1	0
6	1	0	0	0	0	0	0
7	6	0	1	0	0	1	0
8	1	0	0	1	0	0	1
9	1	1	1	0	1	1	0
с привлечением базы синтаксических правил							
1	1	1	0	0	1	0	0
2	1	1	1	0	1	1	0
3	15	3	2	7	2	2	7
4	15	2	2	4	1	1	4
5	5	0	1	0	0	1	0
6	11	0	9	4	0	1	4
7	1	0	0	0	0	0	0
8	1	0	0	1	0	0	1
9	1	1	1	0	1	1	0

Наилучшими здесь оказались результаты для *n*-грамм на связях слов по TF-IDF с отбором фраз в аннотацию на основе оценки (5), где в число четырёх результирующих вошла фраза: *«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»*. Соотнося представление о *знании* из исходной фразы с *моделью знания*, данная фраза посредством местоимённого наречия «*как*» также позволяет строить перифразы вида *«определяется как ⇔ понимается как»*. Рассматриваемая фраза была в числе результирующих и в эксперименте с отбором фраз на основе «наиболее сильных» связей слов исходной фразы по TF-IDF. При этом в дополнение к результатам аннотирования по *n*-граммам здесь также удалось выделить ряд связей понятий из исходной фразы (в первую очередь – для понятия *информация*) с другими понятиями той же предметной области, например: *«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»*.

Точность выделения составляющих образа исходной фразы на основе *n*-грамм можно наглядно оценить пословным сравнением наиболее значимых (т.е. «наиболее сильных») связей и *n*-грамм, отвечающих кластеру наибольших значений оценки (3) из формируемых *Алгоритмом 1*, относительно документов из числа максимально релевантных одновременно и по критерию (4), и по «наиболее сильным» связям. В эксперименте, результаты которого приведены в табл. 10, для исходной фразы №1 из табл. 2 указанные связи и *n*-граммы по составу слов полностью совпадают. В то же время для фраз №1, 7 и 9 из табл. 3 не нашлось документов, релевантных одновременно по двум вышеназванным критериям.

Представленный же в табл. 11 эксперимент дал полное совпадение состава слов рассматриваемых биграмм и *n*-грамм одновременно по фразе №4 из табл. 2 и фразе №1 табл. 3. Тем не менее, одновременно релевантных документов по критерию (4) и «наиболее сильным» связям здесь не нашлось для большего числа фраз: №3, 6, 7, 9 из табл. 2, а также фраз №7 и 8 табл. 3.

Как видно из примера, введение связей слов по TF-IDF как альтернативы синтаксическим отношениям в рассматриваемом ранжировании документов позволяет в большей степени учитывать термины, что немаловажно для предметных областей, где их доля сравнима с долей общей лексики в текстах.

Помимо сравнения наиболее значимых связей и *n*-грамм, эффективность предложенного в настоящей работе метода контекстно-зависимого аннотирования, как и в [2], оценивается сопоставлением с результатами поиска фраз, близких исходной, на готовом синтаксически размеченном текстовом корпусе на основе слов и их сочетаний (табл. 13), предварительно выделенных экспертом и представляющих термины предметной области.

Табл. 10. Сравнение наиболее значимых связей и n-грамм для отбора фраз (без привлечения синтаксических правил, оценка (5))

№ 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*

Табл. 11. Сравнение наиболее значимых связей и n-грамм для отбора фраз (с привлечением синтаксических правил, оценка (5))

№ 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*	

Как видно из табл. 6, 7 и 12, наряду с автоматизацией поиска указанных слов и сочетаний, предложенный метод позволяет здесь в среднем в 17 раз сократить выход фраз, не релевантных исходной ни по описываемому фрагменту знания, ни по языковым формам его выражения, что на 11,77 % выше аналогичного показателя в работе [2]. В настоящей работе этот показатель рассчитывался как усреднённое значение соотношения величины N из табл. 12 и среднего значения указанной величины, соответственно, в табл. 6 и 7 (для разных вариантов выделения связей слов и оценок (5) либо (6)) по отдельным исходным фразам из табл. 2 и 3. Общее число отобранных фраз, т.е. N, выбрано нами для сравнения в предположении худшего случая, когда среди отбираемых фраз нет релевантных исходной.

Табл. 12. Отбор релевантных фраз из текстов Национального корпуса русского языка [11]

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
для исходных фраз из табл. 2									
N	13	73	2	15	83	33	79	224	20
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	2	5	0	1	5	3	3	2	2
N_3^1	2	6	0	2	4	3	3	2	2
для исходных фраз из табл. 3									
N	56	1	1	1	24	17	21	5	2
N_1	0	0	0	0	0	0	0	0	0
N_2	0	0	0	0	0	0	0	0	0
N_3	0	0	0	0	0	0	0	1	0
N_3^1	0	0	0	0	0	0	0	1	0

Табл. 13. Слова и их сочетания для отбора релевантных фраз из Национального корпуса русского языка

№	**Слова и сочетания слов**
по исходным фразам из табл. 2	
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*
по исходным фразам из табл. 3	
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*

4. Некоторые технические детали и допущения

Классическая постановка задачи кластерного анализа [8] предполагает, что каждый элемент последовательности, разбиваемой на кластеры с применением *Алгоритма 1*, представлен в ней ровно один раз. Как и в [1, 2], с целью наглядности изложение предлагаемого в настоящей работе метода неявно содержит предположение о выполнении данного условия.

Извлечение текста из PDF-файла, морфологический анализ словоформ, а также выделение синтаксических связей выполнялись теми же методами, что и в [2]. Для выделения границ предложений в тексте по знакам препинания использовалась обученная модель классификатора, построенного с применением интегрированного пакета Apache OpenNLP [12]. Обучение классификатора распознаванию границ предложений осуществлялось на основе размеченных данных в виде газетных текстов на русском языке (2010 г., всего 1 млн. фраз) из Leipzig Corpora [13]. Здесь, как и в работе [2], при выборе исходных данных для обучения авторы ограничились максимальным по объёму русскоязычным текстовым корпусом.

Заключение

Основной *результат* данной работы – *совершенствование представленного в [2] метода формирования корпуса тематических текстов, релевантных фрагменту знаний и формам его выражения в языке*.

Отметим, что предложенный в настоящей работе вариант контекстно-зависимого аннотирования ориентирован в первую очередь на поиск форм выражения связей понятий в текстах той предметной области, где доля общей лексики сравнима с долей терминов. Поэтому для максимально эффективного решения основной задачи его следует использовать именно как дополнение результатов работ [1, 2].

Тема отдельного рассмотрения – скорость и точность морфологического анализа, необходимого для выделения связей слов. Здесь представляет интерес реализация предложенного в работе метода на языке Python с привлечением библиотеки NLTK [14] и морфологического анализатора Pymorphy [15] как альтернатива реализованному авторами решению на базе библиотеки русской морфологии [16].

Благодарности

Работа выполнена при поддержке Министерства образования и науки РФ (базовая часть госзадания), а также гранта РФФИ (№16-01-00004).

Литература

1. **Михайлов, Д.В.** Выделение знаний и языковых форм их выражения на множестве тематических текстов: подход на основе меры TF-IDF / Д.В. Михайлов, А.П. Козлов, Г.М. Емельянов // Компьютерная оптика. – 2015. – Т. 39, № 3. – С. 429-438. – DOI: 10.18287/0134-2452-2015-39-3-429-438.
2. **Михайлов, Д.В.** Выделение знаний, языковых форм их выражения и оценка эффективности формирования множества тематических текстов / Д.В. Михайлов, А.П. Козлов, Г.М. Емельянов // Компьютерная оптика. – 2016. – Т. 40, № 4. – С. 572-582. – DOI: 10.18287/2412-6179-2016-40-4-572-582.
3. **Шеннон, К.** Работы по теории информации и кибернетики / К. Шеннон; пер. с англ. – М.: Иностранная литература, 1963. – С. 669-686.
4. **Sidorov, G.** Syntactic dependency based N-grams in rule based automatic English as second language grammar correction / G. Sidorov // International Journal of Computational Linguistics and Applications. – 2013. – Vol. 4(2). – P. 169-188.
5. **Кудинов, М.С.** Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей / М.С. Кудинов // Машинное обучение и анализ данных. – 2013. – Т. 1, № 6. – С. 714-724. – ISSN 2223-3792.
6. **Москович, В.А.** Дистрибутивно-статистический метод построения тезаурусов: современное состояние и перспективы / В.А. Москович. – М., 1971. – 66 с.
7. **Tanimoto, T.T.** An elementary mathematical theory of classification and prediction / T.T. Tanimoto. – New York: International Business Machines Corporation, 1958. – 10 p.
8. **Загоруйко, Н.Г.** Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Издательство института математики, 1999. – 270 с.
9. **Frantzi, K.** Automatic recognition of multi-word terms: the C-value/NC-value method / K. Frantzi, S. Ananiadou, H. Mima // International Journal on Digital Libraries. – 2000. – Vol. 3, Issue 2. – P. 115-130. – DOI: 10.1007/s007999900023.
10. **Бродский, А.** Алгоритмы контекстно-зависимого аннотирования Яндекса на РОМИП-2008 / А. Бродский, Р. Ковалев, М. Лебедев, Д. Лещинер, П. Сушин, И. Мучник // Труды РОМИП 2007-2008. – 2008. – С. 160-169.
11. Национальный корпус русского языка [Электронный ресурс]. – URL: http://www.ruscorpora.ru/ (дата обращения 09.03.2017).
12. Apache OpenNLP [Электронный ресурс]. – URL: https://opennlp.apache.org/ (дата обращения 10.03.2017).
13. Leipzig Corpora Collection Download Page [Электронный ресурс]. – URL: http://wortschatz.unileipzig.de/en/download (дата обращения 10.03.2017).
14. Natural Language Toolkit [Электронный ресурс]. – URL: http://www.nltk.org (дата обращения 17.03.2017).
15. Pymorphy – NLPub [Электронный ресурс]. – URL: https://nlpub.ru/Pymorphy (дата обращения 17.03.2017).
16. Russianmorphology: Russian Morphology for lucene [Электронный ресурс]. – URL: http://code.google.com/p/russian-morphology/ (дата обращения 19.03.2017).

Сведения об авторах

Михайлов Дмитрий Владимирович, 1974 года рождения, в 1997 году окончил Новгородский государственный университет имени Ярослава Мудрого (НовГУ) по специальности 2204 «Программное обеспечение вычислительной техники и автоматизированных систем». В 2003 году защитил диссертацию на соискание ученой степени кандидата, а в 2013 году – доктора физико-математических наук. В настоящее время работает доцентом кафедры информационных технологий и систем (ИТиС) в федеральном государственном бюджетном образовательном учреждении высшего образования «Новгородский государственный университет имени Ярослава Мудрого». Опубликовал более

80 научных работ (из них более 20 статей в рецензируемых журналах из списка ВАК). Область научных интересов: интеллектуальный анализ данных, компьютерная лингвистика. E-mail: Dmitry.Mikhaylov@novsu.ru .

Козлов Александр Павлович, 1989 года рождения, в 2011 году окончил НовГУ по специальности «Программное обеспечение вычислительной техники и автоматизированных систем», аспирант кафедры ИТиС НовГУ. Область научных интересов: интеллектуальный анализ данных, компьютерная лингвистика. E-mail: caleo@yandex.ru .

Емельянов Геннадий Мартинович, 1943 года рождения, в 1966 году окончил Ленинградский электротехнический институт им. В.И. Ульянова (Ленина) по специальности «Математические и счётно-решающие приборы и устройства». В 1971 году защитил диссертацию на соискание ученой степени кандидата технических наук. Доктор технических наук (1990). В настоящее время – профессор кафедры ИТиС НовГУ. Его научные интересы включают построение проблемно-ориентированных вычислительных систем обработки и анализа изображений. Автор более 150 научных работ. E-mail: Gennady.Emelyanov@novsu.ru .

ГРНТИ: 28.23.11, 28.23.15, 20.23.19
Поступила в редакцию 10 апреля 2017 г. Окончательный вариант – 1 июня 2017 г.

AN APPROACH BASED ON ANALYSIS OF N-GRAMS ON LINKS OF WORDS TO EXTRACT THE KNOWLEDGE AND RELEVANT LINGUISTIC MEANS ON SUBJECT-ORIENTED TEXT SETS

D.V. Mikhaylov [1], A.P. Kozlov [1], G.M. Emelyanov [1]
[1] Yaroslav-the-Wise Novgorod State University, Velikii Novgorod, Russia

Abstract

In this paper we look at two interrelated problems of extracting knowledge units from a set of subject-oriented texts (the so-called corpus) and completeness of reflection of revealed actual knowledge in initial phrases. The main practical goal here is finding the most rational variant to express the knowledge fragment in a given natural language for further reflection in the thesaurus and ontology of a subject area. The problems are of importance when constructing systems for processing, analysis, estimation and understanding of information. In this paper the text relevance to the initial phrase in terms of the described fragment of actual knowledge (including forms of its expression in a given natural language) is measured by estimating the coupling strength of words from the initial phrase jointly occurring in phrases of the analyzed text together with classifying these words according to their values of TF-IDF metrics in relation to text corpus. The paper considers an extension of links of words from traditional bigrams to three and more elements for the revelation of constituents of an image of the initial phrase in the form of combinations of related words. Variants of link revelation with and without application of a database of known syntactic relations are considered here. To describe more completely the fragment of expert knowledge revealed in corpus texts, sets of the initial phrases mutually equivalent or complementary in sense and related to the same image are entered into consideration. In comparison with the search of components of the analyzed image on a syntactically marked text corpus the method for text selection offered in the current paper can reduce, on average, by 17 times the output of phrases which are irrelevant to the initial ones in terms of either the knowledge fragment described or its expression forms in a given natural language.

Keywords: pattern recognition, intelligent data analysis, information theory, open-form test assignment, natural-language expression of expert knowledge, contextual annotation, document ranking in information retrieval.

Citation: Mikhaylov DV, Kozlov AP, Emelyanov GM. An approach based on analysis of *n*-grams on links of words to extract the knowledge and relevant linguistic means on subject-oriented text sets. Computer Optics 2017; 41(3): 461-471 DOI: 10.18287/2412-6179-2017-41-3-461-471.

Acknowledgements: The work was partially funded by the Russian Federation Ministry of Education and Science (the basic part of the state task) and the Russian Foundation of Basic Research, grant No. 16-01-00004.

References

[1] Mikhaylov DV, Kozlov AP, Emelyanov GM. An approach based on TF-IDF metrics to extract the knowledge and their linguistic forms of expression on the subject-oriented text set [In Russian]. Computer Optics 2015; 39(3): 429-438. DOI: 10.18287/0134-2452-2015-39-3-429-438.

[2] Mikhaylov DV, Kozlov AP, Emelyanov GM. Extraction of knowledge and relevant linguistic means with efficiency estimation for the formation of subject-oriented text sets [In Russian]. Computer Optics 2016; 40(4): 572-582. DOI: 10.18287/2412-6179-2016-40-4-572-582.

[3] Shannon CE. Prediction and entropy of printed English. Bell System Technical Journal 1951; 30(1): 50-64.

[4] Sidorov G. Syntactic dependency based N-grams in rule based automatic English as second language grammar correction. IJCLA 2013; 4(2): 169-188.

[5] Kudinov MS. Shallow parsing of Russian text with conditional random fields [In Russian]. Machine Learning and Data Analysis 2013; 1(6): 714-724.
[6] Moskovich WA. Distributive-Statistical Method of Thesaurus Construction: The State of the Art and Perspectives [In Russian]. Moscow: The Scientific Council «Cybernetics» of the USSR Academy of Science; 1971.
[7] Tanimoto TT. An elementary mathematical theory of classification and prediction. New York: International Business Machines Corporation; 1958.
[8] Zagoruiko NG. Applied methods of data and knowledge analysis [In Russian]. Novosibirsk: Institute of Mathematics SD RAS; 1999.
[9] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. Int J Digit Libr 2000; 3(2): 115-130. DOI: 10.1007/s007999900023
[10] Brodskiy A, Kovalev R, Lebedev M, Leshchiner D, Sushin P. Yandex algorithms of contextual annotation at ROMIP 2008 [In Russian]. Russian Information Retrieval Evaluation Seminar (ROMIP) 2008; 160-169.
[11] Russian National Corpus [In Russian]. Source: ⟨http://www.ruscorpora.ru/⟩.
[12] Apache OpenNLP. Source: ⟨https://opennlp.apache.org/⟩.
[13] Leipzig Corpora Collection Download Page. Source: ⟨http://wortschatz.uni-leipzig.de/en/download⟩.
[14] Natural Language Toolkit. Source: ⟨http://www.nltk.org ⟩.
[15] Pymorphy – NLPub. Source: ⟨https://nlpub.ru/Pymorphy⟩.
[16] Russianmorphology: Russian Morphology for lucene. Source: ⟨http://code.google.com/p/russianmorphology/⟩.

Authors' information

Dmitry Vladimirovich Mikhaylov (b. 1974) graduated from Yaroslav-the-Wise Novgorod State University in 1997, majoring in Software of Computers and Automated Systems. Obtained his PhD (Kandidat nauk) and his Doctoral (Doktor nauk) degrees in Physics and Mathematics in 2003 and 2013, respectively. Currently he works as the Docent of the Department of Information Technologies and Systems at the same university. Author of more than 80 scientific papers. Research interests are intelligent data analysis and computational linguistics. E-mail: *Dmitry.Mikhaylov@novsu.ru* .

Alexander Pavlovich Kozlov (b.1989) graduated from Yaroslav-the-Wise Novgorod State University in 2011, majoring in Software of Computers and Automated Systems. Now he is post-graduate student of the same university. Research interests are intelligent data analysis and computational linguistics. E-mail: *caleo@yandex.ru* .

Gennady Martinovich Emel'yanov (b. 1943) graduated from the Leningrad Institute of Electrical Engineering in 1966, majoring in Mathematical and Calculating Instruments and Devices. Obtained his PhD (Kandidat Nauk) and his Doctoral (Doktor Nauk) degrees in Technical Sciences in 1971 and 1990, respectively. Now he is a Professor of the Department of Information Technologies and Systems at the Yaroslav-the-Wise Novgorod State University. Scientific interests include the construction of problem-oriented computing systems of image processing and analysis. He is the author of more than 150 publications. E-mail: *Gennady.Emelyanov@novsu.ru* .

Received April 10, 2017. The final version – June 1, 2017.