УДК 004.622

МЕТОДИКА АВТОМАТИЗИРОВАННОГО ФОРМИРОВАНИЯ ТЕЗАУРУСА НА ОСНОВЕ ГРАММАТИЧЕСКИ РЕЛЕВАНТНЫХ ЕДИНИЦ

***Александр Александрович Мелихов**, аспирант, ассистент кафедры,*
e-mail: megadelphin@mail.ru,
***Ольга Сергеевна Смирнова**, младший научный сотрудник, ассистент кафедры,*
e-mail: mail.olga.smirnova@yandex.ru,
Московский технологический университет (МИРЭА),
https://www.mirea.ru

Рассматривается задача формирования первичного тезауруса на основе автоматизированного анализа публикаций в рамках предметной области. Представлена методика, позволяющая сократить объем первичного тезауруса за счет исключения иррелевантных элементов, основанная на применении грамматики естественного языка как эвристики для алгоритма фильтрации.

Ключевые слова: информационная система; поддержка инновационной деятельности; тезаурус; обработка естественного языка; инженерия знаний; бионические информационные ресурсы.

Исследование выполнено федеральным государственным бюджетным образовательным учреждением высшего образования «Московский технологический университет» (МИРЭА) за счет гранта Российского научного фонда (проект № 14-11-00854).

DOI: 10.21777/2312-5500-2016-4-41-51

Введение

Современное состояние научно-технического прогресса принято называть информационной эпохой. Это обусловлено тем, что информация является важным ресурсом, с помощью которого можно существенно оптимизировать затраты ресурсов иного рода. Основным источником такой полезной информации является научно-исследовательская деятельность, в сфере бионики, направленная на поиск в живой природе решений различных технических задач. Так, накопление опыта, его систематизация и формализация позволяют затем изменять характер практической деятельности в сторону большей оптимальности по тем или иным параметрам. В этом смысле бионика как научное направление обладает интересной особенностью, связанной с тем, что организация научных исследований, как правило, проходит по единой схеме: на основе практических наблюдений строится модель, описательные и предсказательные возможности которой соотносятся с имеющейся картиной мира. В случае если новая модель более точна и удобна в использовании, то она берется на вооружение. Подобный подход логичен, однако современные гипотезы бывают настолько сложны, что проверить их на практике с различными начальными условиями не всегда представляется возможным. Целью же является поиск алгоритмов решения оптимизационных задач, основанных на имитации поведения или механизмов, используемых в природе

А.А. Мелихов

О.С. Смирнова

[1–3], т. е. поиск уже существующих в природе решений актуальных задач. Следует отметить, что бионика является междисциплинарным направлением, в котором задействованы представители различных специальностей: биологи, нейрофизиологи, физики, химики, инженеры и др., при этом достижение указанной цели требует организации информационного обмена, однако даже с учетом современного уровня развития информационных технологий решение подобной задачи является затруднительным в связи с тем, что передача знаний неизбежно требует их корректной формализации. По мере вовлечения все большего числа исследователей в данное научное направление появляется потребность в формировании единого информационного пространства, документы в котором организованы с учетом их семантического содержания [4].

Информационные системы подобного рода характеризуются наличием многоуровневой модели описания содержащихся в них информационных ресурсов и наличием специальных алгоритмов работы с такой моделью. Так, вид документа (текст, изображение и т. д.) определяет набор инструментов для его обработки, структура документа задает некоторую эвристику, необходимую для интерпретации его содержательной части (например, разделы книги), а метаописание определяет смысловую связь конкретного документа с дискурсом в целом.

Основной сложностью, возникающей при реализации подобных систем, является создание формализованной модели научного дискурса, относящегося к рассматриваемой предметной области. Сложность создания такого описания заключается в необходимости проведения комплексного анализа текущего состояния предметной области, требующего совместной работы представителей данного направления, т. е. обобщения и систематизации всего накопленного опыта. Таким образом, возникает некоторый замкнутый круг: для организации совместной работы множества специалистов требуется модель изучаемой ими научной дисциплины, создание которой может быть успешно проведено только при условии координации действий между научными коллективами. В рамках решения данного вопроса предлагается создание информационной системы поддержки инновационной деятельности, ориентированной на развитие бионических технологий, описанной в [5, 6].

Актуальность подобных разработок связана, в первую очередь, с высокой степенью интеграции различных научных коллективов в общемировое научное сообщество. Так, благодаря современным средствам связи возможен обмен данными между отдельными научными сотрудниками и целыми исследовательскими институтами, расположенными на разных материках. Подобное обстоятельство неизбежно влечет за собой необходимость в выборе некоторого универсального языка научного общения, однако практический опыт показывает, что мультикультурализм научного сообщества неизбежно влечет за собой локализацию коллективов по различным признакам, таким как приверженность определенной научной школе и язык общения.

Так, вследствие определенных геополитических причин, в зависимости от выбора языка коммуникации можно выделить ряд крупных групп. Наиболее широкую из них представляет собой англоговорящее сообщество, однако также крупными являются группы, для которых языками общения являются испанский (Латинская Америка и Испания), французский, русский (Россия и страны СНГ) и немецкий. При этом стоит отметить, что даже внутри коллективов могут возникать различные трудности информационного обмена, так как не для всех участников дискурса выбранный язык является родным, что накладывает существенные ограничения на формализацию и интерпретацию знаний.

В связи с этим в качестве решения описанной проблемы предполагается разработка систем, позволяющих организовывать информационный обмен инвариантно языку. В первую очередь необходимо создавать условия для поиска и адаптации иноязычных источников, поэтому в качестве постановки задачи предполагается рассмотреть возможность представления русскоязычному пользователю публикаций, выполненных

на английском языке, являющемся в настоящий момент самым распространенным в научной среде. Целью является разработка методики автоматизированного формирования первичного информационно-поискового тезауруса (АФПИПТ), пригодного для его дальнейшего применения в многоязычной интеллектуальной системе поддержки научных исследований.

Основные положения методики АФПИПТ

При формировании базы знаний первичный список терминов может быть сформирован на основе информационно-поискового тезауруса (ИПТ), основным преимуществом которого является возможность его извлечения из текста в полностью автоматическом режиме. Как правило, ИПТ крайне неоднороден и избыточен ввиду того, что множество запросов от поисковых систем является открытым, он может содержать в себе как семантически релевантные дискурсу единицы, так и слова и выражения, непосредственно не отражающие суть текста. Таким образом, для ИПТ требуется некоторое преобразование, целью которого является снижение его размерности и увеличение релевантности входящих в него единиц. Данное преобразование может осуществляться как на этапе формирования первичных данных, так и для готового ИПТ, однако в любом случае фильтрация также требует введения дополнительных эвристик (перечень стоп-слов, общеязыковые и специальные тезаурусы).

В общем виде алгоритм формирования первичного тезауруса можно представить следующим образом [7]:

– формирование корпуса текстов;
– преобразование текстов к единому формату;
– анализ документов с формированием ИПТ;
– фильтрация ИПТ.

Процесс формирования корпуса текстов подробно рассмотрен в [4], а вопросы, связанные с унификацией формата документов, – в [7], поэтому рассмотрим подробно задачи формирования и фильтрации ИПТ.

Формирование исходных данных для информационно-поискового тезауруса

В ранних поисковых системах индексируемые последовательности слов извлекались из текста в порядке следования, что порождало необходимость применения алгоритмов, позволяющих получать результаты по приблизительному соответствию пользовательскому запросу. Однако, несмотря на простоту реализации, данные алгоритмы обладают рядом очевидных недостатков: перебор индексов с учетом редакционных расстояний характеризуется существенными накладными затратами на ресурсы памяти и машинного времени (что до начала широкого применения параллельных вычислений представляло серьезную проблему), эффективность индексирования напрямую зависит от объема текста, при этом подобные системы критичны к скорости добавления новых данных. В настоящий момент приоритетной тенденцией в сфере разработки информационных систем является дальнейшее развитие аппаратного обеспечения и адаптация существующих алгоритмов с учетом его особенностей, однако вместо увеличения вычислительных возможностей по мере роста объемов обрабатываемой информации возможно изменение способа формирования исходных данных.

Так, в [8] представлено решение данной задачи, основанное на применении грамматики естественного языка в качестве эвристики, позволяющей определить и исключить из набора заведомо нерелевантные данные. Суть заключается в следующем: применение естественного языка (ЕЯ) для обеспечения коммуникаций предполагает формализацию некоторых мысленных образов в линейную последовательность языковых символов, пригодную для передачи и приема через визуальный (текст) и акустический (речь) каналы. В ЕЯ системы формализации письменной и устной речи могут совпадать или незначительно различаться либо не иметь аналогов вовсе. В первом случае порядок слов может не соответствовать их семантической связанности, т. е. стоящие рядом сло-

ва могут относиться к разным синтаксическим группам. Например, в допустимом с точки зрения правил грамматики предложении «Уеду я далеко» слово «далеко» относится не к подлежащему, а к сказуемому. Для разрешения неопределенностей, связанных с несовпадение порядка слов и их синтаксических ролей, предлагается применение синтаксических анализаторов. Рассмотрим принцип работы такого анализатора на примере Stanford Parser, который реализует два вида представлений синтаксических отношений: универсальные зависимости и синтаксическое дерево, представляющие разные подходы к формализации грамматики ЕЯ.

В основе универсальных зависимостей (англ. Universal Dependencies, UD) лежит гипотеза о том, что элементы предложения любого ЕЯ связаны попарно, причем вне зависимости от конкретного языка множество таких связей ограничено [9]. Возьмем для примера сравнительно простое предложение «*Business biomimetics is the latest development in the application of biomimetics*» и определим для него множество бинарных отношений (рис. 1).

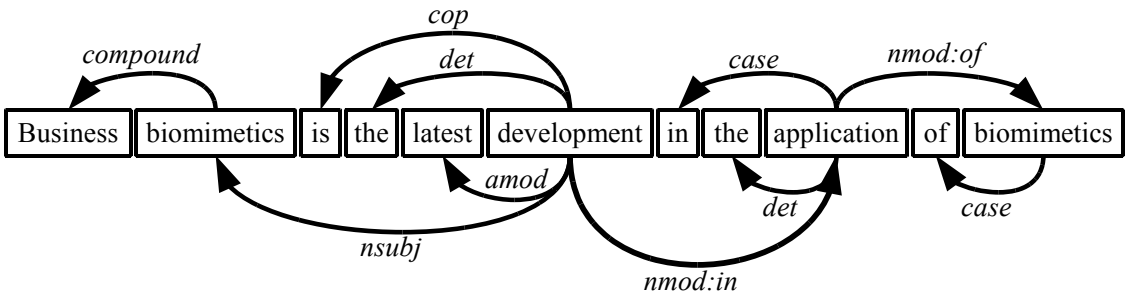

Рис. 1. Бинарные отношения

В более компактном виде данные отношения можно записать следующим образом:
1) compound (biomimetics-2, Business-1);
2) nsubj (development-6, biomimetics-2);
3) cop (development-6, is-3);
4) det (development-6, the-4);
5) amod (development-6, latest-5);
6) root (ROOT-0, development-6);
7) case (application-9, in-7);
8) det (application-9, the-8);
9) nmod:in (development-6, application-9);
10) case (biomimetics-11, of-10);
11) nmod:of (application-9, biomimetics-11).

Во втором случае структура данных, описывающая грамматические отношения, представляет собой дерево, терминальными элементами которого являются отдельные слова, а ветви, в свою очередь, образованы синтаксическими группами, допускающими вложенность (рис. 2).

Дерево синтаксического разбора обладает большей наглядностью и позволяет получать n-граммы различных порядков. Основным же его недостатком является четкая привязка к принятым языковым нормам, поэтому даже в пределах одной языковой грамматики возможны некоторые вариации.

Обход графа позволяет определить количество и состав синтаксических групп, формирующих предложения. Каждая такая синтаксическая группа содержит один или более элементов предложения, при этом элементы внутри группы согласованы между собой по различным грамматическим признакам, таким как род, число, падеж (явный или неявный), время и т. п. Например, из указанного предложения можно извлечь следующие *n*-граммы, представленные в табл. 1.

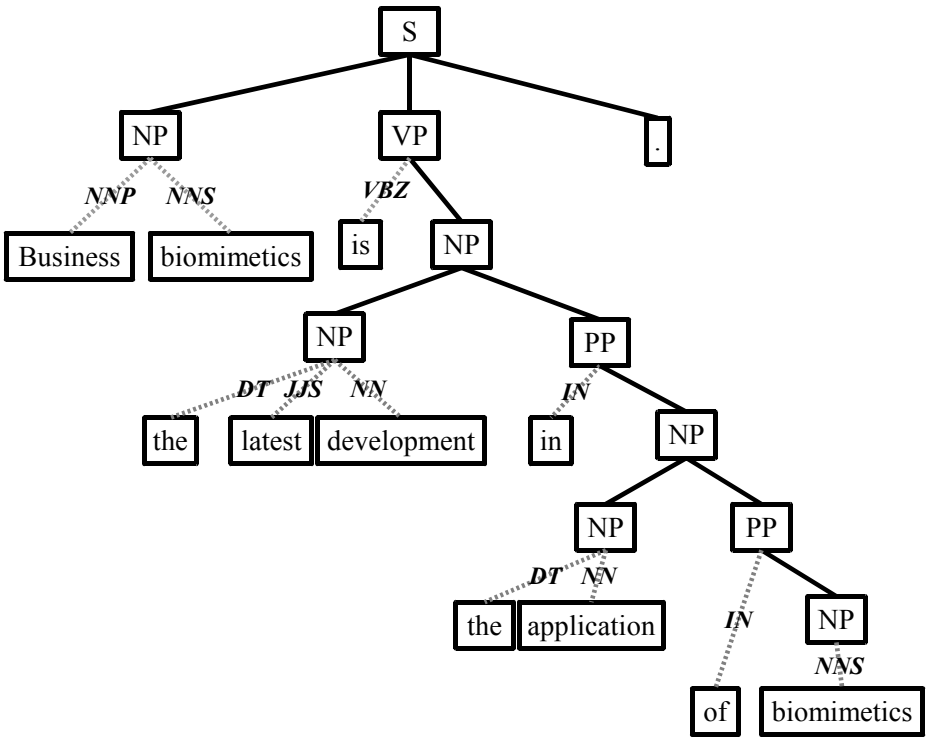

Рис. 2. Дерево синтаксического разбора предложения

Таблица 1

Полный список n-грамм

N-грамма	Порядок
of biomimetics	2
the application	2
business biomimetics	2
the latest development	2
the application of biomimetics	4
in the application of biomimetics	5
the latest development in the application of biomimetics	8
is the latest development in the application of biomimetics	9
business biomimetics is the latest development in the application of biomimetics	12

Сравним между собой наборы извлеченных разными способами биграмм (табл. 2).

Таблица 2

Перечень биграмм

Позиция	Универсальные зависимости	Дерево разбора
Business biomimetics	biomimetics, Business	of biomimetics
biomimeticsis	development, biomimetics	the application
isthe	development, is	business biomimetics
thelatest	development, the	
latestdevelopment	development, latest	
developmentin	application, in	
inthe	application, the	
theapplication	development, application	
applicationof	biomimetics, of	
ofbiomimetics	application, biomimetics	

Левая колонка отображает подход, характерный для ИПТ ранних поисковых систем. Средняя колонка отображает биграммы, полученные в результате определения взаимных зависимостей слов. Правая колонка иллюстрирует набор, полученный на ос-

нове дерева разбора.

В общем виде число элементов в первой колонке можно рассчитать как L–1, где L – число элементов в предложении (его длина). Аналогичные вычисления производятся и для второй колонки. Таким образом, *применение универсальных зависимостей не сокращает общее число биграмм, однако позволяет сформировать их набор с учетом синтаксических отношений.*

Рассмотрим n-граммы более высокого порядка (табл. 3, 4).

Таблица 3

Перечень триграмм

Позиция	**Дерево разбора**
Business_biomimetics_is	the_latest_development
biomimetics_is_the	
is_the_latest	
the_latest_development	
latest_development_in	
development_in_the	
in_the_application	
the_application_of	
application_of_biomimetics	

Таблица 4

Перечень квадрограмм

Позиция	**Дерево разбора**
Business_biomimetics_is_the	the_application_of_biomimetics
biomimetics_is_the_latest	
is_the_latest_development	
the_latest_development_in	
latest_development_in_the	
development_***in_the_application***	
in_the_application_of	
the_application_of_biomimetics	

Таким образом, общее число всех возможных позиционных n-грамм определяется по формуле

$N = L - (n - 1) = L - n + 1$, где

L – длина предложения;

n – длина n-граммы.

Полученные сочетания могут быть представлены в виде таблиц абсолютной и/или относительной встречаемости, которые затем могут быть проанализированы специалистом в предметной области.

Оценка релевантности полученных данных

Для оценки релевантности n-грамм целесообразно применение метрики TF-IDF, важной особенностью которой является возможность расчета весовых коэффициентов для извлекаемых n-грамм независимо для каждого документа [10].

Так, для каждого сочетания его «вес» в документе (*tf* – term frequency) определяется исходя из частоты его повторения по формуле

$tf(t,d) = \frac{n_i}{\sum_k n_k}$, где (1)

n_i – число вхождений;

$\sum_k n_k$ – общее число единиц.

Далее определяется обратная частота документа (*idf* – inverse document frequency):

$idf(t,D) = \log \frac{|D|}{|(d_i \supset t_i)|}$, где

|D| – число документов в корпусе;
$|(d_i \supset t_i)|$ – количество документов, в которых встречается t_i (при $n_i \neq 0$).
Общий «вес» определяется как произведение *tf* и *idf*:
$tf\text{–}idf\,(t, d, D) = tf\,(t, d) \times idf\,(t, d)$.

Интерпретация полученных значений и фильтрация

Рассмотренный пример показывает, что описанная методика позволяет уменьшить число учитываемых n-грамм, но никак не влияет на ИПТ, сформированный из единичных слов. Однако использование синтаксических анализаторов позволяет также определять контекст употребления отдельных слов и словосочетаний, что в конечном итоге позволяет представлять эксперту по знаниям максимально релевантный ИПТ.

Определение части речи (P-o-S tagging) позволяет снизить ранг для элементов, обладающих омонимией (например, «пила – существительное» и «пила – гл. прош. вр., 3 л., ед. ч.»), позволяя заметно сократить число омонимов в конечном наборе данных. Ведение статистики сочетаемости слов может способствовать поиску устоявшихся сложных по составу терминов.

Описание эксперимента и интерпретация результатов

В качестве исходного корпуса выступили публикации и препринты, представленные в табл. 5.

Таблица 5

Перечень публикаций, входящих в корпус

№	Авторы	Название статьи	Вид	Год издания
1	Zhao M.	A Clustering Based Approach for Realistic and Efficient Data-Driven Crowd Simulation	Статья*	2015
2	Amos M., Hodgson D. A., Gibbons A.	Bacterial self-organisation and computation	Статья*	2007
3	Versos C., Coelho D.	A Bi-Directional Method for Bionic Design with Examples	Статья*	2013
4	Bar-Cohen Y.	Biomimetics–using nature to inspire human innovation	Статья*	2006
5	Neurohr R., Dragomirescu C.	Bionics in Engineering – Defining new Goals in Engineering Education at «Politehnica» University of Bucharest	Статья*	2008
6	Walker I. D., Carreras C., McDonnell R., Grimes G.	Extension versus Bending for Continuum Robots	Статья*	2006
7	Emmelmanna C., Sander P., Kranz J., Wycisk E.	Laser Additive Manufacturing and Bionics: Redefining Lightweight Design	Конф.**	2011
8	Pershin Y. V., Di Ventra M.	Memcomputing and swarm intelligence	Статья*	2014
9	Felbrich B., Nönnig J. R., Wiesenhütter S.	Rigid Folding in Robotic Multi-agent Systems	Конф.**	2014

Примечания:
Статья – статья в периодическом журнале;*
*Конф.** – материалы конференции.*

Для каждого текстового документа составлены таблицы вхождения n-грамм порядков от 2 до 4. Для каждой n-граммы определена абсолютная встречаемость и вычислена по формуле (1) относительная встречаемость. В качестве контроль-

ного примера выступают n-граммы, полученные путем токенизации предложения, в качестве контрольного примера – n-граммы, полученные в результате грамматического разбора, при этом отдельно учитывались сочетания по двум разным признакам: грамматической роли всего предложения и частям речи отдельных слов. Данный способ группировки вызван необходимостью разрешения ситуаций омонимии.

После анализа полученных данных был составлен следующий ИПТ. В табл. 6–8 представлены соответственно би-, три-, квадрограммы, полученные двумя различными способами. Левый столбец – контрольный пример, средний – с учетом грамматической роли, правый – с учетом части речи входящих в сочетание элементов.

Таблица 6

ИПТ для биграмм

Позиция	**Дерево разбора (по грамматической роли)**	**Дерево разбора (с учетом частей речи)**
biological_solution	NP pairwise_interactions	NN_interaction NNS_patterns
pairwise_ interactions	NP artificial_intelligence	JJ_pairwise NNS_interactions
memristive_network	NP biological_solutions	NN_pattern NN_formation
memristive_systems	NP interaction_patterns	JJ_state-action NNS_samples
bionic_solution	NP pattern_formation	JJ_biological NNS_solutions
biological_creatures	NP bionic_design	JJ_environmental NNS_conditions
conventional_solution	NP effective_communication	NN_hydrogen NN_peroxide
corridor_scenario	NP natural_materials	JJ_artificial NN_intelligence
functional_principles	NP performance_features	JJ_bionic NN_design
bionic_approach	NP piezoelectric_actuators	JJ_comparative NN_analysis
collision_avoidance	NP state-action_samples	JJ_dynamic NNS_disturbances
performance_features	NP biological_systems	JJ_gripping NNS_objects
piezoelectric_actuators	NP electroactive_polymers	JJ_memristive NNS_systems
biodegradable_materials	NP memristive_elements	JJ_physical NNS_models
biological_systems	NP memristive_systems	JJ_technical NNS_solutions
biological terms	NP natural_models	NN_gel NN_structure
data-driven model	NP optimization_problems	NN_manufacturing NNS_processes
	NP swarm intelligence	NNP_applied NNPS_sciences
	NP technical drawings	VBG_branching NN_structure
	NP variable curvatures	
	NP virtual reality	

Стоит отметить, что для большинства выбранных терминов явно проявляется паттерн «прилагательное + существительное», при этом роль словосочетания в предложении однозначно определяется как элемент группы существительного («noun phrase» или кратко NP).

По мере увеличения числа входящих в сочетание элементов увеличивается число предлогов (табл. 7).

Данное явление объясняется интралингвистическими факторами английского языка: паттерн «артикль + прилагательное + существительное» может быть интерпретирован как «конкретный объект, имеющий некоторое значащее отличие от аналогов». Например, словосочетание «shortest path problem» является устойчивым и не разрыва-

ется в тексте (столбец 1 табл. 7). Однако оно указывает на некоторое абстрактное понятие, которое может быть уточнено добавлением определенного артикля (табл. 8).

Таблица 7

ИПТ для триграмм

Позиция	Дерево разбора (по грамматической роли)	Дерево разбора (с учетом частей речи)
the biological solution	NP the biological solution	DT_the JJ_biological NN_solution
the ant colony	NP the bionic solution	DT_the JJ_core NN_neighbor
ant colony optimization	NP the environmental impact	DT_the JJ_pairwise NNS_interactions
the memristive network	NP the pairwise interactions	DT_the NN_design NN_problem
the shortest path	NP nature ’s capabilities	DT_the NN_ground NN_truth
the bionic solution	NP a conventional solution	JJ_social NN_force NN_model
the design problem	NP effectiveness of communication	DT_the JJ_bionic NN_solution
colony optimization algorithm	NP the conventional solution	DT_the JJ_relative NNS_positions
degrees of freedom	NP the functional principles	DT_a JJ_coordinate NN_system
effectiveness of communication	NP the human brain	DT_the JJ_influencing NN_factor
bionic design method		DT_the JJ_neural NN_network
current-controlled memristive system		DT_an JJ_iterative NN_approach
internal state variables		DT_an NN_interaction NN_pattern
requirements and restrictions		DT_the JJ_environmental NN_impact
adjoining mesh facets		DT_the JJ_inverse NN_length
bionic pedals vehicle		DT_the JJ_memristive NN_network
neural network classifier		DT_the JJ_observed NN_complexity
sense of smell		DT_the NN_core NN_neighbor
shortest path problem		DT_the NN_corridor NN_scenario
		JJ_average NN_area NN_error
		JJ_average NN_position NN_error
		JJ_main NN_motion NNS_directions

Таблица 8

ИПТ для квадрограмм

Позиция	Дерево разбора (по грамматической роли)	Дерево разбора (с учетом частей речи)
of the biological solution	NP environmental and ecological variables	DT_the JJ_internal NN_state NNS_variables
the ant colony optimization	NP the shortest path problem	DT_a JJ_neural NN_network NN_classifier
ant colony optimization algorithm		DT_the JJ_average NN_speed NN_error
environmental and ecological variables		

Таким образом, поиск осмысленных словосочетаний требует от эксперта знания грамматики языка, при этом автоматизация процесса извлечения допустимых с точки зрения грамматики элементов текста позволяет значительно снизить объем анализируемого ИПТ.

Заключение

Интерпретация практических результатов позволяет сделать описанные ниже выводы. Синтаксический разбор текста позволяет извлекать из предложений естественного языка синтаксически связные n-граммы. Для их извлечения можно использовать как универсальные зависимости, так и деревья разбора, однако только в последнем случае достигается снижение размерности выходных данных. Процесс синтаксического разбора затратен по времени, однако не требует, помимо грамматической модели, никаких дополнительных предметно-ориентированных баз знаний.

Составленный на основе n-грамм информационно-поисковый тезаурус пригоден для формирования первичного тезауруса предметной области, при этом процесс его формирования не поддается полной автоматизации, т. е. требуется привлечение эксперта по знаниям. Однако объем получаемых описанным способом данных для дальнейшей обработки значительно меньше, чем при использовании ИПТ, полученного стандартным методом. Результат также может быть улучшен путем формирования списка стоп-слов и их последующего исключения из первичного ИПТ.

На основе полученных результатов эксперимента следует отметить, что предполагаемая методика АФПИПТ позволяет осуществлять автоматизированное формирование первичного тезауруса в области бионики на основе корпуса документов на английском языке. Использование данной методики при разработке интеллектуальной системы информационной поддержки процессов создания и развития перспективных бионических технологий значительно упростит процесс формирования и пополнения базы знаний [11], что, в свою очередь, приведет к значительному увеличению пертинентности результатов поисковых запросов, полученных посредством данной системы.

Литература

1. *Баранюк В. В., Смирнова О. С.* Роевой интеллект как одна из частей онтологической модели бионических технологий // International Journal of Open Information Technologies, 2015. Т. 3. № 12. С. 13–17.

2. *Баранюк В. В., Смирнова О. С.* Детализация онтологической модели по роевым алгоритмам, основанным на поведении насекомых и животных // International Journal of Open Information Technologies, 2015. Т. 3. № 12. С. 18–27.

3. *Смирнова О. С., Богорадникова А. В., Блинов М. Ю.* Описание роевых алгоритмов, инспирированных неживой природой и бактериями, для использования в онтологической модели // International Journal of Open Information Technologies, 2015. Т. 3. № 12. С. 28–37.

4. *Sigov A. S., Nechaev V. V., Baranyuk V. V., Koshkarev M. I., Melikhov A. A., Smirnova O. S., Bogoradnikova A. V.* Architecture of domain-specific data warehouse for bionic information resources // Ecology, Environment and Conservation Paper, Nov. 2015. Vol. 21. Suppl. Issue. P. 181–186.

5. *Sigov A., Nechaev V., Baranyuk V., Smirnova O., Melikhov A., Koshkarev M., Bogoradnikova A.* Bionic-oriented information system for innovation activities // Indian Journal of Science and Technology, 2016. Vol. 9. No. 30. Article 98743. 6 p.

6. *Баранюк В. В., Смирнова О. С., Богорадникова А. В.* Интеллектуальная система информационной поддержки развития перспективных бионических технологий: основные направления работ по созданию // International Journal of Open Information Technologies, 2014. № 12. С. 17–19.

7. *Мелихов А. А., Нечаев В. В.* Пополнение базы знаний интеллектуальной системы информационной поддержки развития перспективных бионических технологий: формирование перечня источников // Информационные и телекоммуникационные технологии, 2015. № 28. С. 16–20.

8. *Мелихов А. А.* Применение дерева синтаксического разбора предложений для повышения релевантности результатов частотного анализа текста // Нейрокомпьютеры: разработка, применение, 2016. № 3.

9. *de Marneffe M.-C., Dozat T., Silveira N., Haverinen K., Ginter F., Nivre J., Manning C. D.* Universal Stanford dependencies: A cross-linguistic typology // Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014).

10. *Robertson S.* Understanding inverse document frequency: on theoretical arguments for IDF // Journal of Documentation, 2004. Vol. 60. No. 5. P. 503–520.

11. *Сигов А. С., Нечаев В. В., Кошкарев М. И.* Архитектура предметно-ориентированной базы знаний интеллектуальной системы // International Journal of Open Information Technologies, 2014. Т. 2. № 12. С. 1–6.

Techniques for automated thesaurus population based on grammatically valid sentence units

***Alexandr Alexandrovich Melikhov**, post-graduate student, assistant of the Department, Federal State Budget Education Institution of Higher Education «Moscow Technological University» (MIREA)*

***Ol'ga Sergeevna Smirnova**, junior research fellow, assistant of the Department, Federal State Budget Education Institution of Higher Education «Moscow Technological University» (MIREA)*

The former article regards the task of forming the initial thesaurus, based on automated in-scope publication analysis. The proposed method implements primary reduction of irrelevant units guided by heuristics based on the inner structure of natural language grammar.

Keywords: information resources on bionics; natural language processing; knowledge engineering; thesaurus.

УДК 004.051

ЭВОЛЮЦИЯ ПОДХОДОВ К УПРАВЛЕНИЮ ИНФОРМАЦИОННЫМИ ТЕХНОЛОГИЯМИ

***Сергей Николаевич Маликов**, канд. техн. наук,*
ст. науч. сотр., зам. генерального директора
по научно-конструкторской работе
e-mail: sergej.malikov@bk.ru,
ОАО «НИИ супер ЭВМ»,
http://www.super-computer.ru

В статье приводится анализ подходов к управлению информационными технологиями организации, дается обзор методик и методов управления. Выявляются особенности применения существующих методологий управления для перехода от традиционной ИТ-архитектуры к сервис-ориентированной архитектуре. Рассматриваются преимущества стандартизованных решений для перехода информационных технологий организации на новый системный уровень.

Ключевые слова: информационные технологии; архитектура предприятия; TOGAF; сервис-ориентированная архитектура; Service Management Methods – SMM

Работа выполнена при финансовой поддержке РФФИ (грант 16-06-00486).

DOI: 10.21777/2312-5500-2016-4-51-58

На начальном этапе развития информационных технологий (ИТ) компании-разработчики предлагали бизнесу программные продукты универсального назначения, что подчас требовало от бизнеса «подстройки», «приведения в соответствие» структуры бизнеса к структуре информационных процессов. Возникли, а затем были стандартизованы [1, 13] понятия «уровень зрелости управления организацией», «уровень зрелости управления ИТ», «критические факторы успеха», «метрики оценки». По мере насыщения рынка учет потребностей конкретного потребителя вышел на первый план. Произошла сегментация рынка информационных технологий. Возросли возможности адаптации информационных продуктов к нуждам бизнеса. Идеология управления ИТ