Моделирование и анализ данных
2024. Том 14. № 1. С. 27–40
DOI: https://doi.org/10.17759/mda.2024140102
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2024. Vol. 14, no. 1, pp. 27–40
DOI: https://doi.org/10.17759/mda.2024140102
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.048

Извлечение научно-технических фактов из отраслевых документов на основе методов семантико-синтаксического и концептуального анализа

Кан А.В.*
ФГБУ «НИЦ «Институт имени Н.Е. Жуковского»
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-9410-406X
e-mail: kanav@nrczh.ru

Козловская Я.Д.**
Московский авиационный институт (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-1780-5687
e-mail: yana_kozlovskaia@mail.ru

Токолова А.А.****
Московский авиационный институт (МАИ)
г. Москва, Российская Федерация
e-mail: tokolovaa@gmail.com

Извлечение научно-технических фактов является трудной задачей с точки зрения корректности получаемой информации. Предлагаемая модель выделения фактов основывается на четких представлениях о смысловой структуре текста, выраженной в виде иерархии синтаксических конструкций единиц смысла, что позволяет выявить межфразовые связи в контактно расположенных предложениях. В качестве смысловых единиц используются отдельные слова, словосочетания, присущие конкретной предметной области, и образующие её понятийный состав. Для обработки исходного текста используются процедуры фразеологического, концептуального и семантико-синтаксического анализа текстов.

Ключевые слова: извлечение фактов, семантико-синтаксический анализ, концептуальный анализ, фразеологический анализ, смысловая структура, единица смысла.

Для цитаты:
Кан А.В., Козловская Я.Д., Токолова А.А. Извлечение научно-технических фактов из отраслевых документов на основе методов семантико-синтаксического и концептуального анализа // Моделирование и анализ данных. 2024. Том 14. № 1. С. 27–40. DOI: https://doi.org/10.17759/mda.2024140102

***Кан Анна Владимировна,** кандидат технических наук, доцент МАИ, начальник аналитического отдела ФГБУ «НИЦ «Институт имени Н.Е. Жуковского», ORCID: https://orcid.org/0000-0001-9410-406X, e-mail: kanav@nrczh.ru

****Козловская Яна Дмитриевна,** студент магистратуры, ФГБОУ ВО Московский авиационный институт (МАИ), ORCID: https://orcid.org/0000-0002-1780-5687, e-mail: yana_kozlovskaia@mail.ru

*****Токолова Алина Александровна,** студент магистратуры, институт «Компьютерные науки и прикладная математика» Московский авиационный института (МАИ), г. Москва, Российская Федерация, e-mail: tokolovaa@gmail.com

1. ВВЕДЕНИЕ

Научные публикации уже давно стали основным источником и способом распространения научно-технических открытий. Согласно отчету [1], с 2018 года ежегодное количество научных публикаций, в том числе опросов, тезисов и обзоров, выросло на 22,4% и за 2022 год достигло значения свыше 5,4 миллионов. Такой объем информации качественно затрудняет ручной поиск релевантных сведений и фактов для их дальнейшей обработки.

Научно-технические факты представляют особый интерес для пользователя и под этим терминов понимается событие, понятие или явление. Для корректного выделения фактов из отраслевой документации используются методы извлечения фактологической информации (ИФИ), среди которых семантико-синтаксический, фразеологический и концептуальный анализ текстов. Выделение наименований понятий (НП) основывается на вычислении меры смысловой значимости, а смысловая структура текста представляется в виде предикатно-актантной структуры (ПАС), получаемой методами семантико-синтаксического и концептуального анализа.

2. ОСНОВНЫЕ ПОЛОЖЕНИЯ ФРАЗЕОЛОГИЧЕСКОГО КОНЦЕПТУАЛЬНОГО АНАЛИЗА ТЕКСТОВ

Фразеологический концептуальный анализ [2] используется для установления единиц смысла предложения, которые и формируют смысловое содержание текста.

Основные положения концепции фразеологического концептуального анализа текстов:

- Смысловое содержание текстов выражается с помощью единиц смысла.
- Понятие – самая устойчивая единица смысла.
- Объекты предложения обладают особыми признаками, выражающимися через предикатно-актантную структуру (ПАС) и набором отношений с другими объектами.
- Сверхфразовые единства формируются из предложений и представляются в виде последовательностей предложений (связного текста).

Предикатно-актантная структура (ПАС) – представляет предложения в виде понятий-предикатов, устанавливающих связи между объектами, и в виде понятий-актантов, содержащих признаки объектов. ПАС обеспечивает автоматический перевод текстов с естественных языков на формализованные и наоборот.

3. ОПРЕДЕЛЕНИЕ ЗНАЧИМЫХ НАИМЕНОВАНИЙ ПОНЯТИЙ В ТЕКСТЕ

Значимые наименования понятий [3,4] выделяются методами статистического, синтаксического и концептуального анализа текстов. Сложность этого процесса заключается в фиксировании границ наименований понятий, несущих ключевую смысловую нагрузку.

Семантические методы позволяют с помощью эталонных словарей идентифицировать значимые слова и словосочетания определенной предметной области.

Статистические методы позволяют определить значимые понятия путем присвоения объектам текста весовых коэффициентов на основе частотного анализа встречаемости в конкретном документе и во всем корпусе документов.

Синтаксические методы позволяют определить, какую синтаксическую роль играют значимые слова и словосочетания, определяя их синтаксическую роль в предложении, как субъект, предикат или объект.

Смысловая значимость слов в контексте документа оценивается статистической мерой *TF•IDF* [5].

$$TF\left(concept, Text\right) = \frac{l_{concept}}{\sum_k l_k} - \text{ частота слова в документе,}$$

где $l_{concept}$ – число вхождений наименования понятия $concept$ в документ;
$\sum_k l_k$ – общее число наименований понятий в данном документе.

$$IDF\left(concept, M\right) = \log \frac{\left|M\right|}{\left|\left\{Text_n \in M \mid concept \in Text_n\right\}\right|}$$

$IDF\left(concept, M\right)$ – обратная частота документа, т.е. инверсия частоты, с которой наименование понятия встречается в корпусе текстов, каждому слово соответствует одно значение IDF, где $\left|M\right|$ – число документов в корпусе;
$\left|\left\{Text_n \in M \mid concept \in Text_n\right\}\right|$ – число документов из массива документов M, в которых встречается понятие $concept$ (когда $l_{concept} \neq 0$).

Эта статистическая мера увеличивает значение слова пропорционально частоте его появления в тексте, однако, уменьшает это же значение при появлении слова в большом количестве документов.

4. МОДЕЛЬ ВЫЯВЛЕНИЯ ФАКТОВ В КОЛЛЕКЦИИ ТЕКСТОВ

Факт $Fact$ состоит из контактно расположенных предложений текста и содержится во фрагменте текста. Таким образом, факт можно представить в виде некоторой последовательности предложений:

$$Fact = \left(Sent_u , Sent_{u+1} , \ldots , Sent_t \right),$$

где $Sent_u , Sent_{u+1} , \ldots , Sent_t$ – последовательность предложений факта;
u – первое предложение факта;
t – последнее предложение факта.

Предложение факта $Sent$ представляет собой кортеж элементов факта:

$$Sent = el_1 , el_2 , \ldots , el_i ,$$

где el_i – элемент факта,
i – количество элементов факта.

Элементом текста el является слово или знак препинания:

$$\text{GPS } el = \left\{ w, pm \right\},$$

где w – слово;
pm – знак препинания.

Каждый элемент текста el обладает грамматическими и синтаксическими признаками:

$$GSP\left(el \right) = \left(GP, SP \right),$$

где GP – грамматические признаки;
SP – синтаксические признаки.

Термином «понятие» определяется некоторый социально значимый мыслительный образ, представляемый в виде отдельного слова или в виде устойчивого фразеологического словосочетания. Понятие $concept$ можно представить математически в виде одного элемента факта или в виде совокупности нескольких элементов факта:

$$concept = \sum \left\{ el_j \right\}_{j=1}^{\text{V– счетное}} .$$

Каждое понятие $concept$ имеет унифицированную и нормализованную формы, получаемые функциями унификации и нормализации:

$Unif : concept \to concept^{un}$ – функция унификации,

$Norm : concept \to concept^{nf}$ – функция нормализации,

где $concept^{un}$ – унифицированная форма понятия;

$concept^{nf}$ – нормализованная форма понятия.

Смысл предложения факта MF выражается через его предикатно-актантную структуру (систему элементарных смысловых триад):

$$MF = \{PSO_1, PSO_2, \ldots, PSO_e\},$$

где e – количество элементарных смысловых триад в предложении.

Элементарная смысловая триада представляется в виде кортежа субъекта S, предиката P и объекта O:

$$PSO = S, P, O,$$

где P – смысловая связь между субъектом и объектом (предикат);
S – главное слово или словосочетание элементарного высказывания (субъект);
O – зависимое слово или словосочетание элементарного высказывания (объект).

Полное описание факта [5] представляется в виде совокупности предикатно-актантных структур предложений факта:

$$FullFact = \sum \{MF_v\}_{v=1}^{\text{V– счетное}}, FullFact \in Fact.$$

5. ПРОЦЕСС ИЗВЛЕЧЕНИЯ НАУЧНО-ТЕХНИЧЕСКИХ ФАКТОВ НА ОСНОВЕ СИНТАКСИЧЕСКОЙ МОДЕЛИ ТЕКСТА

Реализованная синтаксическая модель текста на основе системы обобщенных синтагм, которые выражают форму единицы смысла и состоят из сочетания символов обобщенных грамматических классов слов, входящих в состав словосочетаний эталонного словаря, позволяет анализировать его структуру, извлекать понятийный состав и определять смысловые отношения. Эта модель также помогает извлекать ключевые синтаксические конструкции предложения и классифицировать их элементы. Значимые именованные сущности регистрируются и обрабатываются в соответствии с их синтаксической ролью. Затем эти элементы унифицируются в соответствии с требованиями системы.

Алгоритм автоматического выявления и формализации фактов в текстах:

Шаг 1. Разделить исходный текст на предложения и выполнить морфологический анализ предложений.

Шаг 2. Определить именные и глагольные словосочетания и установить их синтаксическую роль в предложении при помощи упрощенного семантико-синтаксического анализа.

Шаг 3. Составить частотный словарь слов и словосочетаний, определить какие из этих слов и словосочетаний являются значимыми для данного текста.

Шаг 4. Присвоить уникальные идентификаторы каждому словосочетанию согласно словарю унифицированных формализованных представлений наименований понятий (УФПНП) и сопоставить их исходные формы с унифицированными формами представления, указав все их местоположения в тексте (номера предложений).

Шаг 5. Определить номера предложений, содержащих значимые слова и словосочетания, используя словарь указателей связей предложений (УСП).

Шаг 6. Определить связи между предложениями, содержащими ключевые понятия, и их окружением. Установить границы описания фактов в тексте, используя разметку текста по указателям смысловых связей и обобщенным наименованиям понятий.

Шаг 7. Присвоить каждому текстовому описанию фактов идентификационный номер, содержащий порядковый номер события, код и тип документа.

Шаг 8. Определить главные и второстепенные члены предложения, границы словосочетаний, построить дерево зависимостей предложения, построить ПАС и сформировать «скелет» предложения, используя информацию, полученную на шаге 2.

Шаг 9. Определить обобщенную синтагму и построить формализованное представления для каждого словосочетания.

Шаг 10. Соотнести полученные методом концептуального анализа словосочетания со словосочетаниями, полученными путем синтаксического анализа.

Шаг 11. Произвести нормализацию.

Шаг 12. Расчленить описание каждого факта на составные элементы – формализованное представление элементов ПАС, «скелет» предложения с указанием номеров словосочетаний в эталонном концептуальном словаре (ЭКС).

Шаг 13. Выполнить генерацию формализованных представлений предложений факта в обобщенное формализованное представление его смысловой структуры.

Шаг 14. Произвести преобразование обобщенного формализованного представления в его машинную форму.

Основным отличием предлагаемой модели от существующих заключается в том, что в рамках этой модели смысловое представление текста выражается в виде иерархии синтаксических конструкций единиц смысла.

Описания конкретных фактов в тексте могут отображаться его контактно расположенной последовательностью предложений, связанных межфразовыми связями [6]. Рассмотрим в качестве такого примера текст, представленный в таблице 1.

Таблица 1

Фрагмент текста

Стратегический бомбардировщик “B -52” более полувека стоит на вооружении ВВС США. В настоящее время продажа за пределами США боинга “B-52” запрещена, так как его летные характеристики многократно превосходят зарубежные аналоги. Самолеты бомбардировщики “Б-52” были введены в строй с 1955 года. Всего было построено 728 бомбардировщиков. По проекту каждый из этих самолетов несет на борту до 51 единицы боеприпасов.

Разобьем исходный текст на предложения и присвоим каждому идентификационный индекс, состоящий из порядкового номера события, номера предложения в нем, источника информации и даты публикации (см. табл. 2).

Таблица 2

Перенумерованные предложения с идентификационными индексами

1. *Стратегический бомбардировщик "В -52" более полувека стоит на вооружении ВВС США.* – ***458496_1***
2. *В настоящее время продажа за пределами США боинга "В-52" запрещена, так как его летные характеристики многократно превосходят зарубежные аналоги.* – ***458496_2***
3. *Самолеты бомбардировщики "Б-52" были введены в строй с 1955 года.* – ***458496_3***
4. *Всего было построено 728 бомбардировщиков.* – ***458496_4***
5. *По проекту каждый из этих самолетов несет на борту до 51 единицы боеприпасов.* – ***458496_5***

В приведенном фрагменте текста межфразовая связь между предложением № 4 и предложением № 5 обусловлена местоименной анафорой ***«их этих»,*** связь между предложениями № 1, № 2 № 3 и № 4 обусловлена синонимичными конструкциями «***Стратегический бомбардировщик – самолет бомбардировщик***». Наличие межфразовых связей является необходимым условием выделения описания конкретного факта в тексте [7]. Дополнительным условием, предоставленным в таблице 3, является смысловая связь (родовидовые и синонимичные отношения) между частью наименований понятий этого фрагмента, находящимися в разных предложениях (в скобках указаны номера предложений).

Таблица 3

Смысловая связь между частью наименований понятий фрагмента, находящихся в разных предложениях

стратегический бомбардировщик "В -52" ***(№ 1)*** *– самолеты бомбардировщики "Б-52"* ***(№ 3)*** *– бомбардировщиков* ***(№ 4)***
стратегический бомбардировщик "В -52" ***(№ 1)*** *– боинга "В-52"* ***(№ 2)*** *– самолеты бомбардировщики "Б-52"* ***(№ 3)***

В рамках разработанной синтаксической модели возможно осуществить автоматическое преобразование исходного текста в его формализованную синтаксическую структуру, аналогичное приведенной выше структуре. В процессе такого преобразования для получения унифицированного представления элементов ПАС необходимо дополнительно выполнить нормализацию форм слов наименований понятий и их унификацию по словарю УФПНП. Результат такой операции приведен в таблице 4.

Таблица 4

Унифицированные формы представления наименований понятий

Наименование понятия	Унифицированная форма представления
стратегический бомбардировщик	самолет
боинг	самолет

Наименование понятия	Унифицированная форма представления
ВВС США	войско
боеприпасы	оружие
вооружение	оружие
США	страна

В таблице 5 указывается идентификатор факта и приводится его исходная текстовая форма, указывается состав ПАС (в мнемонике грамматических классов слов), унифицированный «скелет», идентификатор факта, номер предложения; приводится формализованное представление ПАС предложения в сокращенной форме в виде символов обобщенных синтагм и в виде унифицированных форм главных слов ПАС; приводится формализованное представления «скелета» (SklRus) предложений факта (в сокращенном виде – только главные слова, в полном виде эти элементы представлены номерами элементов словаря ЭКС); приводится формализованное представление предложений (SenRus) в виде последовательности нормальных слов предложения [7].

Таблица 5

Результаты семантико-синтаксического и концептуального анализа предложений описания факта, приведенного в таблице 1

Исходное предложение № 1	
Стратегический бомбардировщик “B -52” более полувека стоит на вооружении ВВС США.	
Формализованное описание элементов ПАС предложения ***PSO={V, N, N}****={стоять; бомбардировщик; вооружение}*	
Формализованное представление элементов ПАС предложений	
Predicate (P)	*лА = стоять #лА=стоять*
Subject (S)	*FA = бомбардировщик #FA=бомбардировщик*
Object (O)	*ЁК = вооружение #ЁК=вооружение*
Формализованное представления ПАС предложений (PSORus)	
PSO= лAFAЁК*= стоять; бомбардировщик; вооружение # лAFAЁК*	
Формализованное представления «скелета» предложений (SklRus)	
N»N»NVFN=FA»»@A9A»»YAлAыAЁКьAьA=*бомбардировщик “B -52” полувек стоять на вооружение ВВС США*	
Формализованное представления предложений (SenRus)	
PiFA»»@A9A»»cAYAлAыAЁКьAьA=*стратегический бомбардировщик “B -52” более полувек стоять на вооружение ВВС США*	
Исходное предложение № 2	
В настоящее время продажа за пределами США боинга “B-52” запрещена, так как его летные характеристики многократно превосходят зарубежные аналоги.	
Формализованное описание элементов ПАС предложения ***PSO={N, N, V, N, V, N}****={продажа; боинг; запрещен; характеристика; превосходит; аналог}*	
Формализованное представление элементов ПАС предложений	
Predicate (P)	*аА =запрещен # аА =запрещен*
	Юр =превосходит # Юр = превосходит

Subject (S)	*uB = продажа #uB =продажа*
	vB = характеристика # vB = характеристика
Object (O)	*FA = боинг # FA =боинг*
	FA = аналог # FA =аналог
Формализованное представления ПАС предложений (PSORus)	
***PSO_1= uBFAaA** = продажа; боинг; запрещен # uBFAaA* ***PSO_2= vBЮpFA** =характеристика; превосходит; аналог# vBЮpFA*	
Формализованное представления «скелета» предложений (SklRus)	
***FNNFN»N»K, NVN =ыАГтvHRWbTKxaИ,, АТежА=** в время продажа за предел боинг запрещен, характеристика превосходить аналог*	
Формализованное представления предложений (SenRus)	
***ыАПiИмиВъААСьAFA»»9A»»aA,, cAcAяAHTvBcAлAHTFA**= в_настоящий_время _продажа_за_пределам_США_боинг_»B-2»_запрещен,_так_как_его_летный _характеристика_многократно_превосходить_зарубежный_аналог*	
Исходное предложение № 3	
Самолеты бомбардировщики "Б-52" были введены в строй с 1955 года.	
Формализованное описание элементов ПАС предложения ***PSO={V, N, N}=**{введен; самолет; строй}*	
Формализованное представление элементов ПАС предложений	
Predicate (P)	*aA = введен # aA =введен*
Subject (S)	*AA = самолет # AA =самолет*
Object (O)	*Dx = строй #Dx=строй*
Формализованное представления ПАС предложений (PSORus)	
***PSO= aAAADx**= введен; самолет; строй # aAAADx*	
Формализованное представления «скелета» предложений (SklRus)	
***N»N»LKFNFN = eW»»@A9A»»ЯAaAыADxщA9=** самолет "Б-52" был введен в строй с 1955 год*	
Формализованное представления предложений (SenRus)	
***AAFA»»9A»»ЯAaAыADxщA9AHA=**самолет_бомбардировщик_"Б-52"_был_введен _в_строй_с_1955 _год*	
Исходное предложение № 4	
Всего было построено 728 бомбардировщиков.	
Формализованное описание элементов ПАС предложения ***PO={V, N}=**{построен; бомбардировщик}*	
Формализованное представление элементов ПАС предложений	
Predicate (P)	*aA = построен #aA=построен*
Subject (S)	*FA = бомбардировщик # FA =бомбардировщик*
Формализованное представления ПАС предложений (PSORus)	
***PO=aAFA** = построен; самолет # aAFA*	
Формализованное представления «скелета» предложений (SklRus)	
***LKN =** был_построен_бомбардировщик*	
Формализованное представления предложений (SenRus)	
***ЧМЯAaA9AFA =** всего_был_построен_728_бомбардировщик*	
Исходное предложение № 5	
По проекту каждый из этих самолетов несет на борту до 51 единицы боеприпасов.	

Формализованное описание элементов ПАС предложения	
PSO={V, N, N}=*{несет; самолет; боеприпас}*	
Формализованное представление элементов ПАС предложений	
Predicate (P)	*ЪY = несет # ЪY =несет*
Subject (S)	*AA = самолет # AA =самолет*
Object (O)	*AA = боеприпас # AA = боеприпас*
Формализованное представления ПАС предложений (PSORus)	
PSO = ЪYAAAA = *несет; самолет; боеприпас #**ЪYAAAA***	
Формализованное представления «скелета» предложений (SklRus)	
FN FN VFNFN= *по_проект_из_этот_нести_на_борт_до_51_единиц*	
Формализованное представления предложений (SenRus)	
xAAAHTфAФAAAЪYыAHAфA9ABBAA =*по_проект_каждый_из_этот_самолет _несет_на_борт_до_51_единица_боеприпас*	

Результатом работы алгоритма является формализованное представление предложений, обобщающее смысловое содержания фактологической информации. Для генерации формализованного представления обобщенного смыслового содержания факта выделялись унифицированные представления элементов ПАС предложений. Результаты такого представления приведены в таблице 6 в виде формализованного и индексного представления элементов ПАС факта. Каждый элемент ПАС сопровождается его весовым коэффициентом, индексом синтаксической роли, номером словосочетания в словаре ЭКС и уникальным идентификатором факта.

Таблица 6. Формализованное и индексное представление элементов ПАС факта текста № 387473. В этих представлениях номер события был заменен на его уникальный идентификационный номер (ID=008154)

Формализованное представление элементов ПАС факта
PSO={V, N, N}->   ***P****={стоять_**064**_065785, запрещен_**049**_087631, превосходит_**057**_075324, введен_**047**_087631, построен_**049**_098713, несет_**058**_075692};*   ***S****={бомбардировщик_**025**_034758, продажа_**039**_062584, характеристика_**039**_07234, самолет_**029**_046273};* ***O****={вооружение_**038**_086759, аналог_**020**_015638, строй_**029**_063348}*
Индексное представление элементов ПАС факта
стоять_{064_065785_P_008154}, запрещен_{049_087631_P_008154}, превосходит_{057_075324_P_008154}, введен_{047_087631_P_008154}, построен_{049_098713_P_008154}, несет_{058_075692_P_008154}, бомбардировщик_{025_034758_S_008154}, продажа_{039_062584_S_008154}, характеристика_{039_07234_S_008154}, самолет_{029_046273_S_008154}, вооружение_{038_086759_O_008154}, аналог_{020_015638_O_008154}, строй_ {029_063348_P_008154}

Таким образом, как видно из таблицы 5 и таблицы 6, основная задача формализации смыслового представления описания факта заключается в генерации совокупности унифицированных представлений ПАС. Полученная совокупность формализованных представлений ПАС, «скелета» и предложения позволяет не только обеспечить возможность поиска по любому элементу их формализованного описания, но и обеспечивает возможность последовательного перехода от каждого нижестоящего элемента формализованного описания к вышестоящему, а также возможность перехода в обратном порядке. Такое представление смысла предложений обеспечивает реализацию всего спектра семантических операций над смысловым содержанием факта.

Полученное формализованное представление обобщенного смыслового содержания факта позволяет производить поиск и сопоставление идентичных или близких по смысловому содержанию фактов, а также обеспечивает возможность их классификации по различным основаниям: по их содержанию, по именам субъектов или объектов фактов или по тональности отношений между ними и др.

6. ЗАКЛЮЧЕНИЕ

В статье описано решение задачи автоматического извлечения фактов из научно-технологических документов на основе методов их семантико-синтаксического и концептуального анализа.

Представленная модель алгоритма автоматического извлечения фактов решает проблему формализации и унификации смыслового содержания фактов. Результатами работы этого алгоритма является представление смысловой структуры фактов, полученных из исходных корпусов текста, в виде совокупности формализованных представлений ПАС, обеспечивающих возможность поиска по любому элементу их формализованного описания. Такое представление смысла предложений обеспечивает реализацию всего спектра семантических операций над смысловым содержанием факта, в том числе классификации и сопоставления, что дает возможность выполнять проверку корректности извлеченной информации.

Литература

1. *Curcic D.* Number of Academic Papers Published Per Year // Wordsrated URL: https://wordsrated.com/number-of-academic-papers-published-per-year/#:~: text=As%20of%202022%2C%20over%205.14,5.03%20million%20papers%20were%20published. (дата обращения: 09.10.2023).
2. *Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А.* Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации. – М.: Изд-во Русский мир, 2004.
3. *Хорошилов Ал-др А., Мусабаев Р.Р., Козловская Я.Д., Никитин Ю.В., Хорошилов А.А.* Автоматическое выявление и классификация информационных событий в текстах СМИ // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2020. № 7. С. 27–38. DOI: 10.36535/0548-0027-2020-07-4
4. *Хорошилов Ал-др. А., Никитин Ю.В., Хорошилов Ал-ей. А., Будзко В.И.* Автоматическое создание формализованного представления смыслового содержания неструктурированных

текстовых сообщений СМИ и социальных сетей // Системы высокой доступности, № 3, том.10, 2014, С. 36–51.

5. *Кан А.В., Козловская Я.Д., Кадушкин Н.А., Хорошилов Ал-р А.* Автоматическая кластеризация документов СМИ на основе анализа их смыслового содержания // Моделирование и анализ данных. 2020. Том 10. № 3. С. 24–38. DOI: https://doi.org/10.17759/mda.2020100302
6. *Богатырев М.Ю.* Извлечение фактов из текстов естественного языка с применением концептуальных графовых моделей // Известия ТулГУ. Технические науки. – 2016. – № 7. – Ч. 1.
7. *Хорошилов Ал-др А., Козловская Я.Д., Мусабаев Р.Р., Красовицкий А.М., Хорошилов Ал-ей А.* Определение тональности сообщений СМИ методом их концептуального анализа // Моделирование и анализ данных. 2019. № 4. DOI: 10.17759/mda.2019090405

Extracting Scientific and Technical Facts from Industry Documents Based on Methods of Their Semantic-syntactic and Conceptual Analysis

Anna V. Kan*
FSBI "National Research Center" Institute named after N.E. Zhukovsky"
Moscow, Russian Federation
ORCID: https://orcid.org/0000-0001-9410-406X
e-mail: kanav@nrczh.ru

Yana D. Kozlovskaya**
Moscow Aviation Institute (MAI), Moscow, Russian Federation
ORCID: https://orcid.org/0000-0002-1780-5687
e-mail: yana_kozlovskaia@mail.ru

Alina A. Tokolova***
Moscow Aviation Institute (MAI), Moscow, Russian Federation
e-mail: tokolovaa@gmail.com

Extraction of scientific and technical facts is a difficult task in terms of correctness of the obtained information. The proposed fact extraction model is based on clear ideas about the semantic structure of the text, expressed as a hierarchy of syntactic constructions of meaning units, which allows identifying interphrase relations in contacted sentences. Individual words, word combinations inherent to a particular subject area and forming its conceptual composition are used as meaning units. The procedures of phraseological, conceptual and semantic-syntactic analysis of texts are used to process the source text.

Keywords: fact extraction, semantic-syntactic analysis, semantic-syntactic analysis, conceptual analysis, semantic triad.

For citation:
Kan A.V., Kozlovskaya Y.D., Tokolova A.A. Extracting Scientific and Technical Facts from Industry Documents Based on Methods of Their Semantic-syntactic and Conceptual Analysis. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2024. Vol. 14, no. 1, pp. 27–40. DOI: https://doi.org/10.17759/mda.2024140102 (In Russ., abstr. in Engl.).

****Anna V. Kan,*** Candidate of Technical Sciences, Associate Professor, Moscow Aviation Institute (MAI), Head of the Analytical Department, Federal State Budgetary Institution "National Research Center" Institute named after N.E. Zhukovsky", e-mail: kanav@nrczh.ru

*****Yana D. Kozlovskaya,*** master's student, Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, ORCID: https://orcid.org/0000-0002-1780-5687, e-mail: yana_kozlovskaia@mail.ru

******Alina A. Tokolova,*** master's student, Institute of Computer Science and Applied Mathematics Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, e-mail: tokolovaa@gmail.com

References

1. Curcic D. Number of Academic Papers Published Per Year // Wordsrated URL: https://wordsrated.com/number-of-academic-papers-published-per-year/#:~:text=As%20of%202022%2C%20over%205.14,5.03%20million%20papers%20were%20published (Accessed: 09.10.2023).
2. Belonogov G.G., Kalinin Y.P., Khoroshilov A.A. Computer linguistics and perspective information technologies. Theory and practice of building systems of automatic processing of text information. – Moscow: Izd-vo Russky Mir, 2004.
3. Khoroshilov Al-Dr. A., Musabaev R.R., Kozlovskaya Y.D., Nikitin Y.V., Khoroshilov A.A. Automatic detection and classification of information events in mass media texts// Scientific and Technical Information. Series 2: Information processes and systems. 2020. № 7. С. 27–38. DOI: 10.36535/0548-0027-2020-07-4
4. Khoroshilov Al-Dr. A., Nikitin Y.V., Khoroshilov Al-ey. A., Budzko V.I. Automatic creation of formalized representation of semantic content of unstructured text messages of mass media and social networks // Systems of High Availability, No.3, Vol.10, 2014, pp. 36–51.
5. Kan A.V., Kozlovskaya Y.D., Kadushkin N.A., Khoroshilov Al-r A. Automatic clustering of media documents based on the analysis of their semantic content // Modeling and Data Analysis. 2020. Vol. 10. No. 3. С. 24–38. DOI: https://doi.org/10.17759/mda.2020100302
6. Bogatyrev, M. Yu. Fact extraction from natural language texts using conceptual graph models // Izvestiya TulSU. Technical Sciences. –2016. – № 7. – Vol. 1.
7. Khoroshilov Al-Dr. A., Kozlovskaya Ya.D., Musabaev R.R., Krasovitsky A.M., Khoroshilov Al-ey A. Determination of the tone of media messages by their conceptual analysis method // Modeling and Data Analysis. 2019. № 4. DOI: 10.17759/mda.2019090405

Получена 04.03.2024
Принята в печать 18.03.2024

Received 04.03.2024
Accepted 18.03.2024