УДК 004.853
DOI: 10.26102/2310-6018/2022.39.4.006

Извлечение морфологических признаков технических систем из русскоязычных патентов по деревьям зависимостей

С.С. Васильев✉, Д.М. Коробкин, С.А. Фоменков, В.О. Ермилов, А.С. Тозик

Волгоградский государственный технический университет,
Волгоград, Российская Федерация
svasilev2012@yandex.ru✉

Резюме. В статье представлена методология извлечения морфологических признаков технических систем в виде компонентов устройства и связей между ними. Объектом анализа для извлечения данных выступает главный пункт формулы изобретения в текстах русскоязычных патентов. Информация о компонентах устройства является наиболее фундаментальной и важной и может использоваться во множестве задач анализа патентного массива, а поиск эффективных подходов по извлечению такой информации все еще продолжается. В настоящем исследовании областью применения указанных данных рассматривается направление автоматизированного изобретательства. Целью работы являлся анализ качества извлечения данных по деревьям зависимостей для русского языка. Деревья зависимостей являются результатом работы систем синтаксической разметки естественного языка. Для сравнения были выбраны следующие синтаксические анализаторы: UdPipe, Stanza, DeepPavlov и spaCy. Выходные данные представлены в виде семантических структур SAO (Subject-Action-Object). Дана оценка качества извлечения данных с помощью метрик точности (precision), полноты (recall) и F1-меры. Для этого вручную было размечено 20 патентных формул с 252 структурами SAO. При текущих методологических ограничениях из тестовой выборки в лучшем случае удалось извлечь 79 % связок SAO в терминах метрики recall при нестрогой оценке данных, т. е. без учета полноты именных групп субъекта и объекта. Значение F1-меры по инструментам несколько ниже и находится в пределах от 48 % до 66 % в зависимости от типа оценки. Сделаны общие выводы по текущему уровню работы синтаксических анализаторов в рамках исследуемой области применения. Материалы статьи представляют практическую ценность при проработке эффективных подходов извлечения структурированных данных из русскоязычного патентного массива.

Ключевые слова: патент, извлечение данных, компоненты устройств, деревья зависимостей, SAO.

Благодарности: исследование выполнено за счет гранта Российского научного фонда № 22-21-20125, https://rscf.ru/project/22-21-20125/, и Администрации Волгоградской области.

Для цитирования: Васильев С.С., Коробкин Д.М., Фоменков С.А., Ермилов В.О., Тозик А.С. Извлечение морфологических признаков технических систем из русскоязычных патентов по деревьям зависимостей. *Моделирование, оптимизация и информационные технологии.* 2022;10(4). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=1246 DOI: 10.26102/2310-6018/2022.39.4.006

Extraction of morphological features of technical systems from Russian patents using dependency tree analysis

S.S. Vasiliev✉, D.M. Korobkin, S.A. Fomenkov, V.O. Ermilov, A.S. Tozik

Volgograd State Technical University,
Volgograd, Russian Federation
svasilev2012@yandex.ru✉

Abstract. The article presents a methodology for extracting morphological features of technical systems in the form of device components and connections between them. The main section of Russian patents claims is chosen as the subject of the study for data extraction. Information about device components is the most fundamental and important part. It can be used in many tasks of computer-aided patent analysis, while the search for effective approaches to extracting such information is still in progress. In the present inquiry, computer-aided development of inventions is considered as a range of applications for such data. The aim of the study was to explore the quality of data extraction using dependency tree analysis for Russian language. The dependency tree is the result of markup by natural language processing tools. Several parsers were chosen for the comparison: UdPipe, Stanza, DeepPavlov and spaCy. The output data are presented in the form of semantic SAO (Subject-Action-Object) structures. The quality of data extraction has been evaluated using precision, recall and F1 metrics. For this purpose, 20 patent claims with 252 SAO structures were manually marked. Under the current methodological constraints, we were able to extract from the dataset 79 % of the SAO structures at best according to the recall metric with a non-strict data evaluation, i.e. without accounting for the completeness of noun groups. The value of F1-measure is lower and ranges from 48 % to 66 % depending on the evaluation type. Conclusions are drawn about the current level of the syntactic analyzer performance within the field of application under review. The results can be useful for developing efficient approaches to extracting structured data from Russian patent arrays.

Keywords: patent, data extraction, device components, dependency trees, SAO.

Acknowledgements: the research was supported by the grant of the Russian Science Foundation No. 22-21-20125, https://rscf.ru/project/22-21-20125/, and the Administration of Volgograd Oblast.

For citation: Vasiliev S.S., Korobkin D.M., Fomenkov S.A., Ermilov V.O., Tozik A.S. Extraction of morphological features of technical systems from Russian patents using dependency tree analysis. *Modeling, Optimization and Information Technology*. 2022;10(4). Available from: https://moitvivt.ru/ru/journal/pdf?id=1246 DOI: 10.26102/2310-6018/2022.39.4.006 (In Russ.).

Введение

Патентный массив является богатым источником технической информации. По сообщениям в литературе, более 90 % всей научно-технической информации содержится в патентах [1]. Анализ патентных данных производится для самых разных задач: технологического прогнозирования [2], выявления нарушения патентных прав [3], построения дорожных карт [4], извлечения знаний для концептуального проектирования [5] и т. д. С постоянным увеличением в мире количества патентных документов продолжается и поиск эффективных подходов к их анализу.

В настоящем исследовании затронуты вопросы извлечения из патентов морфологических признаков технических систем (ТС) – признаков структуры изобретений. К таким отнесем элементы конструкций и связи между ними, как наиболее полно раскрывающие сущность изобретения. Данные признаки могут быть использованы в морфологических методах проектирования новых устройств, системах автоматизированного изобретательства – CAI (Computer Aided Innovation) [6]. При этом, будучи системообразующими, морфологические признаки могут использоваться и в других задачах патентного анализа.

В своей предыдущей работе [7] авторы представили систему по извлечению морфологических признаков, которая во многом основана на использовании эвристик. Однако необходимость повышения общего качества извлечения информации ведет к пересмотру возможных подходов. Одним из самых доступных является извлечение данных по результатам работы синтаксических анализаторов – деревьям зависимостей (dependency tree) или деревьям разбора.

Целью данного исследования является анализ качества работы современных систем синтаксической разметки русского языка применительно к задаче извлечения

морфологических признаков ТС. В задачи исследования входят формирование методологии извлечения данных, реализация системы извлечения и подготовка данных для оценки качества работы, а также анализ уровня качества по отношению к существующим подходам.

Общеизвестен ряд факторов, препятствующих эффективному анализу патентных документов с помощью традиционных средств обработки естественного языка – это и специфичная терминология, и чрезмерная длина предложений, и сложность структуры патентных формул [7-9]. Однако мотивом к данному исследованию послужил тот факт, что инструменты обработки естественного языка в последнее время сильно прогрессируют. Так, синтетические оценки ряда моделей анализаторов по метрике LAS (Labeled Attachment Score) уже достигли уровня более 90 % для русского языка [10].

Непосредственным источником данных, аналогично [7], также предлагается рассматривать формулу изобретения в тексте патента на устройство (далее ФИУ) как выражающую сущность и наиболее значимые признаки изобретения. При этом рассматривается только русскоязычный патентный домен.

Предшествующий уровень техники

Рассмотрим работы по извлечению технических знаний из патентов, которые наиболее близки к текущей задаче извлечения морфологических признаков ТС. Подходы по извлечению таких знаний можно условно разделить на две больших группы: извлечение на основе методов машинного обучения и на основе знаний грамматики и правилах. Последнее включает часто используемый исследователями формализм SAO (Subject-Action-Object).

Например, в [11] из патентов извлекается техническая информация, в частности, компоненты устройства, в виде структур SAO исключительно на основе морфологической разметки и грамматических правил.

В статье [12] отмечается важность установления основных компонентов изделия и предлагается метод для их извлечения из текстов патентов. Основной компонент в [12] определяется как находящийся в более тесном контакте с другими, расположенный в более выгодном месте и выполняющий основные функции. При этом также использовался формализм SAO и структуры извлекались по POS-тегам (Part Of Speech) с помощью инструмента spaCy [13].

В [3] определяются компоненты из формул изобретения разных типов патентов с целью дальнейшего расследования нарушения патентных прав. При этом методика извлечения компонентов формулы основана на анализе дерева разбора.

Работа [14] посвящена построению графа знаний о конструкциях изобретений на основе анализа китайских патентов. Концепты для графа знаний разделены на три категории: сущности, действия и атрибуты. Данные извлекались в виде SAO-структур с помощью комбинированного анализа деревьев зависимостей. Средняя заявленная точность (precision) и полнота (recall) извлечения сущностей по [14] составила 0,9465 и 0,9217 соответственно. Показатели являются довольно высокими, однако стоит отметить некоторую специфику китайских патентов, например, заявлено, что в них есть явные описания иерархических отношений между компонентами патентуемого объекта, что значительно упрощает анализ. В русскоязычных патентах такой информации в явной форме нет.

В последнее время все активнее развиваются подходы на базе машинного обучения. В обзоре [15], рассмотрены подходы глубокого обучения применительно к задачам анализа патентов и выделено восемь классов таких задач. Текущую задачу можно отнести к разряду вспомогательных, т. е. результаты которых могут быть

использованы для анализа патентов на следующих этапах [15]. В статье [16] для идентификации сущностей и извлечения семантических отношений из патентных документов используют модели BiLSTM-CRF и BiGRU-HAN соответственно. Авторы [16] заявляют, что им удалось достичь уровня F1-меры в 92,2 % при определении сущностей и 51,5 % при определении связей между ними.

В работе [17] заявлен передовой уровень извлечения отношений между сущностями патентного документа. Метод реализован на основе определения ключевых признаков терминов домена по улучшенному алгоритму Text-Rank и использованию их в модели BiLSTM для классификации отношений. Патентная терминология определялась по другому алгоритму. Всего классификация проводилась для семи типов отношений, таких как «компонент-целое», «целое-компонент», «продукт-материал» и т. д., при этом среднее значение F1-меры на тестовой выборке в [17] составило 90,5 %.

В [8] с помощью гибридной нейронной сети BiLSTM+CNN+CRF из формул изобретений на английском языке извлекались связки глаголов с их аргументами. Маркировка сущностей производилась модифицированной BIO-разметкой (Inside–Outside-Beginning). По оценкам авторов [8] своего метода F1-мера составляет 97,5226 %.

В [15] также отмечается потребность в больших наборах данных для обучения, при этом наблюдается тенденция к уменьшению объема обучающих данных с помощью обучения на вспомогательных задачах.

Методология извлечения данных

Единицей извлечения данных выступает структура SAO, семантически описывающая компоненты устройства и связи между ними. При этом извлекаются только связки с определенными семантическими классами глагольных форм, выделенных ранее в [7]: характеризующие наличие конструктивного элемента (например, «содержать», «включать», «снабжать» и т. д.) и наличие связи между элементами (например, «устанавливать», «соединять», «располагать» и т. д.).

При анализе дерева разбора необходимо понимать, как именно маркируются связи между целевыми элементами дерева. Для выбранного инструментария, обсуждение которого представлено в разделе 4 данной работы, все семантические связи описаны в формате CoNLL-U[1]. В работе [18] были выделены общие связи для целей извлечения SAO на примере анализатора UdPipe [19], однако в данном случае необходимо учесть специфику формул изобретения, а именно написание в одно предложение со множеством распространенных оборотов.

После анализа примеров формул изобретений, размеченных инструментами синтаксического анализа, были выделены основные целевые связи, представленные на Рисунке 1.

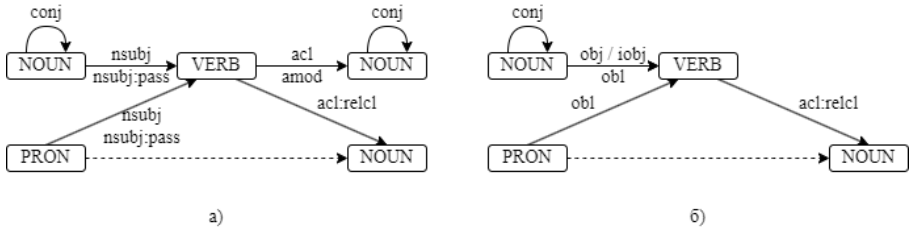

Рисунок 1 – Целевые связи актантов: а) для субъекта; б) для объекта
Figure 1 – The target links of the actants: a) for a subject; b) for an object

1 https://universaldependencies.org/

В блоках схематично выделены глагольные формы (или предикаты) VERB и их актанты – слова, заполняющие семантическую или синтаксическую валентность предиката. В общем случае актанты представлены существительными NOUN или местоимениями PRON. Направления стрелок указаны от дочернего слова к родительскому. Субъекты целевой глагольной формы (Рисунок 1а) могут быть выражены явно связями nsubj, nsubj:pass или же неявно. В последнем случае выделяются придаточные приложения – связь acl, а также модификаторы существительных – связь nmod. При этом возможны цепочки однородных существительных через сочинительную связь conj. Также в дереве разбора возможно разрешить ситуацию с указательными местоимениями (например, «который»). При наличии связи acl:relcl, указательное местоимение допустимо заменить на соответствующее ему существительное, что показано пунктирной линией на схеме.

Объекты целевой глагольной формы (Рисунок 1б) выражены либо явно через связи obj или iobj, либо с помощью связи неосновных глагольных модификаторов obl. По аналогии с субъектом, так же возможна ситуация разрешения действительного актанта по указательному местоимению с помощью связи acl:relcl. Примеры фрагментов ФИУ, поясняющие разницу связи acl:relcl в случае субъектов и объектов, представлены на Рисунке 2.

Рисунок 2 – Пример разметки со связью acl:relcl: а) для субъекта; б) для объекта
Figure 2 – Example of markup with acl:relcl link: a) for a subject; b) for an object

Общий алгоритм извлечения данных состоит из следующих этапов:

1. Предобработка текста: удаляются вводные конструкции, скобки, html-теги, номера пункта ФИУ.

2. Выделение сегментов. Предложение ФИУ разделяется на сегменты по сильным признакам для более корректной работы синтаксических анализаторов на коротких текстах. К таким признакам отнесём: а) фразу «отличающийся тем, что», разграничивающую ограничительную и отличительную части формулы; б) точку с запятой; в) фразы «при этом», «причём».

3. Поиск целевых глагольных форм в сегменте. Каждый токен (слово) в сегменте проверяется по вхождению его леммы в словарь целевых глаголов, упомянутых выше. Если целевой глагол найден, выполняется следующий этап.

4. Извлечение субъекта. В соответствие со схемой (Рисунок 1а) проверяются связи для извлечения субъекта глагольной формы, а также однородных членов. При нахождении субъекта выполняется следующий этап.

5. Извлечение объекта. В соответствие со схемой (Рисунок 1б) проверяются связи для извлечения объектов. При нахождении объекта выполняется следующий этап.

6. Формирование связок SAO. Путем комбинаций целевого глагола и всех найденных к нему субъектов и объектов формируется выходной список извлеченных SAO.

Рассмотрим процесс нахождения компонентов SAO на примере фрагмента ФИУ (см. рис. 3). Стрелки показаны инверсно относительно схемы рис. 1, т. е. в направлении зависимого слова.

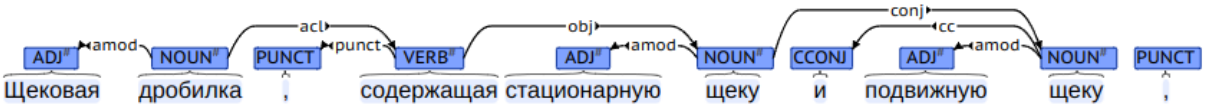

Рисунок 3 – Пример разметки сегмента для извлечения SAO
Figure 3 – Example of segment markup for SAO extraction

Фрагмент содержит целевую глагольную форму – причастие «содержащая». По схеме (Рисунок 1а) для определения субъекта имеем связь acl от глагольной формы VERB «содержащая» к субъекту в форме существительного NOUN «дробилка». Для определения объекта по схеме (Рисунок 1б) имеется связь obj к существительному «щеку», а также к однородному члену «щеку» через связь conj. Таким образом находятся вершины именных групп всех актантов. Сами именные группы собираются ниже по дереву для их актантов с учетом исключения из групп пунктуации и союзных слов в начале. В конечном итоге получим следующие SAO для фрагмента на Рисунке 3:

– «щековая дробилка» - «содержащая» - «стационарную щеку»;

– «щековая дробилка» - «содержащая» - «подвижную щеку».

Аналогичным образом определяются и другие возможные типы связей для актантов.

Вычислительный эксперимент

Для сравнительного анализа авторы ограничились следующими анализаторами русского языка:

1) UdPipe [19] (модель russian-syntagrus-ud-2.10-220711 по API);

2) Stanza [20] (версия библиотеки 1.4.0);

3) DeepPavlov [21] (пакет syntax_ru_syntagrus_bert, версия библиотеки 0.17.4);

4) spaCy [13] (модель ru_core_news_lg).

Список не претендует на всеобщность. Основными критериями выбора инструментов являлись: поддержка русского языка; доступность моделей; простота их запуска; возможность получения результата строкой в формате CoNLL-U. Указанные выше инструменты отвечают заданным критериям и имеют интерфейс на языке Python.

Для оценки качества извлечения данных вручную было размечено 20 главных пунктов формул изобретения. Патенты взяты в основном из групп B02C1/02, B01F7/28 и A21C1/06 по международной патентной классификации. В документах выделялись связки, относящиеся к компонентам изобретения и указывающие на связи между ними. Общая численность составила 252 таких SAO-структур. Данные размечались по аналогии с BIO-разметкой и сохранялись в виде json-файла (см. Рисунок 4):

```
{
    "id": 2,
    "sbj": {
        "text": "щековая дробилка",
        "bio": "I R"
    },
    "act": {
        "text": "содержащая",
        "lemma": "содержать",
        "type": "component"
    },
    "obj": {
        "text": "подвижную щеку",
        "bio": "I R"
    }
},
```


Рисунок 4 – Фрагмент разметки данных
Figure 4 – Fragment of data markup

где id – номер связки SAO в документе; sbj и obj – словари с информацией по субъекту и объекту соответственно, при этом поле text содержит полный текст именной группы в неизменном виде, а поле bio содержит маркеры для отличия частей именной группы; каждый маркер соответствует отдельному токену в именной группе, расшифровка маркеров следующая: R (Root) – вершина именной группы, I (Inside) – второстепенное слово именной группы, P (Preposition) – предлог; для глагольных форм словарь act, помимо первоначального текстового представления text, содержит поле леммы слова lemma, а так же семантический класс глагола, где зарезервировано два типа: component и link для обозначения компонентов и связей между ними соответственно.

Программный скрипт для каждого из выбранных анализаторов реализовывал алгоритм извлечения, описанный в разделе 3 настоящей работы. После извлечения связок производился подсчет метрик точности, полноты и F1-меры по аналогии с работой [7]. При этом строгая оценка так же учитывала полноту именных групп актантов, а нестрогая оценка – только корректность вершины именной группы субъекта и объекта. Извлеченные связки SAO автоматически сравнивались с ранее размеченными данными. Результат оценки работы анализаторов представлен в Таблице 1.

Таблица 1 – Результаты оценки извлечения SAO
Table 1 – Results of SAO extraction assessment

Система разметки	Строгая оценка			Нестрогая оценка		
	precision	recall	F1-score	precision	recall	F1-score
UdPipe	0,46	*0,63*	**0,53**	0,56	0,76	0,64
Stanza	0,41	0,56	0,47	0,54	0,73	0,62
DeepPavlov	0,44	*0,63*	0,52	0,56	*0,79*	**0,66**
spaCy	0,42	0,55	0,48	0,54	0,67	0,61

Обсуждение результатов

По Таблице 1 лучшее значение F1-меры при нестрогой оценке у DeepPavlov и составляет 0,66. При строгой оценке лучшее значение 0,53 у UdPipe. В целом, разброс значений F1 между всеми библиотеками не превышает 6 %. Можно заметить, что точность precision по обоим типам оценки несколько ниже полноты recall. Отчасти, это можно объяснить извлечением нецелевых связок SAO по одним и тем же формальным признакам (связям). Рассмотрим пример на Рисунке 5:

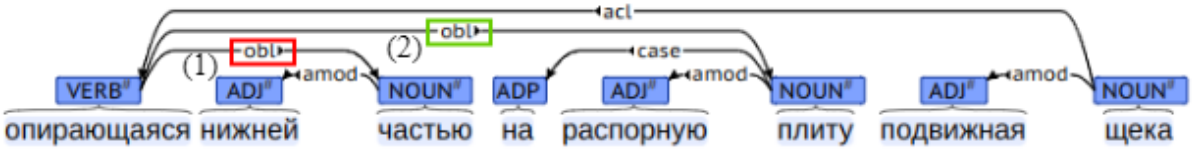

Рисунок 5 – Пример разметки с неучтенной связью
Figure 5 – Example of a mark-up with an uncounted link

Причастие «опирающаяся» семантически выражает связь между элементами, и в данном примере содержит два неосновных глагольных модификатора obl. По схеме связей (см. Рисунок 1) их можно отнести к объектам SAO. В разметку однозначно опадет модификатор (2) «плиту», так как является основной валентностью – показывает с чем связана «подвижная щека». Модификатор (1) лишь уточняет способ этой связи и не является ключевым, поэтому мог отсутствовать в валидационной разметке. Извлеченная связка SAO из примера будет следующей:

– «подвижная щека» - «опирающаяся» - «на распорную плиту».

Вопрос учета всех валентностей относится уже к степени формализации данных, связан с усложнением модели анализа дерева разбора и не совсем укладывается в формализм SAO. В данном случае основной целью было оценить общий уровень качества извлечения.

В связи с этим, для оценки качества извлечения более показательны значения полноты recall. При нестрогой оценке у DeepPavlov эта характеристика достигает значения 0,79. Таким образом, при текущих методологических ограничениях потенциально можно извлечь до 80 % морфологической информации об изобретениях на русскоязычном домене патентов.

При этом остаются следующие проблемы анализа деревьев зависимости при обработке ФИУ:

– некорректная привязка однородных членов или придаточных частей предложения к своим вершинам (проблема чрезмерной длинны ФИУ);

– некорректное определение частей речи специфичных слов и конструкций (специфика домена);

– сложности детального анализа связей (например, obl) и уточнения валентностей, т. е. нужно дополнительно проверять и отсеивать заполнители по связи obl.

Так как модели синтаксических анализаторов обучаются, как правило, на общедоступных размеченных корпусах – публицистике, художественной литературе, – то значительный прирост качества работы таких инструментов на домене патентов в ближайшем будущем маловероятен.

Заключение

В данной работе представлены результаты оценки извлечения структур SAO из русскоязычных патентов, а именно формул изобретений, с помощью анализа деревьев

зависимостей. Целевые структуры SAO описывали компоненты ТС и связи между ними. Сравнивались следующие инструменты: UdPipe, Stanza, DeepPavlov и spaCy. Разница F1-меры между синтаксическими анализаторами на тестовой выборке не превысила 6 %. По метрике recall эксперименты показали, что потенциально возможно извлечь до 80 % искомой морфологической информации об изобретениях. В то же время трудности анализа деревьев разбора и несовершенство работы инструментов на домене патентов порождают множество ошибок и избыточных данных, и среднее значение F1-меры находится в пределах 48 %-66 %. Лучшее значение F1-меры при строгой оценке у UdPipe и составило 0,53; при нестрогой оценке лучший результат показала модель от DeepPavlov со значением 0,66. Для ряда задач по обработке патентной информации такого уровня качества может быть недостаточно.

Автоматизация извлечения знаний из патентного массива, а также корректная формализация этих знаний являются одними из необходимых предпосылок для создания действительно интеллектуальных систем автоматизированного изобретательства. И сегодня, судя по интенсивности публикаций, поиск эффективных методов для реализации таких задач все еще продолжается. Подходы на основе машинного обучения уже достигли значительных успехов, однако требуют больших объемов размеченных данных. А общие инструменты обработки естественного языка – на примере синтаксических анализаторов, – хотя и представляют более простой способ извлечения данных, но в силу объективных причин не могут обеспечить передового уровня качества извлечения.

Новизна проведенного исследования заключается в апробации типового метода извлечения структур SAO по деревьям зависимостей применительно к задаче извлечения морфологических признаков ТС из пунктов формул изобретения в текстах патентов. Теоретическими результатами данной работы является разработка методологии извлечения морфологических признаков ТС из русскоязычных патентов по деревьям зависимостей. Практические результаты заключаются в получении количественной оценки качества работы наиболее известных синтаксических анализаторов русского языка на домене патентов по представленной методологии.

Дальнейшее направление исследований авторам видится в поиске путей увеличения качества извлечения морфологических признаков ТС и проработке их формализации.

СПИСОК ИСТОЧНИКОВ

1. Li X., Song H., Zhang X., Xu Q. Fine-grained Construction of Semantic Technology Network for Technology Evolution Analysis. *Proc. of the 3rd International Conference on Computer Science and Application Engineering*. 2019:1–7. DOI: 10.1145/3331453.3361638.
2. You H., Li M., Hipel K.W. et al. Development trend forecasting for coherent light generator technology based on patent citation network analysis. *Scientometrics*. 2017;111:297–315. DOI: 10.1007/s11192-017-2252-y.
3. Kim S., Yoon B. Patent infringement analysis using a text mining technique based on SAO structure. *Computers in Industry*. 2021;125:103379. DOI: 10.1016/j.compind.2020.103379.
4. Feng L., Niu Y., Wang J. Development of Morphology Analysis-Based Technology Roadmap Considering Layer Expansion Paths: Application of TRIZ and Text Mining. *Applied Sciences*. 2020;10(23):8498. DOI: 10.3390/app10238498.

5. Liu L., Li Y., Xiong Y., Cavallucci, D. A new function-based patent knowledge retrieval tool for conceptual design of innovative products. *Computers in Industry*. 2020;115:103154. DOI: 10.1016/j.compind.2019.103154.
6. Зарипова В.М., Петрова И.Ю., Цырульников Е.С. Классификация автоматизированных систем поддержки инновационных процессов на предприятии (Computer Aided Innovation – CAI). *Прикаспийский журнал: управление и высокие технологии*. 2012;1(17):26–35. Доступно по: https://elibrary.ru/download/elibrary_17708904_61173989.pdf (дата обращения: 20.10.2022).
7. Васильев С.С., Коробкин Д.М., Фоменков С.А. Метод формирования информационного обеспечения синтеза новых технических решений на основе анализа патентного массива. Часть 1. *Вестник компьютерных и информационных технологий*. 2021;18(11):3–12. DOI: 10.14489/vkit.2021.11.pp.003-012.
8. Boting G., Wenqing W. Open Relation Extraction in Patent Claims with a Hybrid Network. *Wireless Communications and Mobile Computing*. 2021;2021(1):1–7. DOI: 10.1155/2021/5547281.
9. Yang S.-Y., Soo V.-W. Extract conceptual graphs from plain texts in patent claims. *Engineering Applications of Artificial Intelligence*. 2012;25(4):874–887. DOI: 10.1016/j.engappai.2011.11.006
10. Lyashevskaya O.N., Shavrina T.O., Trofimov I.V., Vlasova N.A. Grameval 2020 Shared Task: Russian Full Morphology And Universal Dependencies Parsing. *Proc. of the International Conference «Dialogue 2020»*. 2020:553–569. DOI: 10.28995/2075-7182-2020-19-553-569.
11. Ki W., Kim K. Generating Information Relation Matrix Using Semantic Patent Mining for Technology Planning: A Case of Nano-Sensor. *IEEE Access*. 2017;5:26783–26797. DOI: 10.1109/access.2017.2771371.
12. Lin W., Liu X., Xiao R. Research on Product Core Component Acquisition Based on Patent Semantic Network. *Entropy (Basel)*. 2022;24(4):549. DOI: 10.3390/e24040549.
13. Honnibal M., Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*. 2017.
14. Yindi S., Wei L., Guozhong C., Qingjin P., Jianjie G., Jiaming F. Effective design knowledge abstraction from Chinese patents based on a meta-model of the patent design knowledge graph. *Computers in Industry*. 2022;142:103749. DOI: 10.1016/j.compind.2022.103749.
15. Krestel R., Chikkamath R., Hewel C., Risch J. A survey on deep learning for patent analysis. *World Patent Information*. 2021;65:102035. DOI: 10.1016/j.wpi.2021.102035.
16. Chen L., Xu S., Zhu L., Zhang J., Lei X., Yang G. A deep learning based method for extracting semantic information from patent documents. *Scientometrics*. 2020;125:289–312. DOI: 10.1007/s11192-020-03634-y.
17. Xueqiang L., Xiangru L., Xindong Y., Zhian D., Junmei H. Relation Extraction Toward Patent Domain Based on Keyword Strategy and Attention+BiLSTM Model (Short Paper). *Proc. of the 15th EAI International Conference, CollaborateCom*. 2019. DOI: 10.1007/978-3-030-30146-0_28.
18. Kolesnikova V., Korobkin D., Fomenkov S., Rayushkin E., Glushkin V. The Analysis of Technology Development Trends Based on the Network Semantic Structure «Subject-Action-Object». *Cyber-Physical Systems: Intelligent Models and Algorithms. Studies in Systems, Decision and Control*. 2022;417:43–53. DOI: 10.1007/978-3-030-95116-0_4.
19. Straka M., Hajič J., Straková J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proc. of the Tenth International Conference on Language Resources and Evaluation*

(LREC'16). 2016:4290–4297. Доступно по: https://aclanthology.org/L16-1680.pdf (дата обращения: 20.10.2022).
20. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Association for Computational Linguistics (ACL) System Demonstrations*. 2020. Доступно по: https://nlp.stanford.edu/pubs/qi2020stanza.pdf (дата обращения: 20.10.2022).
21. Burtsev M. et al. DeepPavlov: Open-Source Library for Dialogue Systems. *Proc. of ACL 2018, System Demonstrations*. 2018:122–127. DOI: 10.18653/v1/P18-4021.

REFERENCES

1. Li X., Song H., Zhang X., Xu Q. Fine-grained Construction of Semantic Technology Network for Technology Evolution Analysis. *Proc. of the 3rd International Conference on Computer Science and Application Engineering*. 2019:1–7. DOI: 10.1145/3331453.3361638.
2. You H., Li M., Hipel K.W. et al. Development trend forecasting for coherent light generator technology based on patent citation network analysis. *Scientometrics*. 2017;111:297–315. DOI: 10.1007/s11192-017-2252-y.
3. Kim S., Yoon B. Patent infringement analysis using a text mining technique based on SAO structure. *Computers in Industry*. 2021;125:103379. DOI: 10.1016/j.compind.2020.103379.
4. Feng L., Niu Y., Wang J. Development of Morphology Analysis-Based Technology Roadmap Considering Layer Expansion Paths: Application of TRIZ and Text Mining. *Applied Sciences*. 2020;10(23):8498. DOI: 10.3390/app10238498.
5. Liu L., Li Y., Xiong Y., Cavallucci, D. A new function-based patent knowledge retrieval tool for conceptual design of innovative products. *Computers in Industry*. 2020;115:103154. DOI: 10.1016/j.compind.2019.103154.
6. Zaripova V.M., Petrova I.Yu., Tsyrulnikov E.S. Classification of automated systems of support for innovation precesses at enterprises (Computer aided innovation – CAI). *Prikaspiiskii zhurnal: upravlenie i vysokie tekhnologii = Caspian journal management and high technologies*. 2012;1(17):26–35. (In Russ.). Available by: https://elibrary.ru/download/elibrary_17708904_18434426.pdf (accessed on: 20.10.2022).
7. Vasiliev S.S., Korobkin D.M., Fomenkov S.A. method of domain ontology automated replenishment for the support of new technical solutions synthesis. Part I. *Vestnik komp'yuternykh i informatsionnykh tekhnologii = Herald of computer and information technologies*. 2021;18(11):3–12. (In Russ.). DOI: 10.14489/vkit.2021.11.pp.003-012.
8. Boting G., Wenqing W. Open Relation Extraction in Patent Claims with a Hybrid Network. *Wireless Communications and Mobile Computing*. 2021;2021(1):1–7. DOI: 10.1155/2021/5547281.
9. Yang S.-Y., Soo V.-W. Extract conceptual graphs from plain texts in patent claims. *Engineering Applications of Artificial Intelligence*. 2012;25(4):874–887. DOI: 10.1016/j.engappai.2011.11.006.
10. Lyashevskaya O.N., Shavrina T.O., Trofimov I.V., Vlasova N.A. Grameval 2020 Shared Task: Russian Full Morphology And Universal Dependencies Parsing. *Proc. of the International Conference «Dialogue 2020»*. 2020:553–569. DOI: 10.28995/2075-7182-2020-19-553-569.
11. Ki W., Kim K. Generating Information Relation Matrix Using Semantic Patent Mining for Technology Planning: A Case of Nano-Sensor. *IEEE Access*. 2017;5:26783–26797. DOI: 10.1109/access.2017.2771371.

12. Lin W., Liu X., Xiao R. Research on Product Core Component Acquisition Based on Patent Semantic Network. *Entropy (Basel)*. 2022;24(4):549. DOI: 10.3390/e24040549.
13. Honnibal M., Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*. 2017.
14. Yindi S., Wei L., Guozhong C., Qingjin P., Jianjie G., Jiaming F. Effective design knowledge abstraction from Chinese patents based on a meta-model of the patent design knowledge graph. *Computers in Industry*. 2022;142:103749. DOI: 10.1016/j.compind.2022.103749.
15. Krestel R., Chikkamath R., Hewel C., Risch J. A survey on deep learning for patent analysis. *World Patent Information*. 2021;65:102035. DOI: 10.1016/j.wpi.2021.102035
16. Chen L., Xu S., Zhu L., Zhang J., Lei X., Yang G. A deep learning based method for extracting semantic information from patent documents. *Scientometrics*. 2020;125:289–312. DOI: 10.1007/s11192-020-03634-y.
17. Xueqiang L., Xiangru L., Xindong Y., Zhian D., Junmei H. Relation Extraction Toward Patent Domain Based on Keyword Strategy and Attention+BiLSTM Model (Short Paper). *Proc. of the 15th EAI International Conference, CollaborateCom*. 2019. DOI: 10.1007/978-3-030-30146-0_28.
18. Kolesnikova V., Korobkin D., Fomenkov S., Rayushkin E., Glushkin V. The Analysis of Technology Development Trends Based on the Network Semantic Structure «Subject-Action-Object». *Cyber-Physical Systems: Intelligent Models and Algorithms. Studies in Systems, Decision and Control*. 2022;417:43–53. DOI: 10.1007/978-3-030-95116-0_4.
19. Straka M., Hajič J., Straková J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016:4290–4297. Available by: https://aclanthology.org/L16-1680.pdf (accessed on: 20.10.2022).
20. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Association for Computational Linguistics (ACL) System Demonstrations*. 2020. Available by: https://nlp.stanford.edu/pubs/qi2020stanza.pdf (accessed on: 20.10.2022).
21. Burtsev M. et al. DeepPavlov: Open-Source Library for Dialogue Systems. *Proc. of ACL 2018, System Demonstrations*. 2018:122–127. DOI: 10.18653/v1/P18-4021.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Васильев Сергей Сергеевич, аспирант, младший научный сотрудник кафедры САПРиПК, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: svasilev2012@yandex.ru
ORCID: 0000-0001-5044-9787

Sergey Sergeevich Vasiliev, Postgraduate Student, Junior Researcher at CAD Department, Volgograd State Technical University, Volgograd, Russian Federation.

Коробкин Дмитрий Михайлович, кандидат технических наук, доцент кафедры САПРиПК, доцент, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: dkorobkin80@mail.ru
ORCID: 0000-0002-4684-1011

Dmitry Mikhailovich Korobkin, Candidate of Technical Sciences, Associate Professor at CAD Department, Volgograd State Technical University, Volgograd, Russian Federation.

Фоменков Сергей Алексеевич, доктор технических наук, профессор кафедры САПРиПК, профессор, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: saf550@yandex.ru

Sergey Alekseevich Fomenkov, Doctor of Technical Sciences, Professor at CAD Department, Volgograd State Technical University, Volgograd, Russian Federation.

Статья поступила в редакцию 20.10.2022; одобрена после рецензирования 15.11.2022; принята к публикации 25.11.2022.

The article was submitted 20.10.2022; approved after reviewing 15.11.2022; accepted for publication 25.11.2022.