DOI: 10.15514/ISPRAS-2025-37(5)-16

CC BY

Набор табличных данных RF-200 и тестирование производительности извлечения фактов из русскоязычных таблиц

Н.О. Дородных, ORCID: 0000-0001-7794-4462 <nikidorny@icc.ru>
А.Ю. Юрин, ORCID: 0000-0001-9089-5730 <iskander@icc.ru>
Институт динамики систем и теории управления имени В.М. Матросова СО РАН, Россия, 664033, г. Иркутск, ул. Лермонтова, д. 134.

Аннотация. В настоящее время огромное количество данных представлено в виде таблиц. Они повсеместно используются при решении различных практических задач в разных областях. Для семантической интерпретации (аннотирования) таблиц и построения на их основе графов знаний разрабатывается специализированное методологическое и программное обеспечение. Эффективное тестирование подобного обеспечения требует создания и использования русскоязычных наборов данных. В данной статье предложен русскоязычный набор табличных данных RF-200, содержащий 200 таблиц из 26 предметных областей, размеченных с использованием платформы Talisman. Приведены результаты тестирования производительности авторского подхода к извлечению фактов из русскоязычных таблиц с использованием RF-200, при которых F-мера достигла значения 0.464, превзойдя традиционные методы извлечения фактов из текстов (F1 = 0.277). Результаты подчеркивают важность специализированных решений для работы со структурированными данными, особенно для русскоязычных источников. Практическая значимость работы заключается в интеграции подхода в платформу Talisman, что расширяет возможности семантической аналитики, проводимой по таблицам. Исследование вносит вклад в автоматизацию обработки таблиц, решая проблему семантической интерпретации в условиях лингвистического разнообразия, и открывает перспективы для интеграции методов глубокого обучения и масштабирования созданного набора данных.

Ключевые слова: граф знаний; разработка графов знаний; пополнение графов знаний; таблица; русскоязычный набор табличных данных; извлечение фактов; тестирование производительности.

Для цитирования: Дородных Н.О., Юрин А.Ю. Набор табличных данных RF-200 и тестирование производительности извлечения фактов из русскоязычных таблиц. Труды ИСП РАН, том 37, вып. 5, 2025 г., стр. 205–224. DOI: 10.15514/ISPRAS–2025–37(5)–16.

Благодарности: Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 1023110300006-9).

Testing the Performance of Fact Extraction from Russian-Language Tables

N.O. Dorodnykh, ORCID: 0000-0001-7794-4462 <nikidorny@icc.ru>
A.Yu. Yurin, ORCID: 0000-0001-9089-5730 <iskander@icc.ru>

Matrosov Institute for System Dynamics and Control Theory of the Russian Academy of Sciences, 134, Lermontov st., Irkutsk, 664033, Russia.

Abstract. Currently, a huge amount of data is presented in the form of tables. They are widely used to solve various practical problems in different domains. Specialized methods and software are developed for semantic interpretation (annotation) of tables and construction of knowledge graphs based on them. Effective testing of such software requires the creation and use of Russian-language datasets. This paper proposes a Russian-language tabular dataset, called RF-200, containing 200 tables from 26 domains labeled using the Talisman platform. The results of testing the performance of our approach for fact extraction from Russian-language tables using RF-200 are presented, in which the F1 reached a value of 0.464, surpassing traditional methods of fact extraction from texts (F1 = 0.277). The results emphasize the importance of specialized solutions for working with structured data, especially for Russian-language sources. The practical significance of the work lies in the integration of the approach into the Talisman platform, which expands the capabilities of semantic analytics carried out on tables. The study contributes to the automation of table processing, solving the problem of semantic interpretation in the context of linguistic diversity, and opens up prospects for the integration of deep learning methods and scaling of the created dataset.

Keywords: knowledge graph; knowledge graph engineering; knowledge graph population; table; Russian-language tabular dataset; fact extraction; performance testing.

For citation: Dorodnykh N.O., Yurin A.Yu. Testing the Performance of Fact Extraction from Russian-Language Tables. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 5, 2025, pp. 205-224 (in Russian). DOI: 10.15514/ISPRAS-2025-37(5)-16.

Acknowledgements. This work was supported by the state assignment of Ministry of Science and Higher Education of the Russian Federation (theme No. 1023110300006-9).

1. Введение

В настоящее время разработка интеллектуальных систем, ориентированных на обработку больших объемов данных для поддержки принятия решений в различных предметных областях, в том числе в условиях нечёткости и неопределенности, является актуальной задачей. Такие системы, интегрирующие методы искусственного интеллекта (обработки естественного языка, машинного обучения и инженерии знаний), находят применение не только в классических областях, таких как корпоративный поиск (например, Apache Solr, Amazon Kendra, Elasticsearch) или конкурентная разведка (например, Виток-OSINT, Babel X), но и в инновационных сферах – от предиктивной аналитики в здравоохранении и до оптимизации цепочек поставок с использованием концепции Интернета вещей (Internet of Things). Например, платформы типа Siemens MindSphere и GE Predix используют семантические модели для интерпретации данных промышленных датчиков, а системы типа Bloomberg Terminal трансформируют финансовые таблицы в прогнозные модели. Ключевым элементом подобных решений выступают графы знаний (knowledge graphs) – динамические структуры, представляющие информацию в виде сети взаимосвязанных узлов-сущностей (например, «*Илон Маск*», «*Tesla*») и их рёбер-отношений (например, «*основал*», «*производит*»), формируя семантическую модель данных [1-2]. В отличие от традиционных реляционных баз данных, графы знаний поддерживают ассоциативный поиск и выявление косвенных зависимостей (например, как два учёных из разных областей связаны через общие проекты), интеграцию разнородных данных и логический вывод, а также семантическую совместимость с форматами Linked Open Data (OWL и RDF) [3], что подтверждается

успешными открытыми глобальными проектами, такими как DBpedia [4] и Wikidata [5]. Кроме того, графы знаний могут быть масштабированы до любого размера, что позволяет эффективно использовать их при обработке и анализе данных больших объемов. Однако создание графов знаний и наполнение их конкретными сущностями (фактами), особенно для узкоспециализированных предметных областей, например, фармакогеномики или патентного права [1, 6], является сложной и трудоемкой задачей, требующей разработки специализированного программного и методологического обеспечения, автоматизирующие этот процесс.

В данном контексте актуальным является автоматизация наполнения графов при помощи обработки и анализа неструктурированных и слабоструктурированных источников, среди которых особый интерес представляют таблицы. В настоящее время таблицы являются удобным и достаточно распространённым способом представления и хранения данных. Так, по оценкам экспертов таблицы в формате Google Sheets используют около 2 миллиардов пользователей ежемесячно, в то время как Microsoft Excel имеет, по оценкам, от 750 миллионов до 1,2 миллиарда ежемесячных пользователей по всему миру [7]. Большинство корпоративных хранилищ содержат данные преимущественно в табличных форматах, таких как XLSX, CSV, HTML, а также PDF (отсканированные таблицы). Кроме того, согласно недавним исследованиям [8], примерно до 40% всех таблиц расположенных в Вебе, обладают реляционной природой и содержат потенциально полезные факты, которые могут быть использованы для формирования графов знаний. Однако таблицы, будучи формально структурированными, обычно лишены явной семантики. В частности, заголовок столбца «*2023*» может обозначать как год, так и номер проекта, а ячейка со значением «*Apple*» требует дискурсивного анализа для дифференциации компании или фрукта. Проблема усугубляется различиями, которые присущи различным областям. Так, финансовые отчеты часто используют иерархические заголовки и матричную компоновку таблицы, научные статьи – многоуровневые сноски в заголовках, а веб-таблицы часто содержат объединённые ячейки и другие HTML-теги. Существующие проприетарные решения типа Talend [9], Trifacta [10] и Microsoft Semantic Link [11] в основном полагаются на более простые методы, такие как синтаксический анализ и сопоставление регулярных выражений для обнаружения ограниченного набора семантических типов, что приводит к ошибкам интерпретации. Поэтому разработка новых методов, моделей и программных средств, позволяющих производить семантическую интерпретацию таблиц и извлекать из аннотированных табличных данных конкретные сущности, их характеристики и связи, является перспективной областью научных исследований.

Данная работа является продолжением проекта по разработке методологического и программного обеспечения для автоматической семантической интерпретации (аннотирования) таблиц и извлечения новых фактов из аннотированных табличных данных с последующим пополнением предметно-ориентированных графов знаний в рамках платформы Talisman [12], в частности, рассматривается задача тестирования производительности подхода при извлечении фактов из русскоязычных таблиц с использованием нового набора данных (*benchmark*) – RF-200. Этапы подхода, программный инструментарий и другие детали подробно представлены в работе [13], где также рассмотрен демонстрационный пример формирования предметных графов знаний на основе табличных данных.

Основной вклад данной работы заключается в следующем:

- Впервые опубликован новый набор данных RF-200 (*ru-facts-200*), содержащий таблицы на русском языке для решения задачи извлечения новых фактов из русскоязычных таблиц (*fact extraction*). Таблицы были отобраны из корпуса таблиц Russian Web Tables (RWT) [14] и размечены с использованием средств платформы Talisman. Полученный набор данных обеспечивает основу для разработки

мультиязычных решений, способных обрабатывать информацию с различными лингвистическими особенностями. Более того, кодовая база созданного набора опубликована для свободного использования.

- Прирост экспериментальной оценки (F-меры) производительности предлагаемого подхода к аннотированию таблиц и извлечению новых фактов из аннотированных табличных данных на основе созданного набора данных RF-200 составил 0,187 относительно базового решения (*baseline*) извлечение фактов из текстов.
- Интеграция решения в промышленную платформу Talisman, что подтверждает прикладную значимость исследования.

Статья организована следующим образом: в разделе 2 представлено современное состояние исследований и существующие наборы табличных данных. В разделе 3 кратко описывается предложенный ранее подход к семантическому аннотированию таблиц. В разделе 4 описывается процесс создания нового набора русскоязычных табличных данных RF-200. Раздел 5 представляет результаты тестирования производительности авторского подхода с использованием созданного набора данных. В заключении дается обсуждение полученных результатов и планы будущей работы.

2. Современное состояние исследований

В силу большого распространения табличных данных в последнее время все больше исследователей обращают внимание на проблематику создания (*knowledge graph construction* [15-16]), пополнения (*knowledge graph population* [17-18]) и расширения графов знаний (*knowledge graph refinement* [19]) за счет этой информации.

Недостатки существующих решений обуславливают необходимость разработки новых методов, сочетающих автоматизированную обработку слабоструктурированных таблиц с интуитивными интерфейсами и средствами семантической верификации. Несмотря на определенные успехи, область остаётся фрагментарной: отсутствует универсальная методология, способная обеспечить комплексную интерпретацию разнородных таблиц. Согласно экспериментальным данным последнего соревнования SemTab-2024 (*Semantic Web Challenge on Tabular Data to Knowledge Graph Matching*) [20], современные системы демонстрируют неудовлетворительную точность при работе с реальными данными. Более того, существующие решения, как правило, не предоставляют возможности дальнейшего использования семантически аннотированных таблиц, например, пропуская этап формирования графов знаний. Указанные ограничения подчеркивают необходимость создания интегрированных платформ, способных не только генерировать предметные графы знаний из таблиц, но и динамически расширять существующие семантические модели новыми фактами. Приоритетными направлениями остаются разработка кросс-доменных алгоритмов, внедрение пользовательских графических интерфейсов для экспертов-непрограммистов и обеспечение открытости инструментария.

Для тестирования производительности подходов к автоматическому пониманию табличной информации (*table understanding*) [21-22] используются наборы данных (*benchmarks*), называемые также «*золотыми стандартами*» (*gold standards*), которые служат эталоном для измерения производительности (качества) различных методов и систем. Они позволяют выявлять сильные и слабые стороны существующих методов, тем самым помогая в продвижении производительности на современном уровне. Большинство доступных наборов данных для табличных задач охватывают широкий диапазон предметных областей, в основном за счет того, что создаются с использованием крупномасштабных открытых веб-ресурсов, таких как Wikipedia или GitHub, и только некоторые из них нацелены на определенную конкретную предметную область (например, медицину, финансы, промышленность).

Наборы данных для задачи семантической интерпретацией таблиц представлены в табл. 1. Указанные наборы содержат таблицы, размеченные семантическими типами (классами, характеристиками и связями между классами), взятых из различных графов знаний общего назначения, для оценки качества семантического аннотирования отдельных элементов таблиц. По полученным аннотациям из ячеек таблиц могут быть извлечены факты, однако эти наборы напрямую не направленны на задачу извлечения фактов и не предоставляют каких-либо метрик оценки для этого. Следует также отметить, что таблицы в этих наборах представлены на английском языке, исключение составляет только RWT-RuTaBERT, содержащий коллекцию русскоязычных размеченных таблиц.

Табл. 1. Статистика по наиболее распространенным наборам данных для задачи семантической интерпретации таблиц.
Table 1. Statistics on the most common datasets for the task of semantic table interpretation.

Набор данных	**Кол-во таблиц**	**Кол-во столбцов**	**Кол-во строк**	**Кол-во семантических типов**	**Граф знаний**
Limaye [23]	6,5 тыс.	–	–	837	Wikidata, Yago
T2Dv2 [24]	234	1,2 тыс.	2,8 тыс.	193	DBpedia
Tough Table (2T) [25]	180	194 тыс.	802 тыс.	540	DBpedia, Wikidata
BiodivTab [26]	50	1,2 тыс.	12,9 тыс.	84	Wikidata
GitTables [27]	962 тыс.	11,5 млн.	13,6 млн.	2,4 тыс.	Schema.org, DBpedia
SOTAB [28]	108 тыс.	–	–	267	Schema.org
VizNet-Sato [29]	–	120,6 тыс.	–	78	DBpedia
WikiTabels-TURL [30]	–	628,2 тыс.	–	225	Freebase
RWT-RuTaBERT [31]	–	1,4 млн.	–	170	DBpedia

Тем не менее, существует небольшой ряд примеров наборов данных, ориентированных на задачу извлечения фактов (сущностей) из таблиц, в частности:

- **SWDE (Structured Web Data Extraction)** [32] – структурированный набор данных, извлеченных из 128 000 HTML-страниц с 80 веб-сайтов. Собранные записи распределены по восьми категориям: «*автомобили*», «*книги*», «*камеры*», «*работа*», «*фильмы*», «*игроки Национальная Баскетбольная Ассоциация (НБА)*», «*рестораны*» и «*университеты*». Для каждой категории задано от 3 до 5 атрибутов (например, для категории «книга» это будет «название», «автор», «ISBN-код», «издатель» и «*дата публикации*»), которые можно сопоставить столбцам данных. При этом количество строк соответствует количеству страниц (сущностей, например, конкретных книг). Набор данных содержит разметку (*ground truth*), созданную с помощью регулярных выражений для определенных атрибутов.
- **DISCOMAT** [33] – содержит 5 883 таблиц в формате CSV, извлеченных из 2 536 научных статей по материаловедению из баз Interglad, SciGlass и Elsevier. При этом

1 475 таблиц были размечены вручную, остальные автоматически. Набор включает четыре основных семантических типа сущностей: «*материал*», «*компонент*», «*процент*» и «*единица измерения*».

- **arXiv Machine Learning Tables** [34] – содержит 122 таблицы в формате LaTeX, извлеченных из 25 научных статей на arXiv по тематике машинного обучения. Набор включает 3 792 аннотированных записей, принадлежащих одиннадцати типам (например, «*метрика*», «*задача*», «*обучающие данные*»).
- **PubMed Chemistry Tables** [34] – содержит 26 таблиц в формате XML, извлеченных из 16 научных статей на PubMed по тематике физических свойств химических соединений. Набор включает записи, принадлежащие трем основным типам: «*единицы измерения*», «*исследуемое соединение*», и «*биологический объект*».

Основная статистика по данным наборам таблиц представлена в табл. 2.

Рассмотренные наборы в основном охватывают относительно простые варианты таблиц, обладающих реляционной природой. Как правило, они содержат небольшое количество семантических типов, относящиеся к какой-то конкретной области или небольшому набору областей. Кроме того, данные в этих таблицах представлены исключительно на английском языке, что ограничивает применение методов для других языков, включая русский. Таким образом, создание новых мультиязычных наборов данных для задачи извлечения фактов из таблиц, относящихся к разнообразным предметным областям и обладающих сложной структурной компоновкой, является актуальным.

Табл. 2. Статистика наборов данных для задачи извлечения фактов из таблиц.
Table 2. Statistics of datasets for the task of fact extraction from tables.

Набор данных	**Кол-во таблиц**	**Кол-во записей**	**Формат**	**Кол-во категорий (типов)**
SWDE	*128 000*	–	HTML	8 категорий: «автомобили», «книги», «камеры», «работа», «фильмы», «игроки Национальная Баскетбольная Ассоциация (НБА)», «рестораны» и «университеты»
DISCOMAT	*5 883*	–	CSV	4 типа: «материал», «компонент», «процент» и «единица измерения»
arXiv Machine Learning Tables	*122*	3 792	LaTeX	11 типов: «метрика», «задача», «обучающие данные» и др.
PubMed Chemistry Tables	*26*	1 498	XML	3 типа: «единицы измерения», «исследуемое соединение», и «биологический объект»

3. Существующий задел

Разработанный авторами подход реализует семантическое аннотирование колонок и отношений между ними, которое заключается в сопоставлении колонкам релевантных *типов характеристик*, определении наиболее подходящего *типа концепта* на их основе, а также выявление *типов связей* между определенными типами концептов. После установления подобной аннотации из строк таблиц последовательно извлекаются новые факты и добавляются в целевой граф знаний. Обобщенная схема подхода приводится на рис. 1.

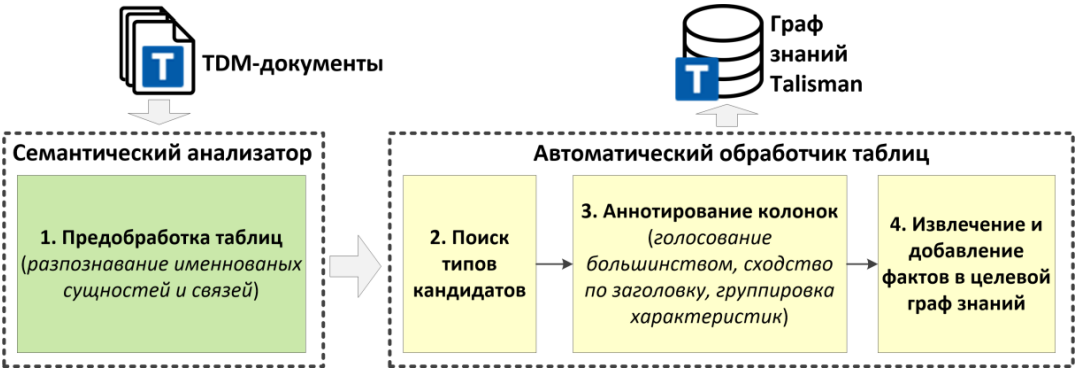

Рис. 1. Обобщенная схема подхода.
Fig. 1. The scheme of the approach.

Подход состоит из четырех основных этапов:

1) **Предобработка таблиц:** Модель XLM-RoBERTa, дообученная на корпусах CoNLL-2003, OntoNotes и DocRED, выполняет распознавание именованных сущностей (персоны, организации, локации и др.) в ячейках таблицы. На основе NER-меток извлекаются базовые факты (*текстовые упоминания* и *значения характеристик*). Важно отметить, что данная модель по умолчанию доступна в форме специального семантического анализатора, входящего в платформу Talisman. Описание гиперпараметров и другие детали модели представлены в работе [35].
2) **Поиск типов кандидатов:** Для каждой колонки таблицы определяется набор возможных *типов характеристик* из *KG*, исключая колонки без извлеченных базовых фактов.
3) **Аннотирование колонок:** Релевантный тип для колонки выбирается с помощью агрегированной оценки, полученной на основе применения комбинации трех эвристических метода (*голосование большинством, сходство по заголовку, группировка характеристик*). Данная оценка определяет итоговую вероятность того, что определенный тип характеристики из набора кандидатов является наиболее подходящим (релевантным) для аннотирования столбца таблицы. Агрегирование осуществляется на основе линейной свертки оценок, полученных каждой эвристикой: $f_{agg}(c_i) = f_1(c_i) \times w_1 + f_2(c_i) \times w_2 + f_3(c_i) \times w_3$, где c_i – целевой столбец для аннотирования; f_1, f_2, f_3 – эвристики аннотирования столбца; w_1, w_2, w_3 – весовые коэффициенты, которые уравновешивают важность каждой оценки.
4) **Извлечение и добавление фактов в целевой граф знаний:** На основе аннотаций извлекаются *концепты*, их *характеристики* и *связи*, пополняя граф знаний Talisman. При этом факты связей формируются только в пределах одной строки.

Разработанный подход реализован в форме специального автоматического *обработчика таблиц* (*tables-annotator*), который использовался в рамках исследовательского проекта Института системного программирования имени Иванникова Российской академии наук (ИСП РАН). В рамках этого проекта решалась задача автоматизированного наполнения предметно-ориентированных графов знаний платформы Talisman [12] новыми фактами, извлеченными, в том числе, из табличных данных. Подробная информация по подходу представлена в работе [13].

4. Набор данных RF-200

Структурная компоновка таблиц и большой охват разнообразных предметных областей важны для создания качественного набора табличных данных. В данной работе использовался крупномасштабный корпус табличных данных – Russian Web Tables (RWT) [14], который был сформирован на основе среза русскоязычной Википедии за 13 сентября 2021 года. RWT представлен набором файлов в формате CSV, содержащих непосредственно таблицы, а также файлы в формате JSON, содержащие метаданные о таблицах. В табл. 3 описывается основная статистика корпуса RWT.

Табл. 3. Статистика корпуса таблиц RWT.
Table 3. The RWT corpus statistics.

Статистика	**Значение**
Количество таблиц	1 266 731
Количество колонок	7 419 771
Количество ячеек	99 638 194
Среднее число ячеек на таблицу	81,78
Размер набора	17 ГБ
Процент практически пустых колонок	6%
Среднее число ячеек в колонке	13,42
Процент колонок содержащих только числовые данные	17%

Из корпуса RWT было отобрано 225 исходных вертикальных таблиц, в которых данные содержаться в форме столбцов (вертикальных колонок), на основе метаданных корпуса. Собранные таблицы принадлежат 26 разным предметным областям и содержат как простые заголовки (заголовок первой строкой), так и сложные иерархические заголовки, которые могут располагаться в произвольном порядке внутри таблицы. Данные в таблицах не были очищены и могут содержать незначительные опечатки и мусорные символы, которые потенциально могут вносить сложность в процесс обработки и анализа этих таблиц.

Разметка данных таблиц осуществлялась в автоматизированном режиме средствами платформы Talisman [12]. В частности, была разработана модель предметной области (*OntoScheme*), отражающая основные понятия, их характеристики и отношения между ними для всех 26 областей. Статистика по созданной модели представлена в табл. 4.

Пример фрагмента модели предметной области, описывающий область музыки (данные по чартам, синглам, информация об исполнителях и музыкальных группах и т.п.), представлен на рис. 2.

Затем применялся специальный обработчик извлечения фактов, к которому вручную определялась конфигурация в формате JSONPath отдельно для каждой таблицы. Данная конфигурация представляет собой набор инструкций, использующий механизм регулярных выражений и созданную модель предметной области, по которым происходило извлечение фактов из таблиц. Извлеченные таким образом факты составили эталонные данные разметки (*ground truth*). В результате было размечено 200 таблиц. Статистика по собранным и размеченным таблицам представлена в табл. 5.

В результате был создан новый набор данных – RF-200 (*ru-fats-200*), содержащий размеченные русскоязычные таблицы. Основная статистика по набору RF-200 представлена

в табл. 6. При этом среднее количество колонок на одну таблицу составило 4,89. Среднее количество ячеек на 1 таблицу составило 97,45, а доля пустых ячеек 8,82%.

Табл. 4. Статистика по созданной модели предметной области.
Table 4. Statistics on the created domain model.

Элемент модели	**Количество**
Тип концепта	29
Тип значения характеристики	9
Тип характеристики концепта	281
Тип связи	50
Тип характеристики связи	22
Всего элементов	**391**

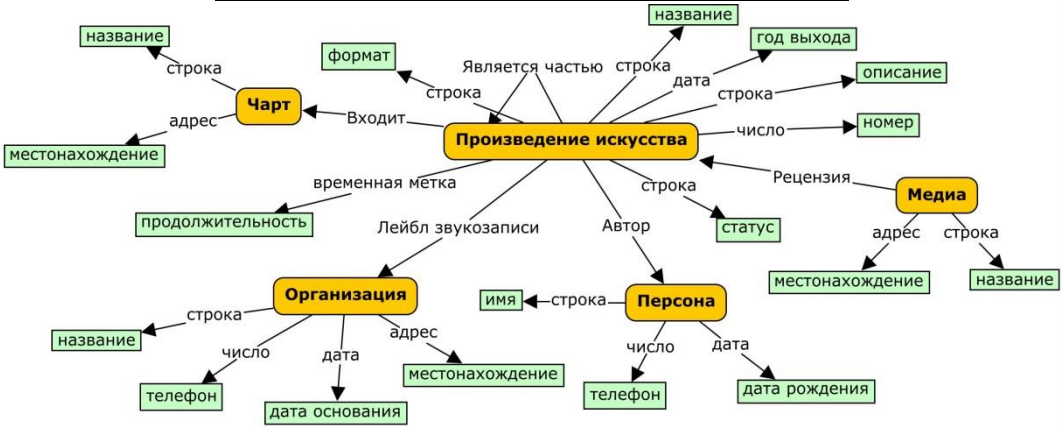

Рис. 2. Пример фрагмента модели предметной области, описывающий область «музыка».
Fig. 2. An example of a domain model fragment ("music").

4. Тестирование производительности

4.1 Настройки и метрики

Тестирование производительности авторского подхода и его программной реализации в форме обработчика платформы Talisman осуществлялась на основе подготовленного набора данных RF-200. В качестве базового решения (*baseline*) для сравнения был выбран классический подход извлечения фактов из текстов, основанный на распознавании именованных сущностей (*Named Entity Recognition*) и извлечении отношений между ними (*Relation Extraction*). Этот подход также реализован в форме специального обработчика – *семантического анализатора* (*semantic analyzer*) в платформе Talisman. Следует отметить, что провести корректное сравнение с другими внешними "*state-of-the-art*" решениями достаточно сложно, так как они направлены на обработку только определенного набора данных, обладающих собственной спецификой (форматом, поддерживаемых категорий типов и др.).

Основная проверяемая гипотеза заключается в ответе на следующие вопросы: «*Пригодны ли классические методы извлечения информации из текстов для таблиц?*» и «*Требуют ли таблицы создания специализированных решений?*».

Табл. 5. Статистика по собранным и размеченным таблицам.
Table 5. Statistics on the collected and labeled tables.

№	Предметная область	Краткое описание	Кол-во отобранных таблиц	Кол-во размеченных таблиц
1	*Локации*	Статистика по странам, отдельным субъектам, городам (население,	34	33
2	*Спорт*	Команды, игроки, виды спорта	29	26
3	*Киноиндустрия и театры*	Фильмы, сериалы, аниме, театральные постановки, актеры	18	16
4	*Политика*	Партии, депутаты, политики, лидеры	16	15
5	*Кинонаграды*	Кинопремии, победители, номинации и	11	8
6	*История*	События, личности и военная статистика	9	9
7	*Музыка*	Песни, синглы, певцы, группы	9	6
8	*Автоспорт*	Ралли, команды, турниры	8	8
9	*Архитектурные*	Строения как старые, так и новые	8	8
10	*Торговля и финансы*	ВВП, кредиты, импорт, экспорт	7	3
11	*Энергетика*	Показатели энергетики, мощности	7	2
12	*Печатные издания*	Книги, манга, журналы	7	7
13	*Измерения*	Эталонные и рекордные измерения различных показателей	6	6
14	*Праздники и мероприятия*	Названия праздников, периоды празднования	6	6
15	*Природные объекты*	Статистика по рекам и озерам	6	6
16	*Авиация*	Самолеты, вертолеты	5	5
17	*Медиа*	Радио, телевидение	5	5
18	*Организации и объединения*	Данные по различным организациям и объединениям	5	5
19	*Продукты питания*	Статистика по составу продуктов питания	4	4
20	*Астрономия*	Астрономические аппараты, звёздные	4	4
21	*Национальности и этносы*	Статистические данные по различным национальностям, этносам и	4	4
22	*Религия*	Статистические данные по различным конфессиям и религиозным течениям	4	4
23	*Телешоу*	КВН, стендап, команды	4	1
24	*Награды и премии*	Статистические данные по различным номинациям, премиям и награжденным	3	3
25	*Реслинг*	Соревнования по реслингу, статистика по победам рестлеров	3	3
26	*Компьютерные игры*	Игры, платформы, игровые издания, киберспорт	3	3
ИТОГО			**225**	**200**

Табл. 6. Статистика по размеченному набору табличных данных RF-200.
Table 6. Statistics on the labeled tabular dataset (RF-200).

Предметная область	**Кол-во таблиц**	**Кол-во колонок**	**Кол-во ячеек**	**Кол-во пустых ячеек**
Локации	33	153	1992	99
Спорт	26	188	4786	614
Киноиндустрия и театры	16	74	1464	44
Политика	15	68	915	81
История	9	33	877	156
Кинонаграды	8	42	588	3
Автоспорт	8	45	431	10
Архитектурные сооружения	8	47	753	124
Печатные издания	7	29	554	48
Музыка	6	21	618	6
Измерения	6	23	248	11
Праздники и мероприятия	6	20	263	12
Природные объекты	6	26	359	21
Авиация	5	28	499	109
Медиа	5	16	221	17
Организации и объединения	5	37	640	55
Продукты питания	4	14	194	0
Астрономия	4	16	800	0
Национальности и этносы	4	12	243	10
Религия	4	11	102	0
Торговля и финансы	3	12	280	0
Награды и премии	3	13	342	52
Реслинг	3	14	472	41
Компьютерные игры	3	14	906	143
Энергетика	2	16	880	63
Телешоу	1	7	63	0
ИТОГО	**200**	**979**	**19490**	**1719**

Интуитивно, метрика оценки должна вычислять разницу между количеством истинных (размеченных) фактов, находящихся в таблице набора RF-200 и количеством извлеченных фактов обработчиком таблиц и семантическим анализатором. Таким образом, экспериментальная оценка была получена отдельно для двух этапов:

1) извлечение *фактов-концептов*, *фактов-значений* и *фактов-упоминаний* (будем обозначать этот этап как «NERC»);
2) извлечение *фактов-характеристик концептов*, *фактов-связей* и *фактов-характеристик связей* (будем обозначать этот этап как «RELEXT»).

В качестве метрик оценки для обоих этапов извлечения фактов использовались стандартные: точность (*precision*), полнота (*recall*) и F-мера (*F1 score*):

$$Precision = \frac{CF}{EF}, \qquad Recall = \frac{CF}{NF}, \qquad F1 = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

где *CF* – количество правильно (т.е. совпадающих с истинными) извлеченных фактов из таблицы обработчиком; *EF* – количество фактов в целом, извлеченных обработчиком из таблицы; *NF* – общее количество фактов, содержащиеся в таблице набора RF-200.

Таким образом, данные метрики считались для каждой таблицы и потом суммировались для всего набора.

4.2 Результаты

Итоговые результаты тестирования производительности извлечения фактов из таблиц на наборе данных RF-200 приведены в табл. 7.

Табл. 7. Результаты экспериментальной оценки на наборе RF-200.
Table 7. The results of experimental evaluation on the RF-200 dataset.

Этап	Семантический анализатор			Обработчик таблиц		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NERC	0,668	0,542	0,554	0,659	0,641	0,623
RELEXT	0,000	0,000	0,000	0,377	0,281	0,306
NERC + RELEXT	0,334	0,271	0,277	0,518	0,461	0,464

Экспериментальная оценка по отдельным предметным областям для NERC-этапа приведена в табл. 8, а для RELEXT-этапа приведена в табл. 9. Далее обсудим ключевые выводы по полученной оценке.

Выводы и замечания по полученной оценки производительности:

- Оценка производилась отдельно по каждой таблице с использованием только определенной части модели предметной области (подмножества типов).
- Точность NERC-этапа оказалась немного выше для семантического анализатора. Это связано с тем, что семантический анализатор может точно выделять необходимые значения в ячейках. В то время как обработчик таблиц всегда выделяет значения ячеек целиком.
- Оценка полноты NERC-этапа оказалась выше для обработчика таблиц за счет того, что семантический анализатор может пропускать некоторые значения ячеек в колонке, особенно если они принадлежат к редким NERC-меткам (например, это могут быть редко встречаемые события, мероприятия, механические системы и т.п.). В то время как обработчик таблиц всегда выделяет все значения ячеек в колонке.

Табл. 8. Экспериментальная оценка по отдельным предметным областям для этапа NERC.
Table 8. The experimental evaluation of selected domains for the NERC stage.

Предметная область	**Семантический анализатор**			**Обработчик таблиц**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Национальности и этносы	0,950	0,854	0,883	0,998	0,937	0,963
Продукты питания	0,815	0,361	0,495	0,845	0,875	0,859
Политика	0,718	0,727	0,708	0,853	0,808	0,812
Медиа	0,703	0,737	0,711	0,768	0,742	0,754
Музыка	0,844	0,632	0,621	0,827	0,792	0,751
Природные объекты	0,819	0,506	0,555	0,893	0,677	0,737
Реслинг	0,767	0,793	0,767	0,737	0,727	0,715
Локации	0,688	0,637	0,602	0,715	0,748	0,707
Праздники и мероприятия	0,625	0,782	0,688	0,649	0,798	0,707
Автоспорт	0,676	0,681	0,657	0,684	0,748	0,700
Спорт	0,580	0,678	0,601	0,547	0,699	0,590
Киноиндустрия и театры	0,669	0,481	0,501	0,561	0,576	0,551
Организации и объединения	0,570	0,617	0,562	0,655	0,779	0,699
Компьютерные игры	0,818	0,596	0,660	0,720	0,727	0,693
История	0,734	0,691	0,699	0,677	0,691	0,673
Энергетика	0,749	0,546	0,607	0,831	0,546	0,647
Печатные издания	0,647	0,662	0,634	0,592	0,727	0,642
Авиация	0,485	0,581	0,499	0,516	0,696	0,580
Торговля и финансы	0,727	0,309	0,414	0,667	0,467	0,524
Религия	0,502	0,406	0,426	0,567	0,505	0,520
Архитектурные сооружения	0,621	0,450	0,477	0,565	0,577	0,513
Награды и премии	0,873	0,373	0,491	0,642	0,389	0,481
Кинонаграды	0,569	0,389	0,415	0,406	0,514	0,428
Телешоу	0,455	0,234	0,309	0,469	0,359	0,407
Астрономия	0,596	0,195	0,258	0,608	0,370	0,392
Измерения	0,187	0,167	0,167	0,134	0,186	0,147
По всем областям	**0,668**	**0,542**	**0,554**	**0,659**	**0,641**	**0,623**

Табл. 9. Экспериментальная оценка по отдельным предметным областям для этапа RELEXT.
Table 9. The experimental evaluation of selected domains for the RELEXT stage.

Предметная область	**Обработчик таблиц**		
	Precision	**Recall**	**F1**
Национальности и этносы	0,750	0,628	0,669
Продукты питания	0,667	0,667	0,667
Политика	0,833	0,526	0,609
Медиа	0,784	0,438	0,534
Музыка	0,627	0,378	0,466
Природные объекты	0,427	0,430	0,422
Реслинг	0,461	0,380	0,407
Локации	0,525	0,297	0,360
Праздники и мероприятия	0,392	0,313	0,342
Автоспорт	0,500	0,250	0,333
Спорт	0,410	0,294	0,325
Киноиндустрия и театры	0,513	0,248	0,321
Организации и объединения	0,312	0,328	0,319
Компьютерные игры	0,448	0,266	0,304
История	0,376	0,271	0,302
Энергетика	0,279	0,312	0,285
Печатные издания	0,198	0,350	0,252
Авиация	0,335	0,205	0,230
Торговля и финансы	0,308	0,154	0,205
Религия	0,197	0,206	0,201
Архитектурные сооружения	0,251	0,165	0,193
Награды и премии	0,111	0,111	0,111
Кинонаграды	0,091	0,091	0,091
Телешоу	0,000	0,000	0,000
Астрономия	0,000	0,000	0,000
Измерения	0,000	0,000	0,000
По всем областям	**0,377**	**0,281**	**0,306**

- Экспериментальные оценки семантического анализатора для этапа RELEXT оказались нулевыми из-за того, что данный обработчик может выделять связи и характеристики только внутри одного текста (ячейки). В то время как обработчик таблиц может выделять связи и характеристики между значениями ячеек разных колонок. Данная оценка наглядно показывает, что классический подход извлечения фактов из текстов слабо применим к табличным данным.
- В целом итоговые оценки (этапы NERC + RELEXT) для предлагаемого подхода (обработчика таблиц) оказалась ожидаемо выше, чем для семантического анализатора за счет своей направленности на обработку таблиц.

Основные причины (проблемы), повлиявшие на не высокую оценку производительности:

- Наличие опечаток и «мусорных» тегов HTML в некоторых значениях ячеек.
- Предлагаемый подход включает этап предварительной обработки таблиц, который основан на результатах распознавания именованных сущностей и извлечении связи (рис. 1). Таким образом, работа обработчика таблиц полностью основана на результатах работы семантического анализатора. Поэтому если семантический анализатор в заданной колонке не нашел NERC-метки, то обработчик таблиц пропустит эту колонку.
- Текущая реализация обработчика таблиц не позволяет извлекать характеристики связей.
- В ячейках с идентифицирующими характеристиками (например, названиями) могут попадаться пустые ячейки или может стоять прочерк или символ «н/д».
- В разных ячейках одной колонки могут быть представлены разные концепты с характеристиками (например, колонка может содержать одновременно как концепты типа «*Персона*», так и «*Организация*»).
- В одной ячейке могут быть представлены разные концепты или характеристики (например, ячейка может содержать регион с его географическими координатами).
- В одной ячейке могут быть представлены множественные значения концептов или характеристик одного типа (например, может быть перечисление имен или организаций).
- Концепт с его идентифицирующей характеристикой (названием) может быть расположен вне таблицы (например, в заголовке), а в самой таблице есть только характеристики этого концепта (например, для песни, название которой вынесено в заголовок, в таблице представлены только продолжительность песни и дата ее записи).
- Названия заголовков являются названиями характеристик концепта или связи.
- Характеристика концепта или связи является составной и распределена в нескольких колонках (например, счет в футбольном матче может быть разбит на несколько ячеек).
- Характеристика концепта или связи является составной и содержится в одной ячейке (например, год, состоящий из диапазона, или карьера игрока, состоящая из множества дат).
- Наличие вычисляемых значений ячеек в колонках (например, расчет времени участников в гонке относительно времени победителя).
- Разные единицы измерения, которые могут приводить к нескольким видам характеристик. Например, в двух таблицах с измерением численности населения указаны «*млн. чел.*» и «*тыс. чел.*».

Для устранения определенных выше проблем требуется улучшить существующие методы семантического аннотирования, в частности, требуется добавить:

- Корректную обработку извлечения идентифицирующих характеристик с учетом пустых ячеек, прочерков или специальных символов.
- Извлечение характеристик связей из таблиц.
- Извлечение фактов составных характеристик из таблиц, которые могут собираться как внутри одной ячейки, так и быть собраны из ячеек разных столбцов.
- Обработку множественных однотипных значений (концептов и характеристик) в ячейках.
- Извлечение фактов из таблиц с использованием внешнего контекста таблицы, в частности, связь с фактами, которые расположены в остальном документе (например, в заголовке названия таблицы).

В целом, полученные результаты показывают перспективность использования разработанного подхода и обработчика таблиц для поддержки процесса извлечения конкретных сущностей (фактов) из семантически аннотированных табличных данных и пополнения ими предметно-ориентированных графов знаний.

5. Заключение

Эффективное тестирование методологического и программного обеспечения для автоматической семантической интерпретации (аннотирования) таблиц и извлечения новых фактов из аннотированных табличных данных требует создания и использования русскоязычных наборов данных.

Основной вклад данного исследования заключается в создании первого русскоязычного набора табличных данных RF-200, охватывающего 26 предметных областей, а также в результатах оценки производительности авторского подхода. Набор опубликован и доступен для свободного использования на GitHub [36]. Программная реализация подхода в форме обработчика платформы Talisman продемонстрировала его превосходство над традиционными методами извлечения фактов из текстов, достигнув F-меры 0.464 на этапах NERC и RELEXT. Полученные результаты свидетельствуют о перспективности использования специализированных решений для работы со структурированными данными, особенно в условиях лингвистического разнообразия.

Результаты исследования имеют как теоретическую, так и практическую значимость. С теоретической точки зрения, предложенный метод аннотирования устраняет субъективность за счёт статистической верификации, что расширяет возможности семантической интерпретации таблиц за пределы числовых данных. С практической точки зрения, созданный набор данных RF-200 позволяет проводить эффективное тестирование производительности современных решений в области обработки таблиц и извлечения фактов. Однако работа выявила ряд ограничений. Во-первых, зависимость от качества распознавания именованных сущностей (NER) может приводить к пропуску колонок с редкими метками. Во-вторых, текущая реализация подхода не поддерживает извлечение характеристик связей и обработку составных значений в ячейках.

Перспективные направления будущих исследований включают интеграцию методов глубокого обучения, основанных на тонкой настройке предварительно-обученных языковых моделей (например, RuTABERT [14]), для повышения точности и автоматизации аннотирования таблиц с более сложной структурой. Созданный набор данных RF-200 будет расширен за счёт включения горизонтальных и матричных таблиц с иерархическими заголовками с объединёнными ячейками, а также поддержку мультиязычности. Кроме того, для подтверждения выводов планируется провести дополнительные статистические тесты,

расчёты доверительных интервалов и измерения межаннотационного согласия на RF-200, с целью определения типов таблиц, для которых предлагаемый подход обеспечивает получение максимальных оценок. Перспективной также является задача обеспечения отображения элементов существующей онтологической схемы графа знаний платформы Talisman в онтологические понятия графов знаний общего назначения такие как Wikidata или DBpedia для увеличения семантической согласованности и упрощения повторного использования созданного набора данных RF-200 за счет предоставления стандартизированной поддержки видов фактов в формате Семантического Веба (RDF/OWL).

Список литературы

[1]. Hogan A., Blomqvist E., Cochez M., d'Amato C., De Melo G., Gutierrez C., Gayo J. E. L., Kirrane S., Neumaier S., Polleres A., Navigli R., Ngomo A.-C. N., Rashid S. M., Rula A., Schmelzeisen L., Sequeda J., Staab S., Zimmermann A. Knowledge Graphs. Springer Nature Switzerland, 2021, 237 p. DOI: 10.1007/978-3-031-01918-0.

[2]. Ji S., Pan S., Cambria E., Marttinen P., Yu P.S. A Survey on Knowledge Graphs: Representation, Acquisition and Applications. IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, 2021, pp. 494-514. DOI: 10.1109/TNNLS.2021.3070843.

[3]. 5-star Open Data, Available at: https://5stardata.info/en/, accessed 22.04.2025.

[4]. DBpedia, Available at: https://www.dbpedia.org/, accessed 22.04.2025.

[5]. Wikidata, Available at: https://www.wikidata.org/, accessed 22.04.2025.

[6]. Villazon-Terrazas B., Garcia-Santa N., Ren Y., Srinivas K., Rodriguez-Muro M., Alexopoulos P., Pan J. Z. Construction of Enterprise Knowledge Graphs (I). Exploiting Linked Data and Knowledge Graphs in Large Organisations, Springer, Cham, 2017.

[7]. Number of Google Sheets and Excel Users Worldwide, Available at: https://askwonder.com/research/number-google-sheets-users-worldwide-eoskdoxav, accessed 22.04.2025.

[8]. Peeters R., Brinkmann A., Bizer C. The Web Data Commons Schema.org Table Corpora. Proc. the ACM Web Conference (WWW'24), New York, NY, USA, 2024, pp. 1079-1082. DOI: 10.1145/3589335.3651441.

[9]. Talend, Available at: https://www.talend.com/, accessed 22.04.2025.

[10]. Trifacta, Available at: https://asana.com/ru/apps/trifacta, accessed 22.04.2025.

[11]. Microsoft Semantic Link, Available at: https://learn.microsoft.com/en-us/fabric/data-science/semantic-link-overview, accessed 22.04.2025.

[12]. Talisman, Available at: http://talisman.ispras.ru, accessed 22.04.2025.

[13]. Dorodnykh N. O., Yurin A. Yu. Automated Extraction of Facts from Tabular Data based on Semantic Table Annotation. Trudy ISP RAN/Proc. ISP RAS, vol. 36, no. 3, 2024, pp. 93-104. DOI: 10.15514/ISPRAS-2024-36(3)-7.

[14]. Fedorov P. E., Mironov A. V., Chernishev, G. A. Russian Web Tables: A Public Corpus of Web Tables for Russian Language Based on Wikipedia. Lobachevskii Journal of Mathematics, vol. 44, 2023, pp. 111-122. DOI: 10.1134/S1995080223010110.

[15]. Kruit B., Boncz P., Urbani J. Extracting novel facts from tables for knowledge graph completion. Proc. the 18th International Semantic Web Conference (ISWC'2019), Auckland, New Zealand, 2019, pp. 364-381. DOI: 10.1007/978-3-030-30793-6_21.

[16]. Zhang S., Meij E., Balog K., Reinanda R. Novel entity discovery from web tables. Proc. the ACM Web Conference (WWW'20), New York, NY, USA, 2020, pp. 1298-1308. DOI: 10.1145/3366423.3380205.

[17]. Zhang S., Balog K. Web Table Extraction, Retrieval, and Augmentation: A Survey. ACM Transactions on Intelligent Systems and Technology, vol. 11, no. 2, 2020, pp. 1-35. DOI: 10.1145/3372117.

[18]. Balog K. Populating Knowledge Bases. Entity-Oriented Search INRE, vol. 39, 2018, pp. 189-222. DOI: 10.1007/978-3-319-93935-3_6.

[19]. Subagdja B., Shanthoshigaa D., Wang Z., Tan A.-H. Machine Learning for Refining Knowledge Graphs: A Survey. ACM Computing Surveys, vol. 56, no. 6, 2024, pp. 1-38. DOI: 10.1145/3640313.

[20]. SemTab-2024, Available at: https://sem-tab-challenge.github.io/2024/, accessed 22.04.2025.

[21]. Bonfitto S., Casiraghi E., Mesiti M. Table understanding approaches for extracting knowledge from heterogeneous tables. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 11, no. 4, 2021, e1407. DOI: 10.1002/widm.1407.

[22]. Zheng M., Feng X., Si Q., She Q., Lin Z., Jiang W., Wang W. Multimodal Table Understanding. Proc. the 62nd Annual Meeting of the Association for Computational Linguistics (ACL'2024), Bangkok, Thailand, 2024, pp. 9102-9124. DOI: 10.18653/v1/2024.acl-long.493.

[23]. Limaye G., Sarawagi S., Chakrabarti S. Annotating and searching web tables using entities, types and relationships. Proceedings of the VLDB Endowment, vol. 3, no. 1-2, 2010, pp. 1338-1347. DOI: 10.14778/1920841.1921005.

[24]. T2Dv2 Gold Standard for Matching Web Tables to DBpedia, Available at: https://webdatacommons.org/webtables/goldstandardV2.html, accessed 22.04.2025.

[25]. Cutrona V., Bianchi F., Jimenez-Ruiz E., Palmonari M. Tough tables: Carefully evaluating entity linking for tabular data. Proc. the 19th International Semantic Web Conference (ISWC'2020), Athens, Greece, 2020, pp. 328-343. DOI: 10.1007/978-3-030-62466-8_21.

[26]. Abdelmageed N., Schindler S., Konig-Ries B. Biodivtab: A table annotation benchmark based on biodiversity research data. Proc. the 20th International Semantic Web Conference (ISWC'2021) – Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab-2021), 2021, pp. 13-18.

[27]. Hulsebos M., Demiralp C., Groth P. GitTables: A Large-Scale Corpus of Relational Tables. Proceedings of the ACM on Management of Data, vol. 1, no. 1, 2023, pp. 1-17. DOI: 10.1145/3588710.

[28]. SOTAB (Web Data Commons - Schema.org Table Annotation Benchmark), Available at: https://webdatacommons.org/structureddata/sotab/, accessed 22.04.2025.

[29]. Zhang D., Suhara Y., Li J., Hulsebos M., Demiralp C., Tan W.-C. Sato: Contextual semantic type detection in tables. Proc. the VLDB Endowment, vol. 13, no. 11, 2020, pp. 1835-1848. DOI: 10.14778/3407790.3407793.

[30]. Deng X., Sun H., Lees A., Wu Y., Yu C. TURL: Table Understanding through Representation Learning. Proc. the VLDB Endowment, vol. 14, no. 3, 2020, pp. 307-319. DOI: 10.14778/3430915.3430921.

[31]. Tobola K. V., Dorodnykh N. O. Semantic Annotation of Russian-Language Tables Based on a Pre-Trained Language Model. Proc. the 2024 Ivannikov Memorial Workshop (IVMEM), 2024, pp. 62-68. DOI: 10.1109/IVMEM63006.2024.10659709.

[32]. Hao Q., Cai R., Pang Y., Zhang L. From one tree to a forest: a unified solution for structured web data extraction. Proc. the 34th international ACM SIGIR conference on Research and development in Information Retrieval, Beijing, China, 2011, pp. 775-784. DOI: 10.1145/2009916.2010020.

[33]. Gupta T., Zaki M., Khatsuriya D., Hira K., Krishnan N. M. A., Mausam. DISCOMAT: Distantly Supervised Composition Extraction from Tables in Materials Science Articles. Proc. the 61st Annual Meeting of the Association for Computational Linguistics (ACL'2023), Toronto, Canada, 2023, pp. 13465-13483. DOI: 10.18653/v1/2023.acl-long.753.

[34]. Bai F., Kang J., Stanovsky G., Freitag D., Dredze M., Ritter A. Schema-Driven Information Extraction from Heterogeneous Tables. Proc. the 61st Annual Meeting of the Association for Computational Linguistics (ACL'2024), Miami, Florida, USA, 2024, pp. 10252-10273. DOI: 10.18653/v1/2024.findings-emnlp.600.

[35]. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. Unsupervised Cross-lingual Representation Learning at Scale. Proc. the 58th Annual Meeting of the Association for Computational Linguistics (ACL'2020), 2020, pp. 8440-8451. DOI: 10.18653/v1/2020.acl-main.747.

[36]. RF-200 (ru-facts-200), Available at: https://github.com/YRL-AIDA/ru-facts-200, accessed 22.04.2025.

Информация об авторах / Information about authors

Никита Олегович ДОРОДНЫХ – кандидат технических наук, старший научный сотрудник Института динамики систем и теории управления им. В.М. Матросова Сибирского отделения РАН (ИДСТУ СО РАН) с 2021 года. Сфера научных интересов: автоматизация создания интеллектуальных систем и баз знаний, получение знаний на основе преобразования концептуальных моделей и электронных таблиц.

Nikita Olegovych DORODNYKH – Cand. Sci. (Tech.), senior associate researcher at Matrosov Institute of System Dynamics and Control Theory named SB RAS (ISDCT SB RAS) since 2021. Research interests: computer-aided development of intelligent systems and knowledge bases, knowledge acquisition based on the transformation of conceptual models and tables.

Александр Юрьевич ЮРИН – доктор технических наук, заведующий лабораторией Информационно-телекоммуникационных технологий исследования природной и техногенной безопасности ИДСТУ СО РАН, профессор Института информационных технологий и анализа данных Иркутского научно-исследовательского технического университета (ИрНИТУ). Его научные интересы включают разработку систем поддержки принятия решений, экспертных систем и баз знаний, использование прецедентного подхода и семантических технологий при проектировании интеллектуальных диагностических систем.

Alexander Yurievich YURIN – Dr. Sci. (Tech.), Head of a laboratory “Information and telecommunication technologies for investigation of natural and technogenic safety” at ISDCT SB RAS and professor of the Institute of information technologies and data analysis of Irkutsk National Research Technical University (INRTU). His research interests include development of decision support systems, expert systems and knowledge bases, application of the case-based reasoning and semantic technologies in the design of diagnostic intelligent systems, maintenance of reliability and safety of complex technical systems.