Адаптация подхода SimCLR к аннотированию столбцов русскоязычных таблиц

Кирилл Тобола[1], Никита Дородных[1]

[1] *Институт динамики систем и теории управления им. В.М. Матросова СО РАН, ул. Лермонтова, д. 134, 664033, Иркутск, Россия*

Аннотация

Таблицы часто лишены явной семантики, что затрудняет их машинную интерпретацию. Для интеграции с графами знаний особенно актуальна задача семантитческой интерпретации столбцов русскоязычных таблиц, однако существующие методы ограничены нехваткой размеченных данных и языковой спецификой. В работе предложен подход на основе адаптации контрастного обучения из работы SimCLR для русскоязычных таблиц, использующий DistilBERT в качестве базового кодировщика для обработки неразмеченных данных корпуса Russian Web Tables. Результаты для задачи семантического аннотирования русскоязычных столбцов превосходят базовые методы, подтверждая эффективность при работе с редкими типами и русскоязычными данными.

Ключевые слова

Русскоязычные таблицы, табличные данные, семантическое аннотирование столбцов, графы знаний, контрастное обучение, табличные представления.

1. Введение

Табличные данные представляют собой фундаментальный формат структурированного представления информации, широко используемый в научных исследованиях и бизнес-аналитике. Однако автоматизированная обработка таких данных сопряжена со значительными трудностями, обусловленными отсутствием явной семантической разметки, что особенно актуально для табличных данных на языках, отличных от английского [1]. В частности, русскоязычные таблицы сталкиваются с дополнительными ограничениями, связанными с недостатком специализированных инструментов и аннотированных корпусов. Существующие методы, основанные на архитектуре BERT [2–4], демонстрируют высокую эффективность, но требуют значительных объемов размеченных данных, которые для русского языка остаются ограниченными. Кроме того, англоязычные решения демонстрируют низкую адаптивность к русскоязычным таблицам из-за различий в токенизации и семантических особенностях языка.

2. Предлагаемый подход

Настоящее исследование предлагает инновационный метод семантической аннотации столбцов русскоязычных табличных данных, основанный на применении контрастного обучения в рамках архитектуры SimCLR [5]. Предлагаемый подход позволяет минимизировать проблему недостаточного объема размеченных данных за счет использования неразмеченных примеров и выявления скрытых семантических закономерностей. Ключевая задача семантической интерпретации столбцов заключается в автоматическом определении их семантического типа (например, «Спортсмен», «Литературный жанр» и аналогичных).

7th International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2025), July 7–11, 2025, Irkutsk, Russia
EMAIL: kirilltobola@icc.ru (A. 1); tualatin32@mail.ru (A. 2)
ORCID: 0009-0006-1014-451X (A. 1); 0000-0001-7794-4462 (A. 2)

ICCS-DE 2025 Workshop Proceedings
DOI: 10.47350/ICCS-DE.2025.34

Методологическая основа исследования заключается в разработке специализированного русскоязычного кодировщика для генерации устойчивых табличных представлений. Полученные векторные представления демонстрируют эффективность при решении задачи семантической интерпретации столбцов в русскоязычных таблицах.

2.1. Алгоритм обучения

Предлагаемый кодировщик табличных представлений был обучен на модифицированной версии набора данных RWT (Russian Web Tables) [6], содержащей 4,6 миллиона неразмеченных столбцов, с использованием метода контрастного обучения.

Контрастное обучение представляет собой метод самообучения, направленный на формирование информативных векторных представлений. Основной принцип данного подхода заключается в максимизации меры сходства между положительными парами данных и одновременной минимизации сходства между отрицательными. Такой механизм обеспечивает эффективное обучение на неразмеченных данных. Процедура обучения включает следующие этапы:

1. Создание двух аугментаций для каждого столбца (положительная пара).
2. Кодирование их в векторные представления.
3. Сближение векторов положительной пары и отдаление векторов отрицательных пар (отрицательными парами считаются аугментации, полученные из различных исходных столбцов).

В качестве двух аугментаций были выбраны: случайное удаление ячеек в столбце и перестановка строк внутри столбца.

2.2. Результаты

Экспериментальная часть исследования была выполнена на вычислительном кластере «Академик В. М. Матросов». Обучение модели кодировщика табличных представлений осуществлялось в течение 100 эпох с использованием четырёх графических ускорителей NVIDIA A100 (затрачено 9 дней 9 часов и 290 ГБ видеопамяти).

После завершения обучения кодировщик был интегрирован в архитектуру модели RuTaBERT [7] и дообучен на размеченном наборе данных RWT-RuTaBERT [8] для решения задачи семантической интерпретации столбцов русскоязычных таблиц (затрачено 2 дня 20 часов и менее 10 ГБ видеопамяти).

В качестве основного критерия эффективности модели использовалась F1-мера. Учитывая несбалансированность датасета RWT-RuTaBERT, применялись микро-F1 и макро-F1 метрики.

Для оценки эффективности предложенного подхода были проведены сравнительные тесты с моделями Doduo [4] и базовой версией RuTaBERT на наборе данных RWT-RuTaBERT. Результаты экспериментальной оценки представлены в Таблице 1.

Таблица 1

Результаты экспериментальной оценки для задачи семантической интерпретации столбцов русскоязычных таблиц на наборе RWT-RuTaBERT.

Модель	Микро F1	Макро F1
DODUO	96,2%	89,1%
RuTaBERT	96,4%	90,0%
Предлагаемый подход	**96,9%**	**91,0%**

Результаты исследования свидетельствуют о перспективности предложенного подхода. Наблюдаемое улучшение показателя макро F1 на 1-2% демонстрирует его повышенную эффективность в прогнозировании редко встречающихся семантических типов. Данное обстоятельство позволяет сделать вывод о более высокой обобщающей способности

разработанного метода, что обусловлено применением устойчивых табличных представлений кодировщика, сформированных в процессе контрастного обучения.

Также использование метода позволило достичь передовых показателей при меньших вычислительных затратах. В частности, предложенный метод требует примерно в три раза меньше видеопамяти в процессе обучения, что делает возможным его применение на стандартных рабочих станциях без необходимости использования специализированного оборудования.

3. Заключение

В данной работе представлен метод семантического аннотирования столбцов русскоязычных табличных данных, основанный на принципах контрастного обучения. Результаты проведенных экспериментов свидетельствуют о том, что предложенный подход позволяет преодолеть зависимость от значительных объемов размеченных данных благодаря механизму самообучения на неразмеченных таблицах.

Сравнительный анализ показал, что разработанный метод демонстрирует более высокую эффективность по сравнению с существующими базовыми решениями. Также достигается значительное повышение вычислительной эффективности: требования к объему памяти графических процессоров сокращаются на 60% относительно аналогов.

Предлагаемый подход создает перспективы для разработки универсальных систем семантической интерпретации табличных данных, что представляет особую актуальность для задач интеграции разнородной структурированной и слабоструктурированной информации.

4. Благодарности

Работа выполнена при поддержке государственного задания Министерства науки и высшего образования Российской Федерации (тема № 1023110300006-9).

5. Список источников

[1] Liu J., Chabot Y., Troncy R., Huynh V.-P., Labbe T., Monnin P. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. Journal of Web Semantics, volume 76, 2023, 100761. doi: 10.1016/j.websem.2022.100761.

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

[3] Yin P., Neubig G., Yih W. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. Proc. the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8413-8426. doi: 10.18653/v1/2020.acl-main.745.

[4] Suhara Y., Li J., Li Y. Annotating Columns with Pre-trained Language Models. Proc. the 2022 International Conference on Management of Data (SIGMOD'22), 2022, pp. 1493-1503. doi: 10.1145/3514221.3517906.

[5] Chen T., Kornblith S., Norouzi M., Hinton G. A simple framework for contrastive learning of visual representations. Proc. the 37th International Conference on Machine Learning (ICML'20), Online, 2020, pp. 1597-1607. doi: 10.5555/3524938.3525087.

[6] Fedorov, P., Mironov, A., Chernishev, G. Russian Web Tables: A Public Corpus of Web Tables for Russian Language Based on Wikipedia. Lobachevskii Journal of Mathematics, 2023, pp. 111-122.

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

[8] RWT-RuTaBERT, 2025. URL: https://huggingface.co/datasets/sti-team/rwt-rutabert.