Semantic Column Annotation for Russian Tables Using Multilingual BERT Language Model

Kirill V. Tobola[1,2], Nikita O. Dorodnykh[2]

[1]*Institute of Mathematics and Information Technologies, Irkutsk State University*

[2]*Matrosov Institute for System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences (ISDCT SB RAS)*

Abstract

Tables are widely used for organizing, presenting, and storing data. Automatically extracting information from these tables is challenging as they often have arbitrary and non-uniform layouts. Recent research has demonstrated significant progress in addressing various challenges in automatic table understanding. But, these are often limited to languages such as English. In this paper, we propose a new RuTaBERT framework, based on a pre-trained language model, for semantic annotation of columns in Russian-language tables. We utilized a preprocessed large-scale corpus of Russian-language web tables, extracted from Wikipedia pages, as training data to fine-tune the language model. Experimental results show that RuTaBERT establishes new state-of-the-art performance on new benchmark with additional language for the column type prediction task with micro F1 score of up to 95%.

Keywords

Semantic table interpretation, Column type annotation, BERT, Pre-trained language model, Knowledge graph, Tables, Russian language

1. Introduction

This paper focuses on improving the comprehension of semantic table structures, particularly for tables written in Russian. We suggest a framework called RuTaBERT, which employs pre-trained language models to enhance semantic column annotation. RuTaBERT is built upon the multilingual BERT model [1] and leverages local table context to boost annotation accuracy. Additionally, we develop a method for automatically labeling source tabular datasets based on table headers. The RuTaBERT codebase and trained models are publicly available, allowing for their application in various systems for table column annotating and performance evaluation.

Overall, the paper offers a significant contribution to the field of table understanding and semantic annotation, with a focus on Russian-language tables. The proposed framework and methods have the potential to increase the efficiency and accuracy of table analysis and knowledge discovery in various domains.

6^{th} International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2024), July 01–05, 2024, Irkutsk, Russia

kirilltobola@gmail.com (K. V. Tobola); nikidorny@icc.ru (N. O. Dorodnykh)

ICCS-DE 2024 Workshop Proceedings (iccs-de.icc.ru)
DOI: 10.47350/ICCS-DE.2024.26

2. Methodology

2.1. Problem statement

A table is a 2D array of rows and columns, where each cell can hold various types of data. In this approach, vertical tables are used as input data, where each column may have a header. The goal is to predict the semantic type of each column, such as "Book" or "Publication Date", rather than standard data types like string or integer. To achieve this, a set of 356 semantic types based on the DBpedia knowledge graph has been created, including classes, data properties, and object properties, and translated into Russian. The task is to annotate Russian-language tables with these semantic types.

Figure 1 illustrates an example of solving the CTA task for the input table.

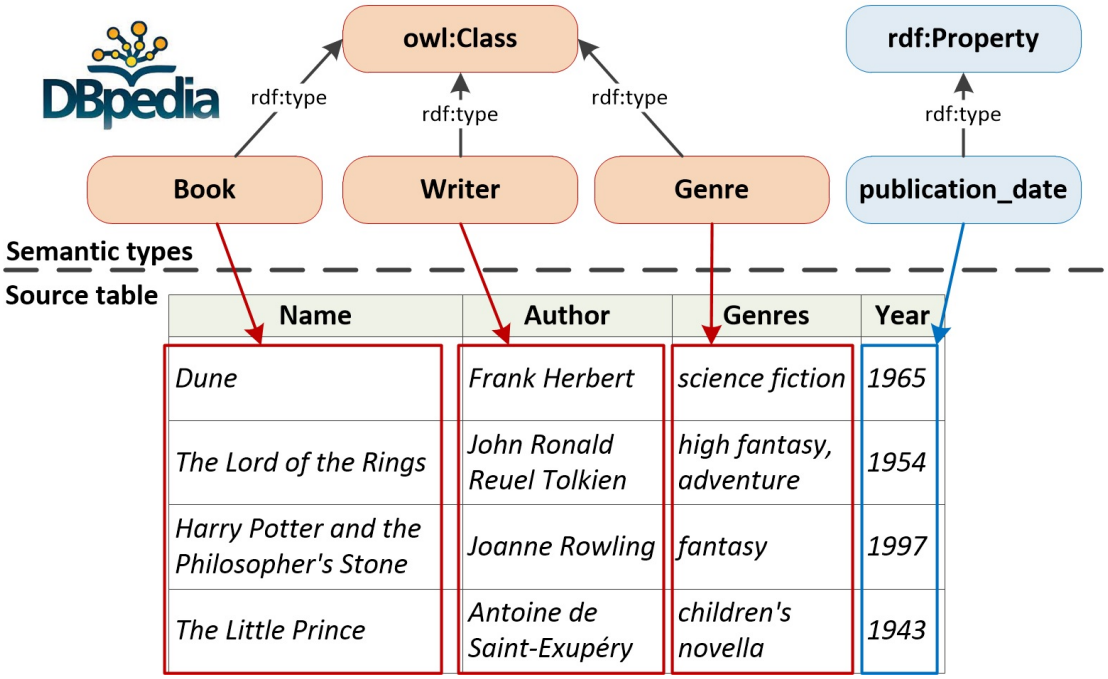

Figure 1: Example of column type annotation problem.

2.2. Model architecture

We suggest a framework called RuTaBERT that fine-tunes the multilingual BERT model for sequence classification tasks, such as column labeling in tables. We also propose two context-aware table serialization methods, multi-column serialization and neighboring column serialization, to convert tables into text sequences for input to the BERT model.

1) *Multi-column serialization* based on the approach [2] and represents data records as follows:

$$S_{mc}(T) = [CLS]tks^1...[CLS]tks^n[SEP]$$

where [CLS] marks the start of a sequence; [SEP] marks the end; tk_i is the tokens for i-column.

In this case, all cell values of a i-column are tokenized: $ci = v_1, ..., v_m$. This method inserts a [CLS] token for each column, allowing RuTaBERT framework to predict a label for each [CLS] token.

2) *Neighboring column serialization* uses neighboring columns as local context:

$$S_{nc}(T) = [CLS]tks^{trg}[SEP]tks^{nb}[SEP]$$

where tks^{trg} represents the tokens of the target column; tks^{nb} represents the tokens for all non-target neighboring columns. Here, only one [CLS] token for the target column is used for prediction, and [SEP] tokens separate non-target columns.

3. Results and discussion

We trained RuTaBERT for 10 epochs with a batch size of 32 using both serialization methods on Russian Web Tables (RWT) [3] dataset. The preprocessed tables from the RWT corpus were split into training, validation, and test datasets with standard proportions: 70%, 20%, and 10%, respectively.

We evaluate RuTaBERT performance on the column type annotation (CTA) task using two serialization methods, multi-column and neighboring column, and compare it to a baseline method based on lexical matching between table header names and semantic types. Table 1 represents the experimental evaluation of RuTaBERT on the CTA task using the RWT dataset.

Table 1
The experimental evaluation for RuTaBERT

Method	Micro F1	Macro F1	Weighted F1
baseline	0.881	0.693	0.902
multi-column	0.952	0.861	0.951
neighboring	0.953	0.856	0.953

According to the results, both serialization techniques have similar performance and achieve better results than the baseline, indicating the framework's potential for semantic annotation of Russian-language tables. However, the models face challenges related to data sparsity in the table corpus, which is evident in the imbalanced distribution of semantic types. This affects the models' ability to capture sufficient signals for minority class semantic types.

Acknowledgments

The reported study was supported by the Ministry of Science and Higher Education of the Russian Federation (№ 1023110300006-9).

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[2] Y. Suhara, J. Li, Y. Li, D. Zhang, c. Demiralp, C. Chen, W.-C. Tan, Annotating Columns with Pre-trained Language Models, Association for Computing Machinery, 2022. URL: https://doi.org/10.1145/3514221.3517906.

[3] P. Fedorov, A. Mironov, G. Chernishev, Russian web tables: A public corpus of web tables for russian language based on wikipedia, 2022. arXiv:2210.06353.