Russian address elements classification using artificial neural networks

Anton Reshetnikov

May 2024

Abstract

This documents provides result of research of token classification task applied to russian place addresses. The task is to recognize elements of address like region, area, city, territory, street. The source code available at `https://github.com/qwazer/ruaddress-elements-classification`.

1 Introduction

Logistics companies have the task of checking and normalizing the address against the incoming string. For example, there is an input string with the address "Москва, Абрат, 1" (in Russian). It is necessary to select address-forming elements, check against the database, that such an address actually exists and return the status of the address and its normalized representation.

Traditionally, IT systems that solve this problem using a rules-based approach (word level tokenization, token classification by regular expressions). The task is to explore the possibilities of solving the same problem using modern deep learning models.

1.1 Team

Anton Reshetnikov - researcher

2 Related Work

[Макаров, 2020] provides overview of traditional algorithms used for address elements recognition. The particular solution for address element recognition described in [Кузнецов, 2020]

3 Model Description

The most modern deep learning models based on Transformer architecture described in the "Attention Is All You Need" article [Vaswani et al., 2023]. There

are several models adapted for russian language exists. For detail see this overview article: [Zmitrovich et al., 2024].

For the purpose of current research cointegrated/rubert-tiny2 model selected. Reasons of such choice are:

1. It based on Transformer architecture;
2. The model is focused on Russian;
3. Small size, which make it suitable for fast experiments.

4 Dataset

4.1 Datasource

The dataset based on freely available "State address register" distributed at `https://fias.nalog.ru/` by Federal Taxation Service of the Russian Federation.

Only address elements related to Administrative division (Административно-территориальное деление) was selected from the "State address register".

4.2 Token classes

Dictionary "Element type" used as token classes for token classification task. Values of dictionary are:

1. REGION
2. REGION_TYPE
3. AREA
4. AREA_TYPE
5. TERRITORY
6. TERRITORY_TYPE
7. CITY
8. CITY_TYPE
9. STREET
10. STREET_TYPE
11. DELIMITER

4.3 Mapping table

Address elements levels of "State address register" mapped to custom dictionary "Element type" with next mapping table:

State address register level (in Russian)	Element type	Example
Субъект	REGION	Омская область
Административный район	AREA	Любинский р-н
Город	CITY	
Населенный пункт	CITY	поселок Камышловский
Элемент планировочной структуры	TERRITORY	Территория СНТ Сибзаводовец-2
Элемент улично-дорожной сети	STREET	7-я аллея
Земельный участок	not used	
Здание (сооружение)	not used	
Помещение	not used	
Помещения в пределах помещения	not used	
Машино-место	not used	

In case of REGION and CITY has the same value (like for "город Москва", "город Санкт-Петербург", "город Севастополь", "город Байконур") CITY tokens are omitted.

4.4 Dataset structure

The idea is to generate own dataset called "Ruaddress" from "State address register" datasource. The "Ruaddress" dataset has 2 columns, described in the next table:

Column name	tokens	classes
Description	list of token words	list of class codes
Example	[Вологодская, Область,,, Грязовецкий, Район, Вохтога, Рабочий, поселок]	[1, 2, 11, 3, 4, 7, 8, 8]

4.5 Augmenation

A place address can be presented in a various forms. To generate address form from address elements the augmentation procedure is used. The procedure receive address-forming elements for a place, then apply series of transformation and output 2 lists: tokens and related token classes.

4.5.1 Formats of address element

Controlled by next options:

Option	Probability	Example for True value	Example for False value
typeFirst	0.75	поселок Мирный	Завьяловский район
commaSeparator	0.75	поселок Мирный,	Завьяловский район
shortType	0.5	п. Мирный,	Завьяловский р-н

4.5.2 Char level augmentation

For adding some noise and corruptions to the words the augmentex tool used. The augmentex tool described in next paper [Martynov et al., 2023]. It used with next config:

```
char_aug = CharAug(
    unit_prob=0.05, # Percentage of the phrase to which augmentations will be applied
    min_aug=0, # Minimum number of augmentations
    max_aug=3, # Maximum number of augmentations
    mult_num=3, # Maximum number of repetitions of characters
    random_seed=42,
    lang="rus", # supports: "rus", "eng"
    platform="pc", # supports: "pc", "mobile"
)
```

4.5.3 Word shortening

For word shortening 2 options are used.

Option	Probability	Example for Красноярский	
hyphen substition	0.15	Красн-кий	
dot truncation	0.15	Красн.	

4.6 "Ruaddress" dataset

The "Ruaddress" dataset consist of 1,5 millions augmented addresses. The dataset was split on train, test, validation parts in 4:1:1 (0.66:0.16:0.16) proportion accordingly.

5 Experiments

The model was trained using Jupiter notebook from the [Gugger et al., 2022] article. The original Jupiter notebook from Hugging face NPL course was adapted for the "Ruaddress" dataset.

5.1 Metrics

Standard metrics for classification task are used: Precision, Recall, F1.

5.2 Experiment Setup

The experiments shown that the model can be effectively trained in 1 epoch. The second epoch and the third epoch increase metrics in less than 0,005 %.

The whole list of parameters are shown in the next code snippet

```
args = TrainingArguments(
    "bert-finetuned-ner",
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    num_train_epochs=1,
    weight_decay=0.01
)
```

6 Results

The trained model solve address token classification task.

Example

Input:

```
Ставропольский край г Лермонтов территория
садоводческого некоммерческого товарищества
имени И.В. Мичурина, ул массив 3 линия 3
```

Output:

```
REGION - Ставропольский
REGION_TYPE - край
CITY_TYPE - г
CITY - Лермонтов
TERRITORY_TYPE - территория
TERRITORY - садоводческого некоммерческого товарищества имени И. В. Мичурина
DELIMITER - ,
STREET_TYPE - ул
STREET - массив 3 линия 3
```

The model published to huggingface at qwazer/rubert-address-elements. The example of Inference API output show on Fig. 1

The comparisons with other models was not performed in scope of the current work.

7 Conclusion

In scope of the current work the “Ruaddress” dataset based on “State address register” was build. It was augmented using different techniques. The model cointegrated/rubert-tiny2 was fine-tined for the token classification task applied

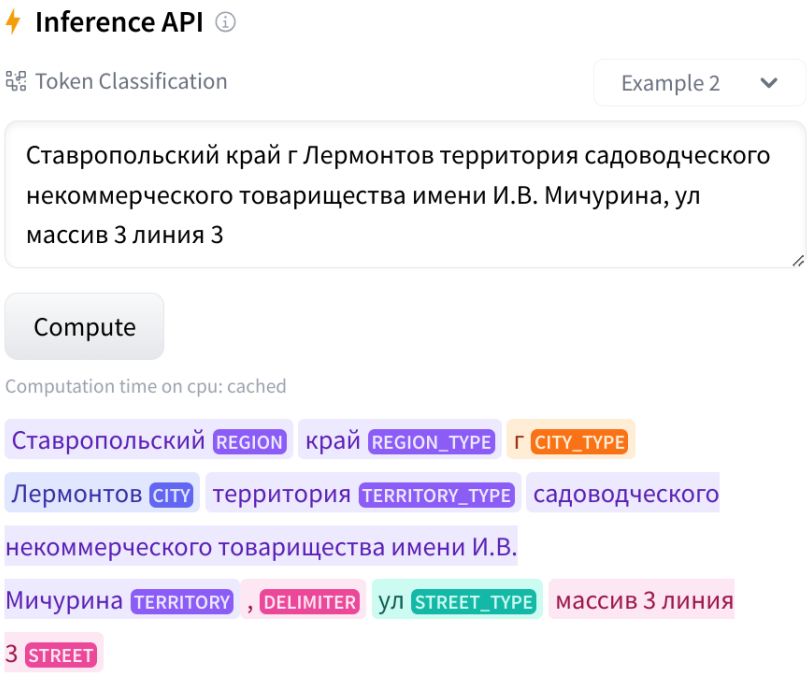

Figure 1: Inference example

to russian place addresses. The resulting model was published to huggingface at qwazer/rubert-address-elements.

7.1 Futher steps and ideas to improve the model

1. Add more tokens for buildings, steads, flats, possible inclusion of person or organization names, etc.
2. More word-level augmentations like word drop, swap, etc
3. Support for translit
4. Build the model bases on Recurrent networks and compare metrics and performance.
5. Try to build multi-classification model, when a word can have multiple tokens.

References

[Gugger et al., 2022] Gugger, S. et al. (2022). Huggingface npl course, token classification chapter. `https://huggingface.co/learn/nlp-course/chapter7/2/`.

[Martynov et al., 2023] Martynov, N., Baushenko, M., Abramov, A., and Fenogenova, A. (2023). Augmentation methods for spelling corruptions. In *Proceedings of the International Conference Dialogue*, volume 2023.

[Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

[Zmitrovich et al., 2024] Zmitrovich, D., Abramov, A., Kalmykov, A., Tikhonova, M., Taktasheva, E., Astafurov, D., Baushenko, M., Snegirev, A., Kadulin, V., Markov, S., Shavrina, T., Mikhailov, V., and Fenogenova, A. (2024). A family of pretrained transformer language models for russian.

[Кузнецов, 2020] Кузнецов, (2020). Нормализация адресов, ГАР ФИАС и Адрессарий. `https://habr.com/ru/articles/672186/`.

[Макаров, 2020] Макаров, (2020). Алгоритмы распознавания почтовых адресов. In *E-Scio*, volume 46, pages 352–357.