

# Разработка алгоритма обработки естественного языка при помощи машинного обучения с использованием нейросетей BERT

Научный руководитель: Н.А. Прокопьев

Студент: С.С. Саидмуродов, 09-852

# Automatic Speech Recognition (ASR) → теряем пунктуацию

**Яндекс Переводчик** Текст Сайты Документы Картинки Для бизнеса

РУССКИЙ

↔ ТАТАРСКИЙ

это пример работы алгоритмы распознавания речи от Яндекса как видимо он может распознавать именованные сущности но при этом знаки препинания не ставишь такой текст сложнее читать и при дальнейшем обработки возникают трудности

225 / 10000

бу-сөйләмне Яндекстан тану алгоритмының эш үрнәге, күрәсең, ул исемдәге асылларны таний ала, ләкин шул ук вакытта уку билгеләре мондый текстны укырга кыенрак куясың һәм алга таба эшкәрткәндә кыенлыклар килеп чыга

Перевести в Google

сообщить об ошибке

# Зачем нужна пунктуация?

- Наличие пунктуации:
  - Влияет на восприятие текста
  - Улучшает работу таких алгоритмов обработки естественного языка как:
    - Машинный перевод
    - Анализ тональности текста
    - Извлечение информации
    - Распознавание именованных сущностей

- Цели:

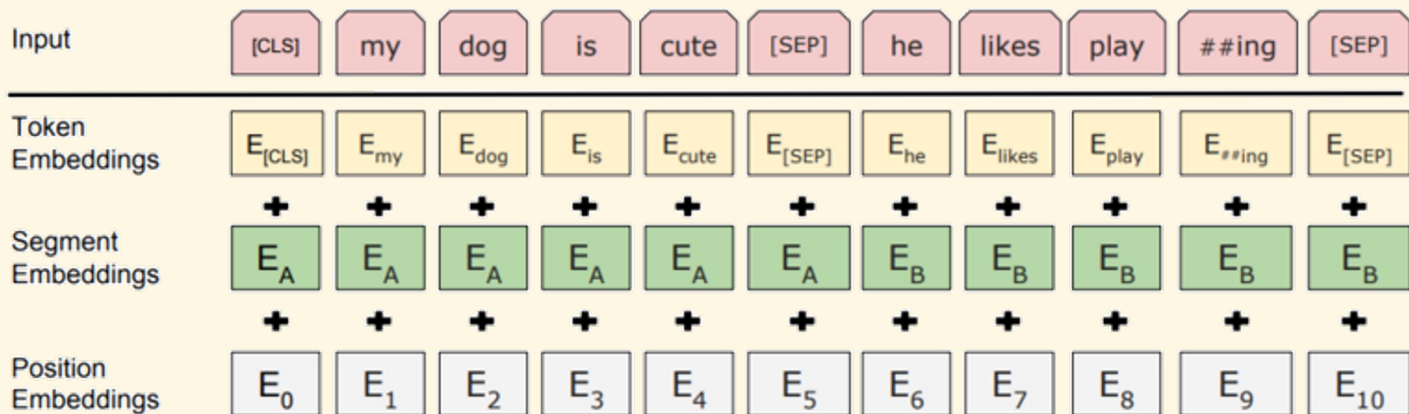
- Изучение возможности реализации программного обеспечения для восстановления знаков пунктуации в тексте для татарского языка

- Задачи:

- Изучение предметной области машинного обучения и нейронных сетей
- Анализ существующих подходов к обработке естественного языка
- Экспериментальная реализация алгоритма
- Разработка модели восстановления пунктуации
- Тестирование по результатам обучения модели

# Анализ существующих подходов

- Акустическая и **языковая** модели
- Bidirectional Encoder Representations from Transformers (BERT)



# Аналоги

- Google Speech-To-Text API

```
curl -s -H "Content-Type: application/json" \
  -H "Authorization: Bearer "$(gcloud auth print-access-token) \
  https://speech.googleapis.com/v1/speech:recognize \
  --data '{
    "config": {
      "encoding": "FLAC",
      "sampleRateHertz": 16000,
      "languageCode": "en-US",
      "enableAutomaticPunctuation": true
    },
    "audio": {
      "uri": "gs://cloud-samples-tests/speech/brooklyn.flac"
    }
  }'
```

- NeMo PunctuationCapitalizationModel

## NVIDIA/NeMo

NeMo: a toolkit for conversational AI



148

Contributors

32

Issues

211

Discussions

4k

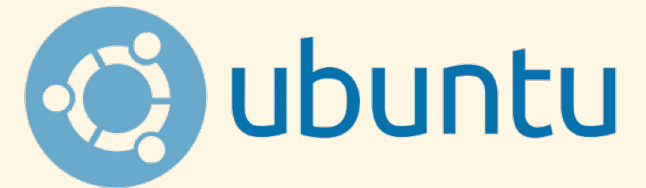
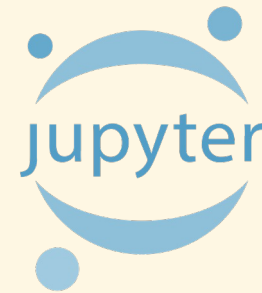
Stars

1k

Forks



# Инструменты разработки



# Архитектура модели

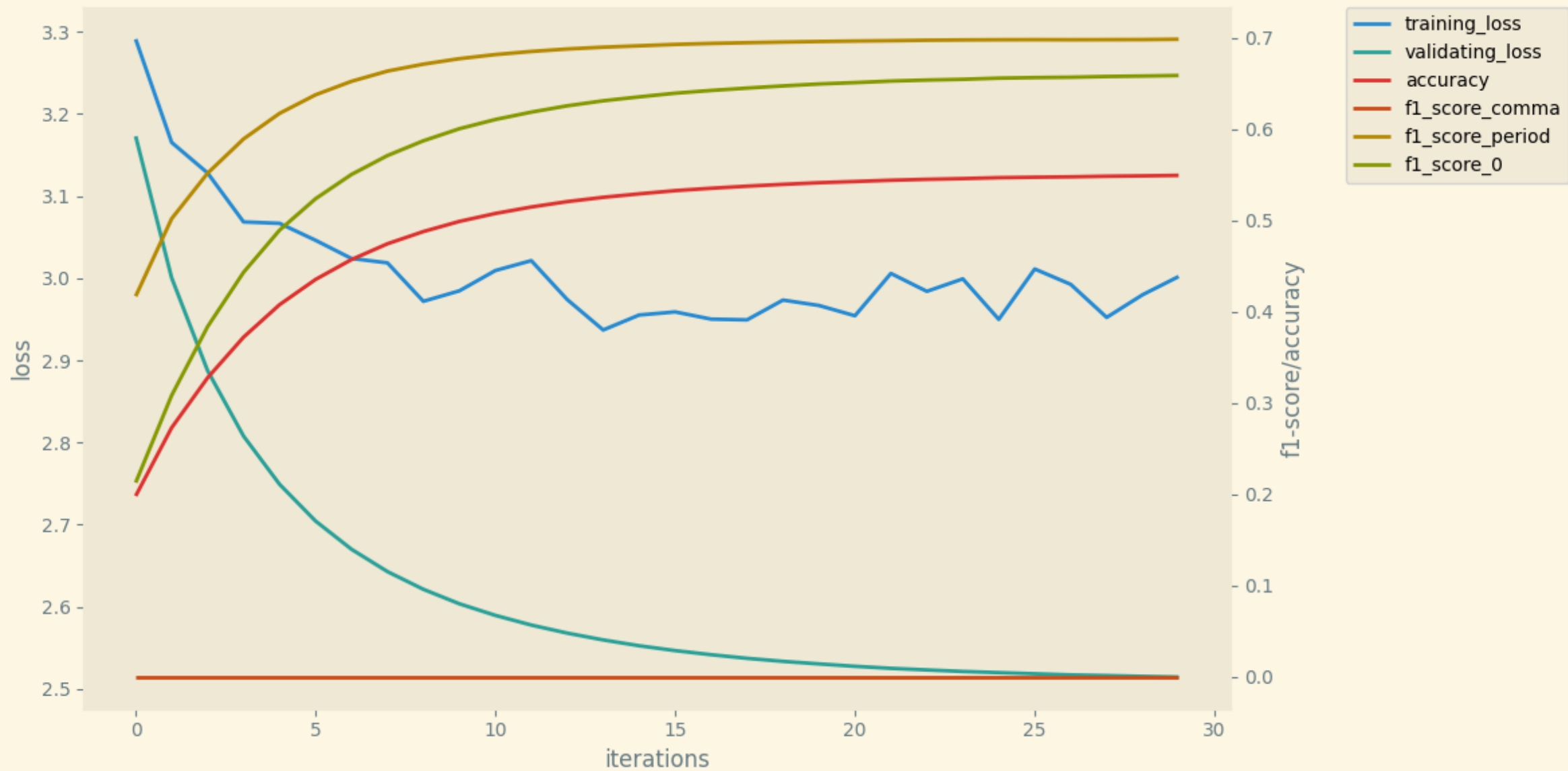
## PuncRec

BERT для татарского  
без выходного слоя

hidden\_layer =  
Linear(768,768)

output\_layer =  
Linear(768,punc\_num)





# Конвейер обработки данных

Необработанный  
текст

татарстан бүгенге  
көндә россиянең  
үзенчәлекле һәм  
мөстәкыйль төбәге  
без яңа эш алымнарын  
һәм заманча  
технологияләрне  
беренче чиратта  
нефть чыгару нефть  
эшкәртү һәм нефть  
химиясе машина төзү  
тармакларында һәм it  
өлкәдә кулланышка  
кертү буенча алдынгы  
урынны биләп торабыз

Токенизация

```
['татарстан', 'б',  
 '##ү', '##ген',  
 '##ге', 'к', '##өн',  
 '##дә', 'россия',  
 '##нең', 'ү',  
 '##зен',  
 '##чә', '##лек',  
 '##ле', 'һәм', 'м',  
 '##ө', '##ст',  
 '##эк', '##ый',  
 '##ль', 'т',  
 '##ө', '##б',  
 '##әге', 'без',  
 'яңа', 'э', '##ш',  
 'ал', '##ым',  
 '##нары', '##н',  
 'һәм', 'за',  
 '##ман', '##ча',  
 'технология',  
 '##ләрне',  
 'беренче', 'чи',  
 '##рат', '##та']
```

Идентификаторы  
токенов

```
[39675, 308, 14776,  
15462, 12998, 316,  
35864, 20610, 26375,  
22473, 359, 55371,  
73085, 28691, 11959,  
11374, 318, 20717,  
11784, 40768, 14738,  
12719, 324, 20717,  
12063, 77475, 13141,  
85887, 335, 11320,  
16804, 13716, 90658,  
10300, 11374, 10242,  
14186, 13291, 83865,  
64404, 60540, 17761,  
23486, 10376]
```

# Конвейер обработки данных

Ненормализованные  
результаты модели

```
[[[-0.1541, -0.3536, -0.1487],  
  [-0.1478, -0.3668, -0.1451],  
  [-0.1541, -0.3536, -0.1487],  
  ...],  
  [[-0.1567, -0.3509, -0.1480],  
   [-0.1547, -0.3546, -0.1458],  
   [-0.1478, -0.3668, -0.1451]],  
  [[-0.1522, -0.3515, -0.1441],  
   [-0.1522, -0.3515, -0.1441],  
   [-0.1519, -0.3519, -0.1428],  
   ...],  
  [[-0.1539, -0.3518, -0.1422],  
   [-0.1459, -0.3634, -0.1405],  
   [-0.1522, -0.3515, -0.1441]],  
  [[-0.1494, -0.3510, -0.1351],  
   [-0.1503, -0.3524, -0.1351],  
   [-0.1504, -0.3533, -0.1346],  
   ...],  
  [[-0.1446, -0.3639, -0.1302],  
   [-0.1494, -0.3510, -0.1351],  
   [-0.1523, -0.3515, -0.1325]]]
```


softmax и argmax

```
[0, 0, 2, 0, 0, 0,  
 1, 0, 0, 2, 0, 0,  
 0, 0, 2, 0, 0, 0,  
 0, 0, 0, 1, 0, 0,  
 2, 0, 0, 2, 0, 0,  
 0, 0]
```


Постобработка



Татарстан бүгенге  
көндә россиянең  
үзенчәлекле һәм  
мөстәкыйль төбәге.  
Без яңа эш  
алымнарын һәм  
заманча  
технологияләрне,  
беренче чиратта,  
нефть чыгару, нефть  
эшкәртү һәм нефть  
химиясе, машина  
төзү тармакларында  
һәм it өлкәдә  
кулланышка кертү  
буенча алдынгы  
урынны биләп  
торабыз


# Демо стенд


 **Hugging Face**


[Models](#) [Datasets](#) [Spaces](#) [Docs](#) [Solutions](#) [Pricing](#)



**Spaces:**  **xakep/TatPunRec** 

 like 0

 See logs

 Running

[App](#) [Files](#) [Community](#) [Settings](#)

## PuncRec - Punctuation and capitalization recovery for tatar language

This model trained using custom dataset of public Tatarstan government documents. It based on tatar BERT. PuncRec recovers punctuation and capitalization of ASR output for tatar (tt).

Языгыз, текст башка билгеләре препинания һәм баш хәрефләр

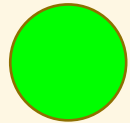
ОЧИСТИТЬ

Исполнить

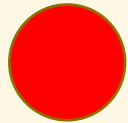
output



# Пример работы



Правильно



Знак стоит там где не должен  
Там где нужно знака нет



Неправильный знак

Закон белән йөкләнгән тикшерү күзәтү функцияләрен гамәлгә ашыру кысаларында бала хокуклары буенча вәкаләтле вәкил эшчәнлегендәге өстенлекле юнәлешләрнең берсе тикшерү гамәлләрен башкарудан гыйбарәт. Бала хокуклары буенча вәкаләтле вәкил законда билгеләнгән вәкаләтләр нигезендә, бала хокукларын һәм мәнфәгатьләрен бозу фактлары турындагы хәбәрләрне дә, татарстан республикасы дәүләт хакимияте органнарының, татарстан республикасында җирле үзидарә органнарының балалар хокукларын һәм мәнфәгатьләрен тәэмин итү эшчәнлеген башкаручы барлык оештыру хокукый рәвешләрендәге һәм милек рәвешләрендәге оешмаларның, аларның вазыйфаи затларының эшчәнлекләрен тикшереп торырга хокуклы. Тикшерүләрнең рәвешләренә килгәндә, Бала хокуклары буенча вәкаләтле вәкил тикшерүләрне мөстәкыйль яисә компетентлы дәүләт органнары һәм аларның вазыйфаи затлары белән берлектә уздырырга хокуклы. Тикшерүләр предметын хакимият органнарының һәм төрле балалар өлкәләрендә эшли торган оешмаларның һәм аларның вазыйфаи затларының, бер яктан, киңкырлы, икенче яктан, төгәл бер юнәлеше буенча үзенчәлекле эшчәнлекләре тәшкил итүне исәпкә алсаң, берләшкән тикшерүләрне оештыру тикшерү нәтижәләренең сыйфатлылыгына һәм дәрәҗәсизлегенә ирешергә мөмкинлек бирер иде.

# Заключение

- В ходе работы была разработана платформа для обучения нейронной сети и была выявлена возможность обучения нейросетевой модели восстановления знаков пунктуации.
- Для улучшения результатов работы модели, необходимо большее количество данных и усложнение архитектуры модели, что приведет к увеличению вычислительной сложности
- Данная модель передана в Институт прикладной семиотики при АН РТ для дальнейшего использования