ОБУЧЕНИЕ МОДЕЛИ BERT ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОДНОЗНАЧНОГО ОПРЕДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ В ТАТАРСКОМ ЯЗЫКЕ

Институт прикладной семиотики АН РТ

Республика Татарстан, Казань

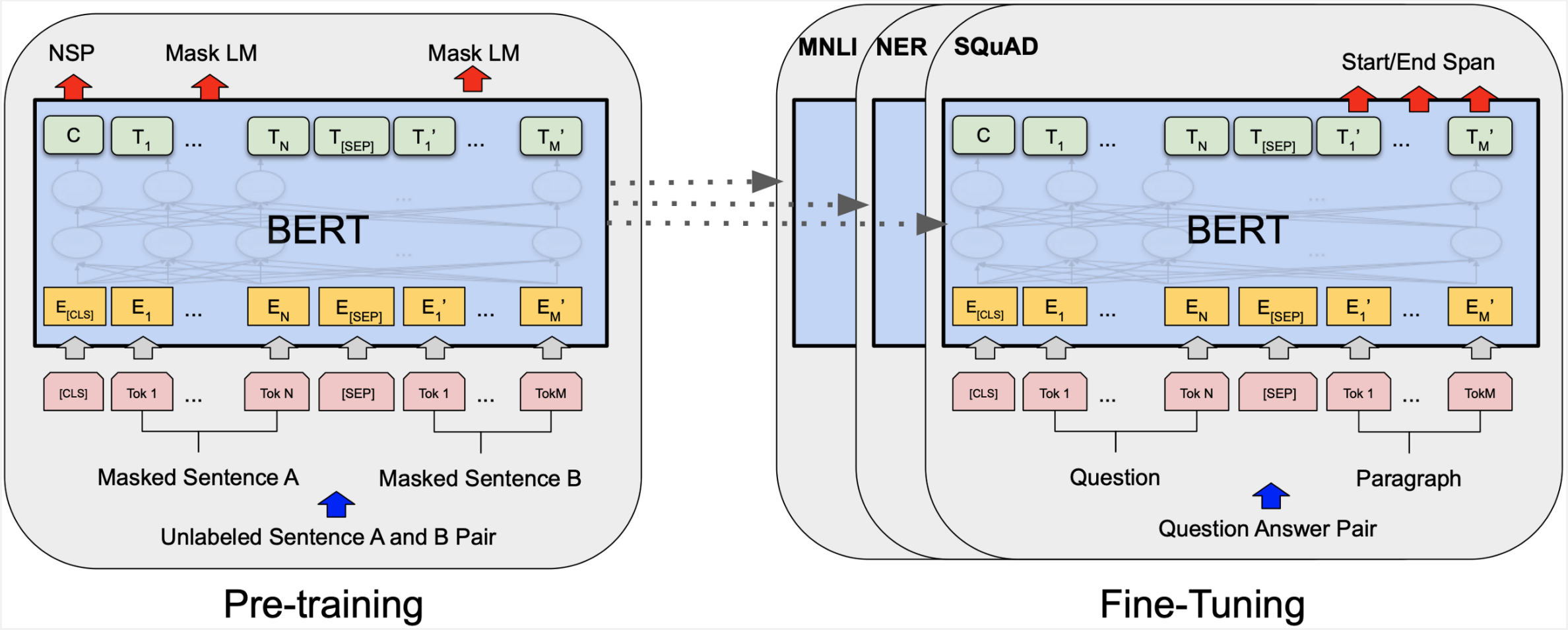
Хусаинов Айдар Фаилович

Насыров Айдар Айратович

ЦЕЛЬ

Создание инструмента для однозначного определения частей речи в предложениях на татарском языке при помощи нейросетевой модели BERT.

ПОЧЕМУ ИМЕННО BERT



ДАННЫЕ ДЛЯ ДООБУЧЕНИЯ

Татарстан республикасында күп фатирлы торак йортларга капиталь ремонт эшләре кырык муниципаль берәмлекнең барлык объектларында да алып барыла.

Егерме сигез муниципаль райондагы кырык алты мәктәптә капиталь ремонт эшләре төгәлләнгән.

Бу турыда бөтендөнья татар конгрессы матбугат хезмәте хәбәр итте

РЕЗУЛЬТАТЫ ДООБУЧЕНИЯ

Франциянең башкаласы булып [MASK] шәһәре тора

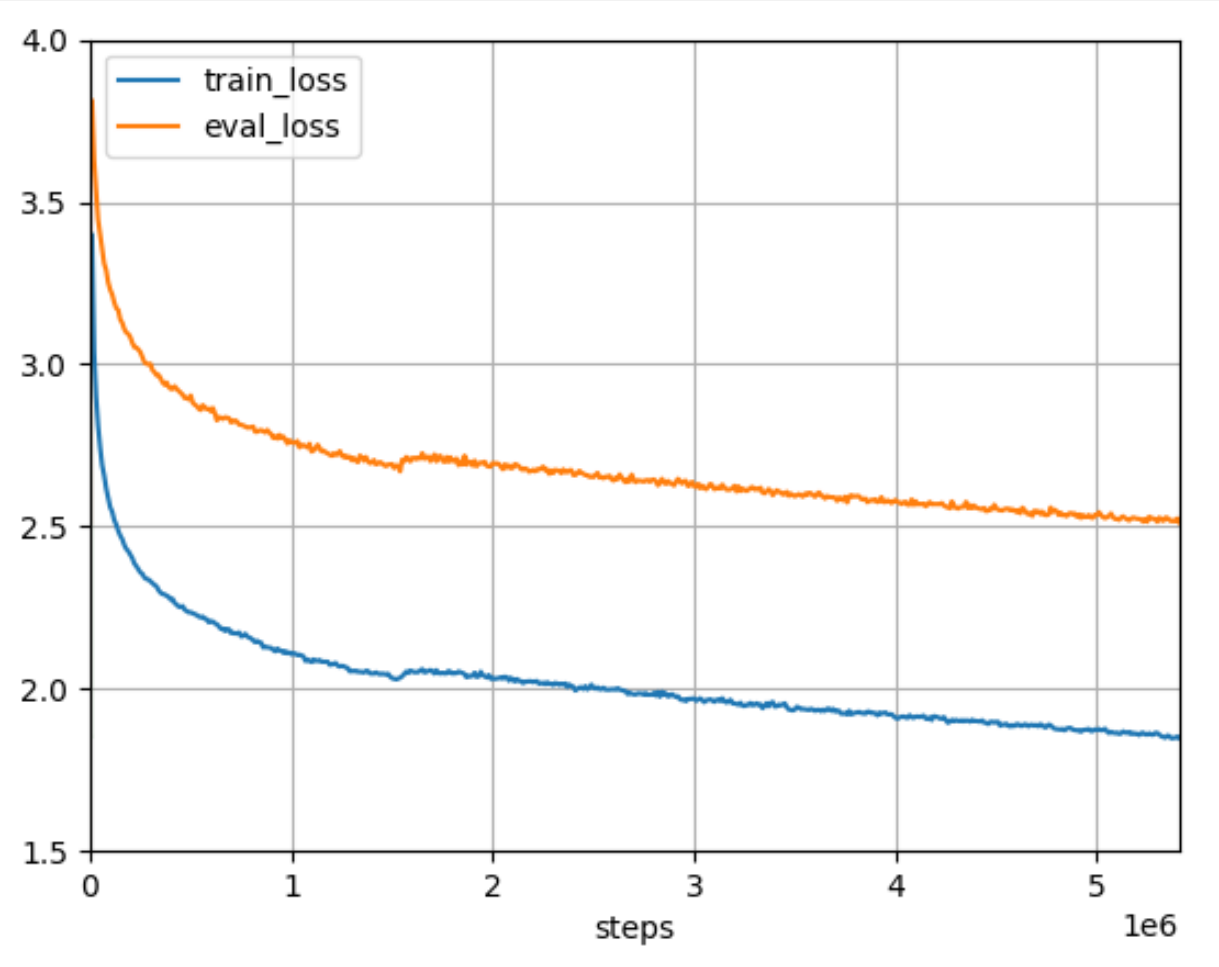
- Париж | 0.652
- Франция | 0.097
- Казан | 0.096
- Рим | 0.019
- Лондон | 0.012

РЕЗУЛЬТАТЫ ДООБУЧЕНИЯ

Мәгариф форумында катнашучыларны ТР Премьер-[MASK] Алексей Песошин сәламләде

- министр | 0.993
- ##ы | 0.004
- Ми | 0.001
- министрлыгы | 0.0
- ы | 0.0

ГРАФИК ДООБУЧЕНИЯ



ПОДГОТОВКА ДАННЫХ

- Поиск и удаление byte order mark symbols:
  - <0xEF>, <0xBB>, <0xBF>
- Замена и удаление символов, отсутствующих в BERT:
  - “ ” – - …
- Замена букв Й и Ё, представляющих собой два символа, а не один:
  - И + ̄
  - Е + ̈

ПОДГОТОВКА ДАННЫХ

- Сократили число объектов классификации до 16 единиц:

'PROP', 'N', 'NR', 'V', 'PUNCT', 'PN', 'PART', 'CNJ', 'MOD', 'Adv', 'POST', 'Adj', 'Num', 'INTRJ', 'IMIT', 'ADV'

РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Предложение:

- ['Бу', 'эш', 'киләсе', 'Экология', 'елында', 'да', 'дәвам', 'итәчәк', '.’]

True:

- ['PN', 'N', 'V', 'N', 'N', 'PART', 'N', 'V', 'PUNCT’]

Predicted:

- ['PN', 'N', 'V', 'N', 'N', 'PART', ‘PN', 'V', 'PUNCT’]

Accuracy = 0.86

РЕЗУЛЬТАТЫ

Высокая точность полученной модели позволяет с уверенностью сказать об успехе применения BERT для решения задачи однозначного определения частей речи в татарском языке, а также возможности использования дообученной модели для задачи снятия многозначности при морфоанализе слов.