1. Тема на слайде
2. Как правило результатом работы алгоритмов распознавания речи является текст без знаков препинания в нижнем регистре
3. Но зачем нам нужна пунктуация? Пунктуация улучшает восприятие текста: текст со знаками препинания читается быстрее; так же такой текст понятнее, можно вспомнить крылатую фразу «казнить нельзя помиловать», где всего одна запятая критически влияет на семантику фразы. Так же пунктуация положительно влияет на работу алгоритмов обработки естественного языка.
4. Цель работы – реализовать алгоритм восстановления пунктуации для татарского языка.
5. Существует два основных подхода решения данной задачи: применение акустических моделей, использующих параметры голоса говорящего, и использование языковых моделей, для обработки текста на естественно языке. В работе использована модель BERT для татарского языка, предоставленная институтом семиотики, данная модель отличается тем, что ее ненужно обучать с нуля, что экономит время и ресурсы.
6. Из рабочих аналогов можно выделить сервис Google для распознавания речи, а также библиотеку NeMO для обработки языка, оба аналога не поддерживают татарский язык.
7. Модель и алгоритм расстановки пунктуации разрабатывались с помощью python. Среди инструментов стоит отметить PyTorch – фрейморк для работы с нейросетями. А также CUDA технология ускорения вычислений с помощью видеокарт Nvidea.
8. Так как BERT изначально обучается для предсказания замаскированного слова в предложении, то необходимо было его модифицировать, удалив выходной слой и добавив кастомные скрытые и выходные слои.
9. Модель обучалась на корпусе публичных документов, предоставленном институтом семиотики. По метрикам можно сказать, что модель неплохо обучилась.
10. Основной алгоритм восстановления пунктуации состоит из конвейера преобразования данных. Исходный текст перед тем, как попасть в модель проходит стадию токенизации.
11. Модель же выдает логиты, которые преобразуются в идентификаторы классов знаков пунктуации, и дальше происходит расстановка знаков препинания и где надо меняется регистр
12. Для демонстрации работоспособности модель была залита на HuggingFace Spaces
13. Что бы сэкономить время, я разобрал пример обработки.
14. В ходе работы был создан алгоритм восстановления пунктуации и регистра, обучена модель. Для дальнейшего улучшения качества работы модели необходимо больше данных и усложнение архитектуры модели. Модель и платформа обучения были переданы в институт прикладной семиотики