Сегодня автоматическое распознавание речи используется в различных областях жизнедеятельности человека, от умных ассистентов до средств ввода текста голосом. Так как мы свою речь явно не включаем пунктуацию, проблемой большинства алгоритмов распознавания речи является потеря пунктуации. (слайд 3) Это в свою очередь приводит к усложнению чтения полученного текста, по некоторым исследованиям текст без пунктуации сложнее понять, чем текст, в котором 20% слов написаны с опечатками. Таким образом наличие пунктуации критично если результат будут читать люди или будет применяться дальнейшая обработка текста.

(слайд 4) Одним из решений проблемы является применение моделей глубокого обучения.Прорывом в области, безусловно являются предобученные модели на основе трансферного обучения для решения различных задач NLP. Предобученные модели позволяют пропустить самый сложный этап первоначального обучения, и заняться точной настройкой модели (fine-tuning) под конкретную задачу. Одним из таких моделей является BERT.

Существует 2 версии Bert: базовая и расширенная. Базовая версия состоит из 12 слоев кодировщиков, расширенная из 24. Данные кодировщики являются одним из компонентов архитектуры Google Transformers. Данная архитектура раньше использовалась для кодирования последовательности данных во время машинного перевода. Bert использует кодировщики для двунаправленного контекстуализирования слова во вложения, то есть числовые векторы представления слов. Каждый слой пытается закодировать текст в свой контекст. Входной текст представляется в виде токенов. Каждый токен – это отдельное слово или метка. Входной текст представляется в виде токенов. Каждый токен – это отдельное слово или метка. Метки бывают 3 видов: [sep],[cls],[pad]. Первая метка обозначает конец предложения, служит для нужд точной настройки классификации, вторая для обозначения вектора для классификации NSP, а третья – позволяет уравнять входные порции текстов до единой длины. Каждый слой Bert пытается закодировать текст в свой контекст. Задача последнего слоя объединить контексты всех слоев. Таким образом на выходе мы можем получить несколько вложений в зависимости от настройки.

Для точной настройки Bert под конкретную задачу создаётся дополнительный слой нейронной сети, который на вход получает вложение. В дальнейшем происходит обучение финального слоя, сама же NLU модель не переобучается.