Здравствуйте, мы выбрали тему обработка естественного языка (POS, NER).

Для решения этой задачи мы решили использовать syntaxnet. Это основанная на TensorFlow библиотека определения синтаксических связей, использующая нейронную сеть. В настоящий момент поддерживается 40 языков, в том числе и Русский. Плюсы syntaxnet это конечно же базирование на гугловском TensorFlow. Syntaxnet уже содержит натренированную модель нейронной сети "Parsey McParseface", пригодная для разбора текста. Код открыт под лицензией Apache 2.0. Точность работы модели "Parsey McParseface" оценивается Google в 94%. Производительность SyntaxNet позволяет обрабатывать приблизительно 600 слов в секунду на обычном настольном компьютере. В качестве сопутствующих инструментов предоставлены средства для анализа лингвистической структуры предложений или высказываний, показывающие функциональную роль каждого слова.

Syntaxnet устанавливается под unix системы, для windows можно использовать докер. Для этого необходимо установить его и скачать образ SyntaxNet. Удобство подхода заключается в том, что в образе SyntaxNet уже настроенный, так же установлены все необходимые зависимости. Далее приведён пример докерфайла, с помощью которого можно запустить контейнер.

Для анализа мы использовали текст на русском языке, взятый с Universal Dependencies. UD это кросс лингвистический проект. Они можно сказать размечают предложения чтобы их потом можно было использовать для парсеров, обучений, исследований и так далее. Чтобы проанализировать наш текст был написан скрипт, который основан на demo скрипте, включённым в образ syntaxnet. Пример приведён на слайде. Скрипты можно запускать не только для файлов но и для отдельных предложений, передавай их в качестве аргумента.

В результате работы скрипта получаем размеченные слова. Формат вывода conll. Пример этого формата приведён на слайде. Наибольший интерес представляют следующие колонки:  
1. порядковый номер слова в предложении,  
2. слово (или символ пунктуации),  
4. часть речи  
7. номер родительского слова (или 0 для корня).

Используя номер родительского слова можно построить граф зависимостей в предложении.

Также с помощью Python были проанализированы полученные данные. Для анализа использовались три основных лингвистических параметра POS LAS UAS. Что это за параметры????