УДК 004.934.1'1

Автоматическое восстановление знаков препинания в задаче распознавания речи

Митрофанов А.А.
(Федеральное государственное автономное образовательное учреждение высшего образования “Национальный исследовательский университет ИТМО”),
Научный руководитель – к.т.н., ведущий н.с., Романенко А.Н.
(ООО “ЦРТ-Инновации”)

В данном докладе рассматривается технология автоматического восстановления знаков препинания с помощью предобученной многофункциональной языковой модели BERT (Bidirectional Encoder Representations from Transformers) в задаче автоматического распознавания речи.

Введение. Системы автоматического распознавания речи имеют широкое применение в самых разных сферах человеческой жизни. Качество расстановки знаков препинания в результатах распознавания напрямую влияет на то, как человек воспринимает и оценивает распознанные тексты. Знаки препинания позволяют разделить текст на осмысленные предложения, делая его более читаемым, а также помогают при дальнейшем анализе текста. В последнее время популярны многофункциональные нейронные языковые модели, которые показывают себя хорошо в большом количестве задач, в том числе и в задаче восстановления пунктуации. Наилучшие известные на данный момент решения для восстановления пунктуации на английском языке - это системы, построенные на базе предобученной языковой модели BERT.

Основная часть. Языковая модель BERT обучается выполнять одновременно несколько разных задач. Это позволяет модели извлекать информацию, полезную для задач разного рода. За счет этого такая языковая модель хорошо адаптируется в том числе к новым задачам, не участвовавшим в обучении.
Задача восстановления знаков препинания может быть сформулирована в терминах классификации. Для каждого слова из входной последовательности требуется предсказать его класс, определяющий, идет ли за ним знак препинания, и если да, то какой. Такая классификация в данной работе была реализована путем дообучения модели BERT на новую задачу. Для этого в модель были добавлены несколько новых слоев, предсказывающих класс каждого слова.
Поскольку BERT учится обобщать информацию из входных текстовых данных, основанная на нем система восстановления пунктуации является устойчивой к ошибкам распознавания речи. Это позволило обучить систему хорошего качества без использования целевых данных (результатов распознавания), а только на обычных текстовых данных, которые значительно легче получить.

Выводы. Представленный метод позволяет построить систему автоматической расстановки знаков препинания для результатов распознавания речи на русском языке, используя только открытые источники данных, и существенно уменьшает ошибку восстановления пунктуации по сравнению с классическими подходами, такими как условное случайное поле (CRF).

Митрофанов А.А. (автор)	Подпись	Романенко А.Н
Романенко А.Н. (научный руководитель)	Подпись	Митрофанов А.А