УДК 004.8

Автоматическая расстановка знаков препинания в распознанном тексте методами машинного обучения

Тирских Д.В. (Иркутский государственный университет)
Научный руководитель – доцент, к.ф.-м.н., Бернгардт О. И. (Института солнечно-земной физики СО РАН)

В данной работе предлагается один из вариантов решения проблемы восстановления пунктуации в неструктурированном тексте. Задача сводится к задаче мультиклассовой классификации — прогноза знака препинания между двумя словами с учетом возможных слов-соседей.

Введение. Задача распознавания человеческой речи в настоящее время решается различными способами. Многие современные алгоритмы распознавания, однако, зачастую не позволяют расставлять знаки препинания. Характерным примером такого сервиса является Youtube, который способен распознавать произносимый на видео текст, но не расставляет знаки препинания. Тем не менее, расстановка знаков препинания оказывается важной при создании различных систем записи речи — от стенографирования длинных текстов до исправления знаков препинания в уже написанном тексте.

Основная часть. Задача решается методом переноса знаний от обобщенной языковой модели и построением оптимального классификатора, прогнозирующего необходимый знак препинания. Задача решается с учетом сильного дисбаланса классов — различные знаки препинания встречаются в текстах с различной частотой. С использованием современных методов машинного анализа текстов и обобщенных моделей языка получается предварительное решение задачи разбиения распознанного голосового текста на предложения, и расстановки знаков препинания в тексте. Использование lданного подхода в теории поможет разрешать ситуации, где постановка знаков препинания не однозначна и зависит от контекста. С такими случаями не справляются современные методы восстановления пунктуации, основанные на прасинге текста и системах правил и словарей. В работе изложен метод подготовки обучающего датасета, выбор обобщенной модели языка, архитектура такой сети, и процесс ее обучения с использованием переноса знаний (методом извлечения признаков). Обсуждается достигаемое качество расстановки знаков препинания.

Выводы. Применение методов машинного обучение позволило получить одно из возможных решение поставленной задачи. Результаты работу могут быть применены для улучшения работы систем автоматического распознания речи и систем машинного перевода.

Тирских Д.В. (автор) Подпись
Бернгардт О.И. (научный руководитель) Подпись