УДК 004.934.1'1

ПОСТРОЕНИЕ МУЛЬТИЯЗЫЧНОЙ СИСТЕМЫ ВОССТАНОВЛЕНИЯ ЗНАКОВ ПРЕПИНАНИЯ

Митрофанов А.А., Митрофанова М.М.
(Университет ИТМО)
Научный руководитель – к.т.н., Романенко А.Н.
(ООО “ЦРТ”)

В данном докладе описывается процесс построения мультиязычной системы восстановления знаков препинания с помощью предобученной многофункциональной языковой модели BERT (Bidirectional Encoder Representations from Transformers) [1].

Введение. Знаки препинания являются неотъемлемой частью естественных языков. Они напрямую влияют на то, как человек воспринимает и оценивает текст. Большинство естественных языков имеют схожие наборы знаков препинания и правила их расстановки. Этот факт позволяет построить мультиязычную систему восстановления знаков препинания, успешно перенося знания из одного языка в другой. Для решения задачи мультиязычного языкового моделирования в последнее время лучше всего показывают себя предобученные модели BERT. Используя такую архитектуру в качестве основы, можно обучить мультиязычную модель, способную корректно расставлять знаки препинания на большом количестве языков [2].

Основная часть. Во время обучения языковая модель BERT обучается сразу на несколько различных задач. Такой подход к обучению модели обеспечивает ее высокую обобщающую способность, за счет чего такая языковая модель хорошо адаптируется к новым задачам.
Задача восстановления знаков препинания может быть сформулирована в терминах классификации слов. Для каждого слова из входной последовательности требуется предсказать его класс, определяющий знак препинания на границе слова. Такая классификация может быть реализована путем дообучения модели BERT на новую задачу. Для этого в модель добавляются новые выходные слои, предсказывающие класс “знак пунктуации” для каждого слова.
Поскольку BERT обучается на большом количестве текстовых данных различных языков, такую модель можно успешно применять для построения мультиязычных систем восстановления знаков препинания. Использования BERT-модели позволяет построить систему хорошего качества без использования большого количества текстовых данных для каждого языка.

Выводы. Представленный метод позволяет построить мультиязычную систему восстановления знаков препинания в тексте с использованием лишь открытых источников данных. Полученная система успешно справляется с расстановкой знаков препинания на таких языках, как русский, английский, французский, казахский, испанский.

Список использованных источников:

1. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // ArXiv. — 2019. — Vol. abs/1810.04805
2. Nagy A., Bial B., Ács J. Automatic punctuation restoration with BERT models // ArXiv. — 2019. — Vol. abs/2101.07343