Морфологическая сегментация для предобученных маскированных языковых моделей

Березникер Алексей Витальевич

Студент (магистр)

Факультет ВМК МГУ имени М.В.Ломоносова, Москва, Россия

E-mail: berezniker@mail.ru

Научный руководитель *— Большакова Елена Игоревна*

В настоящее время всё чаще решение прикладных задач обработки естественного языка опирается на использование языковых моделей, которые способны оценивать вероятность различных языковых единиц так, чтобы эти вероятности отражали знание языка. В отличие от классической статистической языковой модели, оценивающей вероятность очередного слова предложения по предыдущим, маскированная языковая модель оценивает вероятность слова с учетом его различных контекстов. Для обработки текстов в маскированных языковых моделях используются алгоритмы сегментации слов на подслова и словари фиксированного размера часто встречающихся слов и подслов (токенов). Поскольку сложные и редкие слова могут отсутствовать в корпусе текстов, на которых обучаются языковые модели, они могут быть не представлены в словаре токенов, что может привести к их некорректной обработке и ухудшению итогового качества решения.

Наиболее используемой маскированной языковой моделью является BERT, состоящая из стека кодировщиков нейросетевой архитектуры трансформер и использующая алгоритм WordPiece сегментации на подслова. В работе [2] показано, что для английского языка в модели BERT редкие и сложные слова, как правило, сегментируются морфологически некорректно, т.е. граница разбиения слова на подслова проходит не между морфемами – минимальными значащими единицами языка. Например, слово «overseasoned» (перегруженный) сегментируется на подслова как «overseas|oned», в то время как морфологически корректной сегментацией является «over|seasoned». Авторами работы предложен подход к словообразовательной сегментации слов текста без изменения словаря токенов, использующегося в модели BERT, и экспериментально показано улучшение качества интерпретации редких и сложных слов английского языка.

Данная работа посвящена применению морфологически корректной сегментации слов русского языка для предобученных маскиро-

ванных языковых моделей BERT, исследовалось влияние сегментации слов на качество решения задачи тематической классификации на два класса. В качестве предобученных языковых моделей рассматривались BERT Base модели: Sber RuBERT и DeepPavlov RuBERT, отличающиеся обучающим корпусом текстов русского языка и используемым алгоритмом сегментации. Для получения морфологически корректной сегментации слов использовался морфологический процессор CrossMorphy [1], который, в том числе, выполняет сегментацию слова русского языка на морфемы с классификацией по типам (корень, префикс, суффикс, окончание), например, «изо|метр|и|я», «терм|о|регул|яци|я», «микро|био|лог|и|я».

В работе предложен способ замены обычной сегментации, используемой в рассматриваемых языковых моделях, на морфологически корректную сегментации слов на морфемы с использованием встроенного словаря токенов этих моделей. Для проведения экспериментов по оценке влияния предложенного способа на качество бинарной тематической классификации был автоматизированно построен набор данных (датасет), состоящий из сложных и редких слов, встречающихся в русскоязычных текстах из двух предметных областей: математика и биология. Каждая из этих тематик в построенном наборе данных представлена 1200 уникальными словами, примерами таких слов являются «аксонометрия», «аддитивность» (математика) и «авитаминоз», «мутагенез» (биология). В качестве классификатора были опробованы однослойные линейные нейронные сети с разными функциями активации. Проведенные эксперименты продемонстрировали значимое улучшение качества бинарной тематической классификации слов по метрике ROC AUC: с 0.798 до 0.837 у модели Sber RuBERT (с BPE сегментацией), с 0.791 до 0.809 у модели DeepPavlov RuBERT (с WordPiece сегментацией). Таким образом, морфологически корректная сегментация слов может быть использована для улучшения качества работы систем классификации на базе языковой модели BERT.

Литература

1. Bolshakova E. I., Sapin A. S. A morphological processor for Russian with extended functionality //Analysis of Images, Social Networks and Texts: 6th International Conference, AIST 2017, Moscow, Russia, 2017, Revised Selected Papers. – С. 22-33.
2. Hofmann V., Pierrehumbert J. B., Schütze H. Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words //arXiv preprint arXiv:2101.00403. – 2021.