**Использование машинного обучения для определения уровня сложности учебных текстов на русском языке по CEFR**

Лимонова Анастасия Андреевна

Студентка Московского государственного университета имени М. В. Ломоносова, Москва, Россия

Определение уровня сложности русскоязычного текста представляется важной задачей для преподавания русского языка как иностранного. В современном мире не возникает проблемы недостатка текстовых материалов, однако не всегда легко определить их сложность. Так как классификация текстов по уровням CEFR (Общеевропейские компетенции владения иностранным языком) достаточно формализована и описана в методических пособиях, существуют сервисы, помогающие автоматизировать этот процесс.

Онлайн-сервис [Текстометр] применяется для анализа сложности текстов на русском языке, используя шкалу CEFR (от A1 до C2). В его основе лежит регрессионная модель, в которую подаётся вектор закодированных лингвистических признаков текстов. Особенности линейных моделей в том, что они более интерпретируемы, чем нейронные сети, однако, они чаще выдают худшие результаты.

Гипотеза нашего исследования заключалась в том, что использование трансформерных моделей поможет превзойти качество работы Текстометра. Однако, вопреки ожиданиям, трансформеры показывали даже более низкую эффективность обучения на [датасете RuFoLa]. Корпус RuFoLa, содержащий учебные материалы по русскому языку как иностранному. Хотя корпус изначально охватывал все уровни CEFR (A1–C2), ввиду ограниченного количества текстов C2 (всего 6) мы исключили этот уровень из нашего анализа, оставив 5 классов (A1–С1).

Суть работы заключается в анализе сложности задачи определения уровня сложности и проблем датасета RuFoLa. Возможной причиной является неоднородность датасета: тексты одного уровня могут включать отрывки из художественных произведений, учебные описания, а также публицистику, что создаёт противоречивые данные для модели. [Сергей Шаров] предполагает, что предварительная классификация текстов по жанрам и стилям перед оценкой сложности могла бы повысить точность предсказаний.

Кроме того, одной из ключевых метрик сложности текста является лексическая сложность, которая традиционно оценивается с использованием лексических минимумов. Однако такие списки часто не учитывают словообразовательные дериваты, что может снижать их эффективность в предсказательных моделях.

Таким образом, результаты исследования показали, что применение трансформерных моделей к имеющемуся датасету не повысило точность определения уровня сложности текстов на русском языке.

В качестве возможного направления для улучшения модели предлагается комбинированный подход: объединение признаков Текстометра с выходами трансформерной модели, а также в исследовании влияния жанровой и стилистической принадлежности на сложность текста. Также, перспективной кажется идея использовать не модели классификации для предсказания класса, а модели регрессии для предсказания числового значения – величины сложности текста. Мы надеемся, что это позволит точнее учитывать лексическую сложность текста и повысить точность определения уровня.

**Литература**

*Laposhina A.* A corpus of Russian textbook materials for foreign students as an instrument of an educational content analysis. Russkii iazyk za rubezhom. 22-28. 2021

*Лапошина А.Н., Лебедева М.Ю*. Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному. Русистика. №3. 2021.

*Sharoff, Serge.* What neural networks know about linguistic complexity. Russian Journal of Linguistics. 371-390. 2022