Агрегирование квантильных моделей для оценивания когнитивной сложности текста

Веселов Арсений Сергеевич

Студент

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

E-mail: veselovas@my.msu.ru

Научный руководитель *— Воронцов Константин Вячеславович*

Для задачи оценки сложности текста было разработано множество индексов удобочитаемости [1]. Большинство из них использует в своей основе линейную комбинацию некоторых статистических параметров текста с лексического и синтаксического уровней. Оценки сложности текста имеют множество приложений: анализ юридических документов [2], составление текстов инструкций к лекарствам, препаратам и техническим средствам. Многие индексы используются для определения понятности учебной литературы, предлагаемой учащимся разных лет обучения [3]. Использование оценки сложности текстов может быть полезно для прогнозирования временных затрат на обработку документов, нормативных актов и учебной литературы.

В данной работе реализуется основанный на квантильном подходе метод оценивания когнитивной сложности текста, впервые представленный в 2019 году [4]. При таком подходе используется референтный корпус текстов, на основе которого создаются модели оценки сложности текста для морфологического, лексического и синтаксического уровней языка. Эти модели оценивают сложность текста как нормированную взвешенную сумму сложностей аномально сложных токенов. Сложность токена считается аномально высокой, если она превышает квантиль эмпирического распределения сложности. Для вычисления сложности отдельного токена в моделях используются частотные (на основе расстояний между одинаковыми токенами) и сложностные (на основе структуры самого токена) функции сложности. Затем с помощью линейной модели агрегируются сложности, выдаваемые моделями с отдельных уровней языка. Выдвигается предположение, что такая агрегированная модель может показывать более высокое качество ранжирования пар текстов по их когнитивной сложности по сравнению с индексами удобочитаемости.

В качестве референтного корпуса для экспериментов использовалась русскоязычная Википедия. Для подготовки наборов данных

для валидации результатов сравнения были использованы учебники по обществознанию. В результате экспериментов агрегированная модель показала качество выше, чем у индексов удобочитаемости.

Индекс	**Точность, %**
FKGL	88.4
FRES	89.9
CLI	87.9
SMOG	86.7
ARI	88.8
LIX	87.8
TI	88.6

Таблица 1: Результаты индексов удобочитаемости на валидации

Точность агрегированной модели на валидации: **92.8%**.

Литература

1. Оборнева И. В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров. дис. ... канд. пед. наук. М., 2006. С. 165.
2. Дмитриева А. В. «Искусство юридического письма»: количественный анализ решений Конституционного Суда Российской Федерации. Сравнительное конституционное обозрение. 2017. Вып. 3 (118). С. 125–133.
3. Solovyev V. D., Ivanov V. V., Solnyshkina M. I. Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics. // Journal of Intelligent & Fuzzy Systems. 2018. 34 (5). P. 3049–3058.
4. Eremeev M. A., Vorontsov K. V. Quantile-based Text Complexity Measure. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2020». Moscow, 2020.