

0.1. Смаль И.А., Морозов Д.А. Анализ значимости синтаксических признаков текста при оценке его удобочитаемости

Задача определения удобочитаемости текста, то есть того, насколько текст сложен для понимания в процессе прочтения, имеет широкую область применения. Однако, несмотря на это, а также на активное развитие инструментов для анализа естественного языка в последние десятилетия, до сих пор, общепринятыми методами оценки удобочитаемости текста являются формулы, предложенные во второй половине 20-го века.

Современные исследования в данной области обычно включают в себя использование машинного обучения и анализ различных признаков текста, чаще всего лексических и синтаксических. Причем, в большинстве работ лексические признаки показывают себя лучше, чем синтаксические. Одной из возможных причин этого является недостаточное количество исследований того, на какие именно синтаксические признаки стоит обращать внимание при оценке удобочитаемости. Именно этот пробел мы и решили восполнить в данной работе.

Нами были проведены эксперименты по оценке удобочитаемости групп предложений на основе их синтаксических признаков с применением методов градиентного бустинга и случайного леса. Эксперименты проводились на двух корпусах – корпусе учебников [1] и корпусе художественной литературы [2]. В процессе экспериментов размеры групп варьировались для получения более достоверных результатов. В обоих корпусах присутствовала разметка, сопоставляющая тексту класс школы, ученики которого без проблем могли понимать этот текст.

Всего выделялось 248 синтаксических признаков для каждой группы предложений. С помощью четырех методов определения значимости признаков в моделях (*Mutual information*, *drop-column importance*, *permutation importance*, *Mean decrease in impurity*) были найдены самые значимые признаки. Для корпуса учебников таких признаков оказалось 12, а для художественной литературы 17, при этом всего 5 из них были общими. Модели, обученные исключительно на лучших признаках, показали себя лишь незначительно хуже (а на корпусе художественной литературы даже лучше), чем модели, использовавшие все 248 признаков.

Научный руководитель — д.т.н. Баранкин В. Б.

Список литературы

- [1] SOLOVYEV V., IVANOV V., SOLNYSHKINA M. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. // Journal of Intelligent & Fuzzy Systems. Vol. 34. N. 2. P. 1–10.
- [2] МОРОЗОВ Д. А., ГЛАЗКОВА А. В., ИОМДИН Б. Л. Сложность текста и лингвистические признаки: как они соотносятся в русском и английском языках // Russian Journal of Linguistics T. 26. № 2. С. 426–448.