

0.1. *Обершт С.Д.* К задаче оценки сложности текста методами машинного обучения: аналитический обзор, выводы, начальные эксперименты

Оценка сложности текста в задачах компьютерной лингвистики может трактоваться как классификация текста в зависимости от уровня его восприятия; традиционно текст градуируется по количеству лет академического обучения, необходимого для понимания и усвоения информации, а также по ряду специфических признаков. В работе дан аналитический обзор существующих на сегодняшний день подходов и моделей оценки сложности текста, разработанных или адаптированных для русского языка, а также описан ряд задач компьютерной лингвистики, в которых используется этот параметр. Применение методов машинного обучения для оценки сложности позволяет значительно повысить точность результатов по сравнению со статистическими методами, в том числе, с использованием формул читабельности (Флеша, Флеша-Кинкейда, тест SMOG и т.п.) [1]. Применение методов глубокого обучения показывает результаты, превосходящие данные традиционных классификаторов, что отражено в ряде актуальных работ [2,3]. При этом сложность текста анализируется не только как объект с точки зрения применимости методов компьютерной лингвистики, но и выступает в качестве оценочной метрики в задачах обработки естественного языка: сравнение естественного и сгенерированного текстов, оценка качества результатов больших языковых моделей, анализ систем вопрос-ответ.

Проведенные авторами данной работы начальные эксперименты по сравнительному анализу самых используемых формул читабельности текста и их адаптированных к русскому языку версий, показали, что эти формулы не могут считаться потенциально применимыми к текстам различных жанров и стилей, а также к текстам, написанным на языках, отличных от тех, для которых модели первоначально разработаны. Использование индексов как группы параметров без предварительной адаптации к набору данных модели может дать критичную для объективного результата погрешность; при этом в задаче оценки сложности должны учитываться как качественные, так и количественные параметры: морфологические, синтаксические, лексические, и параметры, связанные с дискурсом [4].

Научный руководитель — д.т.н., к.филол.н. Гавенко О. Ю.

Список литературы

- [1] LYASHEVSKAYA O., PANTELEeva I., VINOGRADOVA O. Automated assessment of learner text complexity // *Assessing writing*. 2021. Vol. 49. N. 100529.
- [2] BLINOVA O., TARASOV N. A hybrid model of complexity estimation: Evidence from Russian legal texts // *Front. Artif. Intell.* 2022. Vol. 5. N. 1008530.
- [3] LEE B.W., JANG Y.S., LEE J. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021. N. 10669–10686.
- [4] FENG L. ET AL. A comparison of features for automatic readability assessment // *Coling 2010: Posters*. 2010. P. 276–284.