УДК 81’42

М. А. Донцов
КФУ, г. Казань

АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ СЛОЖНОСТИ УЧЕБНЫХ ТЕКСТОВ: ИСТОРИЯ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

В последнее время учеными активно изучается проблема определения сложности текста и прежде всего с использованием автоматизированных программ-анализаторов. Также исследуется вопрос построения внутреннего механизма и этапов работы автоматизированных анализаторов. В статье рассматривается история и перспективы развития автоматизированного анализа сложности учебных текстов.

Автоматизированный анализ сложности текста является одним из наиболее актуальных направлений в области обработки естественного языка (Natural Language Processing, NLP). Автоматизированный анализ сложности текста имеет длинную историю развития, начиная от первых попыток автоматизации обработки естественного языка до современных систем, способных обрабатывать огромные объемы текстовых данных [5]. Рассмотрим основные этапы этой истории и их значение для современной технологии машинного обучения.

Первые шаги в развитии автоматизации анализа сложности текста были предприняты еще в 1960-х годах. Одним из первых проектов была разработка системы, которая использовала статистические модели для определения вероятности появления определенных слов в тексте. Однако, этот проект был неудачным из-за ограничений в вычислительных возможностях того времени. Следующий важный шаг был сделан в 1970-х годах, когда появились первые коммерческие системы машинного перевода. Эти системы использовали статистический подход для определения вероятности правильного перевода слов и выражений. Однако эти системы все еще имели некоторые ограничения, связанные с точностью и полнотой перевода. В 1980-х годах появилась новая технология машинного обучения, называемая «глубокими нейронными сетями». Эта технология позволила создавать более точные модели, способные автоматически определять значимость каждого слова в тексте. Это привело к появлению новых возможностей в автоматическом анализе сложности текста, таких как выделение ключевых понятий и связей между ними.

Однако развитие технологий машинного обучения было ограничено недостаточным количеством доступных данных для обучения моделей. Поэтому в 1990-х годах появился новый подход к анализу сложности текста, называемый «обучением с учителем». Этот подход использует данные, полученные из других источников, чтобы обучить модель на основе этих данных.

Сегодня автоматизированный анализ сложности текста является одной из самых популярных областей машинного обучения. Системы, способные обрабатывать огромные объемы текстовых данных, стали неотъемлемой частью многих приложений, включая обработку естественного языка, компьютерное зрение и голосовые помощники.

В автоматизированном анализе сложности текста выделяют несколько этапов.

Первым этапом является предварительная обработка текста. На нем происходит удаление стоп-слов (часто встречающихся слов, которые не несут смысловой нагрузки), лемматизация (приведение слова к его базовой форме) и токенизация (разделение текста на отдельные слова или токены). Также может проводиться стемминг (приведение слов к их базовой форме без учёта регистра) для улучшения качества распознавания речи [1].

Далее следует этап анализа синтаксической структуры предложения. Для этого используются алгоритмы обработки естественного языка, такие как TF-IDF (term frequency-inverse document frequency) или Word2Vec. Эти алгоритмы позволяют определить важность каждого слова в предложении и оценить его частоту использования.

После этого начинается этап семантического анализа текста. Здесь используются методы глубокого обучения, такие как рекуррентные нейронные сети или сверхточные нейронные сети. Они позволяют выделить ключевые понятия и связи между ними в тексте.

Наконец, последний этап — это классификация текста по заданному классу. Для этого используется алгоритм машинного обучения, который обучается на размеченных данных с указанием класса.

В современном мире, где образование играет ключевую роль в успехе каждого человека, вопрос о том, как сделать процесс обучения наиболее эффективным и комфортным, становится особенно актуальным. Одним из важных инструментов, способствующих повышению качества обучения, является автоматизированный анализ учебных текстов на предмет их сложности.

Сложность текста определяется его структурой, уровнем абстракции и сложностью используемых в нем понятий и терминов.

Анализ сложности учебных текстов позволяет выявить потенциальные проблемы, связанные с усвоением материала учащимися, и принять меры для их решения. Это может быть сделано путем упрощения структуры текста, замены сложных понятий на более простые, а также использование различных методик обучения, учитывающих индивидуальные особенности учащихся.

Использование автоматизированных методов анализа сложности текстов позволяет существенно сократить время, затрачиваемое на этот процесс, и повысить его точность. В основе таких методик лежит анализ структуры текста, его лексических и грамматических особенностей, а также применение алгоритмов машинного обучения и искусственного интеллекта [2; 3].

На сегодняшний день существует ряд компьютерных программ, разработанных для разных языков, которые способны определять сложность текста. Например, для английского языка представлена программа Coh-Metrix, которая помимо количественных показателей: длина слова и длина предложения, рассчитывает также синтаксическая простота, повествовательность, референциальная когезия, конкретность слов и также «глубокая» когезия [4; 6].

Особое внимание следует также уделить программе Rulingva, разработанной в «Казанском федеральном университете». Данная программа представляет собой автоматизированный анализатор для текстов на русском языке, который позволяет получить анализ по ряду показателей, напрямую связанных с уровнем сложности. К основным показателям можно отнести индекс читабельности, показатель лексического разнообразия и ряд других показателей. Основное внимание на данный момент в данном анализаторе уделено учебному материалу для начальной ступени обучения.

После проведения анализа сложности текста можно получить результаты, которые помогут лучше понимать особенности языка и его структуру. Например, если текст содержит много сложных синтаксических конструкций, то это может указывать на то, что он написан профессионалом в своей области. Если же текст имеет мало сложных элементов, то это может говорить о том, что автор не обладает достаточными знаниями в этой области.

Говоря о перспективах развития анализа сложности текста, стоит сказать, что одним из способов улучшения эффективности автоматизированного анализа сложности текста является использование более точных и разнообразных данных. Например,

можно использовать большие объемы текстовых данных, чтобы выявить связи между словами и фразами, а также учитывать контекст и семантику предложений. Также важно проводить обучение модели на большом количестве примеров различных стилей письма и тематик.

Кроме того, развитие технологий машинного обучения позволяет создавать более сложные модели анализа текста. Например, нейронные сети могут использоваться для классификации текстов по различным категориям или предсказания вероятности события на основе имеющихся данных.

В целом автоматизация анализа сложности текста имеет большой потенциал для улучшения качества работы специалистов в области обработки естественного языка. Однако необходимо продолжать работу над улучшением методов и алгоритмов, а также расширять доступ к качественным данным и технологиям машинного обучения.

Автоматизированный анализ сложности текста является важным инструментом для исследователей в области NLP. Он помогает улучшить качество результатов и сократить время, затрачиваемое на обработку большого объема текстовых данных.

Список использованных источников

1. Автоматическая обработка текстов; тематическая сегментация учебных текстов / М. И. Солнышкина, И. Э. Ярмакеев, Э. В. Гафиятова, Ф. Х. Исмаева // Вестник Самарского государственного технического университета. Серия: Психолого-педагогические науки. — 2019. — № 3(43). — С. 158-173.

2. Кисельников, А. С. К проблеме характеристик текста: читабельность, понятность, сложность, трудность / А. С. Кисельников // Филологические науки. Вопросы теории и практики. — 2015. — № 11-2(53). — С. 79-84.

3. Лингвистическая сложность учебных текстов / А. Я. Вахрушева, М. И. Солнышкина, Р. В. Куприянов [и др.] // Вопросы журналистики, педагогики, языкознания. — 2021. — Т. 40, № 1. — С. 89-99.

4. Солнышкина, М. И. Параметры сложности экзаменационных текстов / М. И. Солнышкина, А. С. Кисельников // Вестник Волгоградского государственного университета. Серия 2: Языкознание. — 2015. — № 1(25). — С. 99-107.

5. Солнышкина, М. И. Обработка естественного языка и изучение сложности дискурса / М. И. Солнышкина, Д. С. Макнамара, Р. Р. Замалетдинов // Russian Journal of Linguistics. — 2022. — Т. 26, № 2. — С. 317-341.

6. McNamara D. S., et al. Automated Evaluationnof Text and Discourse With Coh-Metrix. Cambridge, Cambridge University Press, 2014. 285 p.