

Рецензия на статью

Табаченкова Андрея

«Оценка сложности отдельных объектов в рамках
вложенных моделей данных текстовой разметки»

«ASSESSING THE DIFFICULTY OF SPANS
WITHIN NESTED TEXT MARKUP DATA MODELS»

Статья посвящена разработке методов оценки сложности текстовых фрагментов для задач цифровых гуманитарных наук. Предлагается подход, сочетающий обработку много-пользовательской разметки, векторизацию фрагментов и применение различных методов оценки сложности. В работе систематизированы подходы к оценке сложности объектов, включая метод Гауссиана, метод отступов и метод дисперсии градиентов.

Эксперименты демонстрируют эффективность методов на данных Kaggle NER Corpus и ADE, с выделением метода Гауссиана как наиболее точного по метрике ROC-AUC. Основными достоинствами работы являются системный подход к решению задачи, проработка адаптации методов к специфике текстовых данных и экспериментальная проверка.

В качестве недостатков отмечено тестирование на ограниченном числе наборов данных и влияние переобучения модели на результаты оценки сложности. Рекомендуется расширить валидацию методов на большем числе разнообразных наборов данных, а также углубить анализ интерпретируемости результатов для лучшего разделения объективной сложности фрагментов и артефактов обучения модели.

Таким образом, несмотря на отмеченные недостатки, представленная работа является законченным и ценным исследованием. Предложенный автором подход обладает несомненной новизной и практической значимостью для области цифровой гуманитаристики. Результаты исследования закладывают основу для создания инструментов, способных повысить эффективность и качество разметки текстовых данных. Статья может быть рекомендована к принятию после устранения указанных замечаний.

Рецензент,
Овсиенко Олеся,
Студентка ВМК МГУ