УДК 004.89

РАЗРАБОТКА ПОДСИСТЕМЫ ПРЕДОБРАБОТКИ ТЕКСТОВЫХ ДАННЫХ И ПОСТРОЕНИЯ ПРОСТРАНСТВА ПРИЗНАКОВ ДЛЯ ДИАЛОГОВОГО ПОМОЩНИКА

Лизунова И.А. (Университет ИТМО)
Научный руководитель – к.т.н. Махныткина О.В.
(Университет ИТМО)

В данной работе рассматриваются методы предварительной обработки текстовых данных и построения пространства признаков, применяемые в рамках разработки системы виртуального диалогового помощника для поддержки проведения дистанционного экзамена.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 619423 «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе аргументационного подхода и глубокого машинного обучения».

В настоящее время в современных информационных системах активное развитие получила сфера разработки диалоговых и вопросно-ответных систем на основе методов глубокого машинного обучения, однако наблюдается недостаточная апробация новейших научных идей для подобных систем, работающих на русском языке. В связи с этим большой интерес представляет разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе аргументационного подхода и глубокого машинного обучения. Важной составляющей данной системы является подсистема предобработки текстовых данных и построения пространства признаков.

При разработке системы диалогового помощника решается множество различных задач, таких как осуществление выделения основных тем и фактов из текстовых данных, анализ тональности и эмоциональности пользовательских комментариев, генерация вопросов на основе текстов, выявление релевантности и корректности ответов пользователей, для каждой из которых используются различные методы предобработки текстовых данных. Предварительная обработка текстов может включать следующие этапы: графематический анализ, морфологический анализ и синтаксический анализ. Графематический анализ представляет собой работу с элементами графематической структуры и включает себя разбиение текстов на абзацы и предложения, разбиение предложений на слова, выявление сокращений, определение регистра слов, выделение знаков пунктуации, цифровых комплексов и иных видов графем. Морфологический анализ основан на определении морфологических характеристик слова и включает в себя стемматизацию, которая представляет собой процесс нахождения основы слова, лемматизацию, в ходе которой определятся начальная форма слова, определение грамматических характеристик слова и получение всех грамматических форм данного слова. На этапе синтаксического анализа осуществляется выявление синтаксических связей слов и грамматической структуры предложений путем построения деревьев грамматик зависимостей, в которой все связи в предложении рассматриваются как подчинительные, а вершиной предложения признается сказуемое или его знаменательная часть. Для решения каждой из рассмотренных задач был определен список необходимы алгоритмов предварительной обработки. Следующим этапом работы является построение векторного пространства признаков, для которого применяется fine-tuning предобученной языковой модели BERT.

В данной работе был выполнен обзор существующих методов предварительной обработки текста и проведены графематический анализ, морфологический анализ, синтаксический анализ текстов на естественном языке, а также выполнено построение пространства признаков, что является необходимой основой для осуществления дальнейших этапов разработки виртуального диалогового помощника.