

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ
Кафедра дифференциальных уравнений и системного анализа

Аннотация к дипломной работе

**МАШИННОЕ ОБУЧЕНИЕ В ЗАДАЧАХ ОБРАБОТКИ ЕСТЕСТВЕННОГО
ЯЗЫКА И ДИСКУРС-АНАЛИЗА**

Филиппович Павел Александрович

Научный руководитель:
кандидат физ.-мат. наук,
доцент кафедры веб-
технологий и
компьютерного
моделирования
доцент А.Х. Перез Чернов

Дипломный проект представлен в виде пояснительной записки объемом 56 страниц, содержит 16 таблиц, 16 рисунков, 40 источников.

Ключевые слова: МАШИННОЕ ОБУЧЕНИЕ, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, ГЕНЕРАЦИЯ ТЕКСТА, СЕНТИМЕНТ АНАЛИЗ, СЕМАНТИЧЕСКИЕ ПРОПОЗИЦИИ, РАЗМЕТКА ДАННЫХ

Цель работы: изучение алгоритмов генерации текста, извлечения семантических пропозиций и анализа тональности текста относительно аспектов. Разработка архитектуры системы для полуавтоматической генерации и разметки данных.

Дипломный проект является обобщением современных подходов к обработке естественного языка с помощью инструментов машинного обучения. В работе рассматривается применение на практике основных алгоритмов, а также предложена архитектура системы для полуавтоматической генерации и разметки данных. Реализация алгоритмов и моделей осуществляется с помощью языка программирования «Python».

Во введении содержится информация о задаче дискурс-анализа и о сфере машинного обучения в контексте обработки естественного языка, приводится постановка задачи.

В главе 1 «Обзор алгоритмов генерации текста» рассмотрены современные алгоритмы генерации текста на естественном языке. Приведены необходимые определения, описаны схожие и отличающиеся составляющие алгоритмов.

В главе 2 «Извлечение семантических пропозиций» описана модель Deep SRL для извлечения семантических пропозиций. Произведена модификация существующей реализации модели и применение модели на практике.

В главе 3 «Определение тональности текста относительно заданных аспектов» описана модель MemNet для решения задачи определения тональности

текста относительно заданных аспектов. В рамках главы была произведена реализация модели, а также проверка результатов, полученных авторами.

В главе 4 «Система для полуавтоматической генерации и разметки данных» разработана архитектура системы для полуавтоматической генерации и разметки данных на основе MaskGAN модели, описанной в главе 1.

Graduating paper is presented in the form of an explanatory note in the volume of 56 pages, contains 16 tables, 16 figures, 40 sources.

MACHINE LEARNING, NATURAL LANGUAGE PROCESSING, TEXT GENERATION, SENTIMENT ANALYSIS, SEMANTIC ROLE LABELING, DATA LABELING

Graduating paper is presented in the form of an explanatory note in the volume of 56 pages, contains 16 tables, 16 figures, 40 sources.

The aim of graduation project: studying algorithms for text generating, semantic role labeling and aspect-based sentiment analysis. Development of a system architecture for semi-automatic generation and data labeling.

Graduation project is a generalization of knowledge associated with machine learning in the tasks of processing natural language. The work includes the application of the main described algorithms, as well as the development of a system architecture for semi-automatic generation and data labeling. Implementation of algorithms and models is carried out using the programming language «Python».

The introduction contains information about the task of discourse analysis and the field of machine learning in the context of natural language processing, and the problem is formulated.

In chapter 1 «Обзор алгоритмов генерации текста», modern algorithms of text generation in natural language are considered. The necessary definitions are given, similar and different components of the algorithms are described.

In chapter 2, «Извлечение семантических пропозиций» the Deep SRL model is described to perform semantic role labeling. The existing implementation of the model was modified and it was tested on practise.

In chapter 3 «Определение тональности текста относительно заданных аспектов», the MemNet model is described for solving the problem aspect-based

sentiment analysis. Within chapter, the model was implemented and the results obtained by the authors were validated.

In chapter 4, «Система для полуавтоматической генерации и разметки данных», a system architecture for semi-automatic generation and data labeling is developed based on the MaskGAN model described in chapter 1.