

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Аннотация к магистерской диссертации

**«Задача нормализации содержащих общепринятые аббревиатуры и
выражения текстов»**

Климук Татьяна Алексеевна

Научный руководитель – кандидат физико-математических наук,
доцент Соболевская Е. П.

Минск, 2019

Реферат

Магистерская диссертация, 47 страниц, 3 рисунка, 8 таблиц, 13 источников.

ЛЕКСЕМА, ВЫРАЖЕНИЕ, НОРМАЛЬНАЯ ФОРМА, НОРМАЛИЗАЦИЯ ТЕКСТА, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, ЗАДАЧА КЛАССИФИКАЦИИ, НЕЙРОННАЯ СЕТЬ.

Объект исследования – методы нормализации текстов на естественном языке, содержащих общепринятые выражения и аббревиатуры.

Цель работы – изучить методы нормализации слов и общепринятых выражений языка, разработать и реализовать алгоритмы нормализации текстов, провести сравнительный анализ полученных результатов, определить возможные направления работы для улучшения качества работы алгоритмов.

Методы исследования – анализ, эксперимент, тестирование, сравнение.

В ходе работы было выделено две подзадачи: классификация выражений естественного языка и построение нормализованных последовательностей. Для каждой из задач был проведён обзор наиболее распространённых методов решения, в частности методы, использующие машинное обучение. Описанные подходы были применены к решению указанных задач для русского и английского языков, был проведён сравнительный анализ алгоритмов и анализ ошибок, были предложены возможные направления дальнейших исследований.

Результатом работы являются множество моделей, построенных в ходе решения поставленных подзадач, и подходы к решению задач обработки текстов, предложенные в работе. Среди моделей отдельно необходимо выделить следующие: нейронные модели классификации выражений, способные с точностью более 99.5% определять тип выражений английского и русского языков, нейронная генеративная модель с механизмом внимания, нормализующие слова и выражения русского языка с точностью 98%, а также гибридная модель, использующая модель классификации, правилый и нейронный подходы, для нормализации текстов на английском языке с точностью 99%.

Область применения – решение задач нормализации текстов естественного языка, распознавания и синтеза речи, машинного перевода текстов, информационного поиска.

Abstract

Master thesis, 47 pages, 3 pictures, 8 tables, 13 references.

LEXEME, EXPRESSION, NORMAL FORM, TEXT NORMALIZATION, NATURAL LANGUAGE PROCESSING, CLASSIFICATION PROBLEM, NEURAL NET.

The object of research – methods of normalization of natural language texts containing common expressions and abbreviations.

The aim of this work is to study the methods of words and common language expressions normalization, to develop and implement algorithms for texts normalization, to conduct a comparative analysis of the results, to identify possible areas of work for improving the quality of the algorithms.

Research methods – analysis, experiment, testing, comparing.

During the course of the research two subtasks were identified: the classification of natural language expressions and the construction of normalized sequences. For each of the tasks a review of the most common methods of solution was conducted, in particular, machine learning methods. The described approaches were applied to the solution of these problems for the Russian and English languages, comparative analysis of algorithms and error analysis were carried out, possible directions of further research were proposed.

The result of the work is a set of models built in the course of solving the identified subtasks and approaches to solving the problems of text processing proposed in the work. It is necessary to single out the following models: neural models of expression classification, capable to determine with an accuracy of more than 99.5% the type of the English and Russian language tokens, a neural generative model with the mechanism of attention, normalizing words and expressions of the Russian language with an accuracy of 98%, as well as the hybrid model using the classification model, rules and neural approaches for normalizing texts in English with an accuracy of 99%.

Field of application – tasks of natural language texts normalization, speech recognition and synthesis, machine translation of texts, information retrieval.