Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК

Калевич
Вячеслав Владимирович

Методы построения схемы машинного обучения для распознавания тематики запросов на русском языке в системах автоматического общения

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

подпись магистранта

Научный руководитель
Смолякова О.Г.
к.т.н., доцент

подпись научного руководителя

Минск 2019

КРАТКОЕ ВВЕДЕНИЕ

Обработка естественного языка (NLP) в последнее время получила большое распространение и популярность в качестве инструмента для анализа человеческого языка в цифровой области. NLP можно встретить в различных областях электронной сферы человека, таких как машинный перевод, обнаружение спама в электронной почте, извлечение информации, обобщение, медицинское обслуживание, ответы на вопросы.

Обработка естественного языка (NLP) – это комплекс искусственного интеллекта и лингвистики, предназначенный для того, чтобы компьютеры понимали утверждения или слова, написанные на человеческих языках. Обработка естественного языка появилась для облегчения работы пользователя и улучшения работоспособности приложений.

Колоссальные объемы текстовой информации, генерируемой в системах электронной коммуникаций, социальных сетях и Всемирной паутине в сочетании с необходимостью быстрого доступа к конкретной информации привели к продвижению и коммерческому внедрению NLP в последние годы. В настоящее время NLP широко интегрируется с веб-приложениями и мобильными приложениями, обеспечивая естественное взаимодействие между человеком и компьютерами.

На основании вышеизложенного можно выделить актуальную проблему расширения сферы применения NLP в системах автоматического общения, разработки новых алгоритмов машинного обучения для распознавания запросов на русском языке, поскольку для русскоязычного региона разработано достаточно небольшое количество систем с интеллектуальным распознаванием тематики запроса.

Диссертационная работа посвящена разработке алгоритмов, методов, анализу и выбору схемы построения этапов процессинга запроса на русском языке для его дальнейшего преобразования и оценки. Создание такой схемы обработки и анализа текста с помощью машинного обучения позволит системе автоматического общения (чат-боту) выдавать пользователю релевантный контент, напрямую относящийся к запросу и его тематике.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является разработка методов построения схемы машинного обучения для распознавания тематики запросов на русском языке на базе библиотеки Apache OpenNLP и мессенджера Telegram.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Проанализировать процесс обработки естественных языков при помощи машинного обучения.
2. Выделить критерии оценивания эффективности алгоритма машинного обучения.
3. Выбрать фреймворк и платформу, на которой будет происходить обучение.
4. Разработать наборы данных для распознавания границ текста и разбиения предложений на лексемы.
5. Разработать алгоритм по приведению слов к начальной форме и выделению ключевых слов в пользовательском запросе.
6. Разработать архитектуру программной системы автоматизированного определения тематики запросов (серверная и клиентская часть).
7. Реализовать чат-бота на базе мессенджера Telegram для выдачи релевантного контента юзеру по запросу.
8. Провести экспериментальные исследования разработанной системы.

Объектом исследования являются существующие системы автоматического общения и используемые в них алгоритмы распознавания запросов.

Предметом исследования является программное обеспечение и алгоритмы, обеспечивающие обучение нейронной сети по обработке предложений на русском языке, алгоритм последовательной обработки текста (от предложения до лексем с обозначением частеречной принадлежности)

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность создания эффективной схемы обработки нативного текста на русском языке для получения релевантного, основанного на распознанной тематике запроса пользователя, контента.

Область исследования

Содержание диссертации соответствует образовательному стандарту высшего образования второй ступени (магистратуры).

Теоретическая и методологическая основа исследования

Основой исследования послужили работы зарубежных ученых в области нейронных сетей, обработки естественного языка, филологии, рассмотрении существующих решений и технологий NLP в различных сферах деятельности, и создании аналогов с улучшенными техническими составляющими.

Информационная база исследования сформирована на основе литературы в области машинного обучения, обработке естественных языков, исследованиях в области распределенного представления предложений, слов, языкового моделирования и его нейро-вероятностного варианта.

Научная новизна

Научная новизна и значимость полученных результатов работы заключается в создании новых наборов данных для преобразования текстов, которые вкупе с проработанной схемой машинного обучения способны категоризировать текст.

Теоретическая значимость работы заключается в разработке методов, моделей и инструментальных средств категоризации текстов на русском языке.

Практическая значимость диссертации состоит в пригодности применения результатов исследований в различных областях, связанных с системами автоматического общения, категоризации текстовой информации, поиска релевантного контента. Полученные наборы данных обеспечат эффективную работу алгоритмов поиска.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя О. Г. Смолякова, заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались и обсуждались на Республиканской научно – технической конференции «Информационные технологии и системы: проблемы, методы, решения» (Минск, Беларусь, 2018) в секции «Информационно – вычислительные процессы в распределенных и параллельных системах».

Опубликованность результатов диссертации

По теме диссертации опубликовано 2 печатные работы, из них 1 статья в международном научном журнале, 1 работа в сборниках трудов и материалов научно – технических конференций.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области, выявлены основные существующие параметры, критерии оценивания схем обработки естественного языка. Вторая глава посвящена разработке архитектуры ПО, самого ПС и алгоритмов обработки текстов, методик создания датасетов (наборов данных). В третьей главе представлены результаты расчета критериев эффективности системы распознавания.

Общий объем работы составляет 59 страниц, из которых основного текста – 50 страниц, 16 рисунков на 10 страницах, 3 таблиц, список использованных источников из 31 наименования на 2 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** описаны задачи, решаемые по средствам обработки естественного языка, указаны критерии оценивания NLP – конвейеров и алгоритмов машинного обучения, проведен обзор аналогов.

Обработка естественного языка (NLP) в последнее время получила большое распространение и популярность в качестве инструмента для анализа человеческого языка в цифровой области. NLP можно встретить в различных областях электронной сферы человека, таких как машинный перевод, обнаружение спама в электронной почте, извлечение информации, обобщение, медицинское обслуживание, ответы на вопросы.

Большая часть работы по обработке естественного языка проводится учеными в сфере информационных технологий, в то время как различные специалисты в других сферах также проявили интерес. К ним относятся лингвисты, психологи, философы. Один из самых иронических аспектов NLP

заключается в том, что он дополняет знание человеческого языка. Область NLP связана с различными теориями и методами, которые имеют дело с проблемой естественного языка общения с компьютерами.

Чаще всего языком программирования для проектов машинного обучения является язык Python и Java. Библиотека Apache OpenNLP - это основанный на машинном обучении инструментарий для обработки текста на естественном языке. Он поддерживает наиболее распространенные задачи NLP, такие как токенизация, сегментация предложений, тегирование части речи, извлечение именованных объектов, разбиение на фрагменты, анализ и разрешение по ключевым словам. Эти задачи обычно требуются для создания более сложных служб обработки текста.

В общем, задачи машинного обучения (и NLP обычно попадают в эту область) измеряются на основе того, насколько хорошо машина может дать правильный ответ на вопрос (часто называемый «прогнозированием»). Точность классификации – это то, что мы обычно имеем в виду, когда используем термин «точность». Это отношение количества правильных прогнозов к общему количеству входных выборок.

Вторая глава посвящена общему методу построения NLP-конвейера, этапам обработки нативного текста, методам построения наборов данных и методиках сбора данных, подготовки данных для дальнейшего преобразования и анализа, а также описание алгоритма работы NLP-конвейера.

Процесс чтения и понимания русского языка очень непрост. Наличие причастных и деепричастных оборотов, многозначность слов и фразеологические обороты делают задачу распознавания достаточно сложной. Чтобы выполнить задачи такой сложности с использованием машинного обучения (МО), обычно строится конвейер. Задача разбивается на несколько очень небольших частей, после чего модели МО решают каждую подпроблему по отдельности. Далее модели соединяются в конвейер, по которому обмениваются информацией, что даёт возможность решать задачи очень высокой сложности. Именно так происходит обработка естественных языков.

Перед началом работы необходимо создать наборы данных для обучения нейронной сети и получения коэффициентов для дальнейшего распознавания тех или иных частей введенного текста. Подготовка набора данных – очень трудоемкая задача, которая таит в себе достаточно большое число проблем. Проблемы с наборами данных машинного обучения могут быть связаны с тем, как строится организация, с установленными рабочими процессами и с тем, соблюдаются ли инструкции среди тех, кто отвечает за ведение записей.

С самого начала важно ориентироваться на большие данные, но большие данные не о петабайтах. Все дело в умении правильно их обрабатывать. Чем

больше набор данных, тем сложнее его правильно использовать и получать информацию.

В связи с тем, что обучение происходит на базе фреймворка Apache OpenNLP, существует определенный формат построения датасета для дальнейшего обучения и получения коэффициентов. OpenNLP имеет инструмент командной строки, который используется для обучения моделей. Данные могут быть преобразованы в формат обучения OpenNLP.

Третья глава посвящена анализу показателей эффективности работы алгоритма в зависимости от объема данных для тренировки нейронной сети.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проведен анализ процесса обработки естественного языка и задач, решаемых по средствам NLP-конвейеров. В рамках анализа были выявлены науки, которые тесно взаимодействуют с областью компьютерной обработки языков, определены базовые понятия данной сферы и проблемы. Одной из основных проблем данной области является неоднозначность фраз и слов, которые употребляются в речи.

2. Осуществлен обзор инструментов системы NLP с последующим выбором определенных для дальнейшего создания алгоритма обработки русскоязычных запросов. Чаще всего при построении NLP-конвейеров используются частеречный анализатор (POS), определитель имен собственных (NER), определитель начальных форм слов (lemmatizer), токинезаторы, разбивающие предложения на слова. Также в рамках обзоры были определены наиболее приоритетные сферы использования естественной обработки языка. Такими сферами являются машинный перевод, текстовая категоризация, спам – фильтрация, извлечение информации, обобщение информации, диалоговая система и медицина.

3. Проведен обзор аналогов фреймворков, специализирующихся на обработке естественного языка с оценкой эффективности, функционала, возможностей для разработчиков и удобства обучения кастомных наборов данных. Библиотека Apache OpenNLP - это основанный на машинном обучении инструментарий для обработки текста на естественном языке. Он поддерживает наиболее распространенные задачи NLP, такие как токенизация, сегментация предложений, тегирование части речи, извлечение именованных объектов, разбиение на фрагменты, анализ и разрешение по ключевым словам.

4. Были проанализированы и определены критерии оценки эффективности работы алгоритмов ML и NLP конвейеров, проанализирован общий метод построения NLP конвейера с детальным описанием каждого из последующих шагов.

5. Выделены методы построения наборов данных. В процессе определения методов был проведен анализ каждого из них по критериям эффективности, простоты обработки и подготовки данных, затраты временных ресурсов, способность работать с большими данными.

6. Разработан алгоритм и программное средство (чат-бот с распознаванием тематики запросов на русском языке) на базе мессенджера Telegram. В рамках разработки были созданы наборы данных для таких частей NLP конвейера как лемматизация, частеречное распознавание, определение имен собственных.

7. На основании разработанной системы были проведены замеры и дана оценка эффективности разработанного алгоритма в задачах распознавания тематики запроса на русском языке. К критериям, по которым производилась оценка относятся точность определения, F1 Score. Данные показатели замерялись при различном уровне обученности нейронной сети и также были проанализированы.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Калевич В.В. Использование микросервисной архитектуры в задачах машинного обучения / Калевич В.В. // Республиканская научно-техническая конференция «Информационные технологии и системы: проблемы, методы, решения» – Минск, 2018. – с. 119 –123.

2. Калевич В.В. Микросервисная архитектура при решении задач машинного обучения / Калевич В.В. // Журнал «Молодой ученый» №23 (261) – июнь 2019 – с. 17 –19.