МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

«Новосибирский национальный исследовательский государственный университет»

(Новосибирский государственный университет, НГУ)

Структурное подразделение Новосибирского государственного университета –

Высший колледж информатики Университета (ВКИ НГУ)

КАФЕДРА ИНФОРМАТИКИ

**РАЗРАБОТКА ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ ДЛЯ ВОССТАНОВЛЕНИЯ ПУНКТУАЦИИ ДЛЯ КОМПАНИИ «ЭМБЛ» НА ПЛАТФОРМЕ PYTORCH**

|  |  |
| --- | --- |
| Руководитель | Кудинов В.Ю.  «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_2022 г. |
| Студент 3 курса  гр. 907 сб | Озерная М.А.  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_2022 г. |

Новосибирск

2022

**СОДЕРЖАНИЕ**

[ПЕРЕЧЕНЬ СОКРАЩЕНИЙ, УСЛОВНЫХ ОБОЗНАЧЕНИЙ И ТЕРМИНОВ 3](#_Toc93987743)

[ВВЕДЕНИЕ 4](#_Toc93987744)

[1 ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ 6](#_Toc93987745)

[2 ПОСТАНОВКА ЗАДАЧИ 8](#_Toc93987746)

[3 АНАЛОГИ 9](#_Toc93987747)

[4 ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ 10](#_Toc93987748)

[СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ 11](#_Toc93987749)

# ПЕРЕЧЕНЬ СОКРАЩЕНИЙ, УСЛОВНЫХ ОБОЗНАЧЕНИЙ И ТЕРМИНОВ

ИНС – искусственная нейронная сеть

(затычка)Интернет-магазин (англ. online shop или e-shop) — сайт, торгующий товарами посредством сети Интернет. Позволяет пользователям онлайн, в своём браузере или через мобильное приложение, сформировать заказ на покупку, выбрать способ оплаты и доставки заказа, оплатить заказ. При этом продажа товаров осуществляется дистанционным способом при этом накладывает ограничения на продаваемые товары.

Интернет-ресурс (синонимы «веб-ресурс, [веб-сайт](https://www.glossary-internet.ru/terms/%D0%92/veb_sayt/), веб-сервис, сайт») – это совокупность интегрированных средств технического и программно-аппаратного характера, а также информации, предназначенной для публикации во Всемирной паутине. Интернет-ресурс может содержать информацию в текстовой, графической и мультимедийной форме. Каждый интернет-ресурс должен иметь уникальный адрес, который позволяет найти его в Сети.

# ВВЕДЕНИЕ

Пунктуация – часть символьного представления речи, увековеченной в текстовой записи, принадлежащей определённой языковой культуре. На протяжении существования человеческой цивилизации текст является наиболее ёмким способом сохранять мысли для их исследования и усовершенствования.

Сложные плоды деятельности головного мозга могут быть закодированы в тексте при помощи слов и пунктуации. Слова могут передать, например, смысл окружающих объектов, их свойств и действий. Пунктуация нужна для правильной группировки слов, знаки препинания не несут информации об окружающем мире, их роль – вспомогательная, но немаловажная. Благодаря им текст приобретает структуру и передаёт больший смысл, чем все употреблённые в нём слова и их сочетания по отдельности.

Темой данной выпускной квалификационной работы является «Разработка искусственной нейронной сети для восстановления пунктуации для компании «ЭМБЛ» на платформе PyTorch». Данная разработка нацелена на внедрение в качестве усовершенствования программного продукта, выпускаемого в ЭМБЛ. Компания ООО "ЭМБЛ" — резидент Технопарка Академгородка. Занимается исследованиями в области нейросетевых технологий и разработкой ПО. Продуктом компании является система анализа речи и эмоций Darwin. Система контролирует качество клиентского сервиса и помогает менеджерам более эффективно проводить переговоры.

Цель данной дипломной работы заключается в построении и тренировке специальной ИНС для решения задачи восстановления пунктуации в тексте на русском языке, в котором отсутствуют знаки препинания. Такая ИНС заполнила бы существенный пробел в современном активном продвижении взаимодействия людей и машин. В частности, голосовых роботов-помощников.

В настоящее время развились и начинают всё шире использоваться технологии машинного обучения, позволяющие получить текст из устной речи. Это специально смоделированные и обученные ИНС, распознающие слова в звуковой записи и переводящие их в текстовую. По причине невозможности распознавания пунктуации только по интонационной составляющей устной речи, которая доступна для анализа таким ИНС, они не могут восстановить пунктуационные знаки. Перестраивать и тренировать такие сети для решения дополнительных задач было бы неоправданно сложно, дорого и невыгодно в плане производительности. Тем не менее, решение данной задачи методами явного программирования чрезвычайно осложнено и практически невозможно, учитывая зависимость пунктуации от семантики, и задача входит в круг решаемых с помощью нейросетевых технологий, а именно: подходит для решения методами глубокого обучения нейронных сетей – имеется большое, то есть достаточное, количество данных для обучения. Именно метод обучения алгоритма глубокой нейронной сети на больших данных и был выбран для достижения поставленной цели.

Значение разработки и внедрения такой технологии сравнимо с обучением голосовых роботов грамотности.

# 1 ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

Точки использовали древние греки и римляне для разделения предложений. Также они разделяли текст на абзацы. “Пунктуация древних имела, главным образом, в виду ораторские требования — произнесение речи и её декламацию” — С. К. Булич.

Современная система пунктуации была заложена одним из первых издательских домов в XVI веке в Венеции. В 1566 году типограф Альд Мануций Младший издал первую книгу о принципах пунктуации. Вместе с его дедом, основателем дома Альда, они считаются учредителями стандартизованной системы пунктуации. С того времени в европейской пунктуации не было сделано существенных изменений.

В современной русской пунктуации задействованы такие знаки препинания как точка, запятая, точка с запятой, двоеточие, многоточие, вопросительный и восклицательный знаки, кавычки, скобки и тире. Их расстановка обусловлена контекстом.

В оформлении текста пунктуация позволяет кратко и наиболее удобно связать сложные смысловые отношения и придать оттенки письменной речи. При помощи пунктуации пишущий достигает адекватного выражения своих мыслей, а читающий имеет возможность их правильно понять.

Набор знаков в тексте не случаен и не хаотичен: “Твердость и стабильность русской пунктуации определяется ее структурным принципом, а возможности передавать богатство и разнообразие смысловых оттенков и эмоций — принципами смысловым и интонационным” — писал К. Паустовский.

Три вышеперечисленных принципа порождают большое количество правил, которые делают расстановку пунктуации в тексте, где она по каким-то причинам отсутствует, сложной задачей, которой занимаются специалисты по обработке естественного языка — науки на пересечении машинного обучения с математической лингвистикой. Цель обработки естественного языка (Natural Language Processing, NLP) – обучение машин работе с обычным человеческим языком. В 60-х годах решением проблем в этой области занимались парсеры со словарями, в 70-х – графовые структуры, а после 80-х начали активно применяться алгоритмы машинного обучения (Machine Learning), а в 90-х годах – n-граммы. В данный момент алгоритмы NLP применяются для поиска, перевода, показа подходящей онлайн-рекламы, при распознавании речи и настроения в чат-ботах и голосовых помощниках.

Машинное обучение – это подраздел искусственного интеллекта, который изучает методы построения алгоритмов, не напрямую решающих задачу, а обучающихся за счёт применения решений множества сходных задач. На входе алгоритм получает данные, из которых на выход передаётся только самое главное, отвечающее за результат. В случае с обработкой естественного языка, а конкретно, с восстановлением пунктуации, данные для обучения представляют собой файлы с большим количеством текста, в котором пунктуация присутствует. А после обучения алгоритм сможет предсказать вероятность пунктуационного знака после конкретного слова в тексте. Благодаря возможности работы с большими объёмами информации на многоядерных CPU и GPU сейчас в машинном обучении развивается технология глубокого обучения (Deep Learning), которая даёт хорошие результаты за счёт новых моделей и алгоритмов, таких как механизм внимания (attention) и модель Трансформер (Transformer), которая его использует.

# 2 ПОСТАНОВКА ЗАДАЧИ

Необходимо разработать нейронную сеть, позволяющую получать из сплошного текста текст с правильной пунктуацией быстро, без применения внешних сервисов. Для этого заказчику наиболее хорошо подходит нейросеть, написанная на библиотеках PyTorch при помощи языка программирования Python.

В ходе решения задачи запланировано осуществить следующие шаги:

1. Проанализировать предметную область;
2. Исследовать имеющиеся технологии NLP;
3. Выбрать подходящую модель нейронной сети;
4. Построить алгоритм для обработки данных, тренировки и оценки нейросетевой модели;
5. Подготовить данные для обучения модели;
6. Произвести тренировку модели;
7. Оценить результаты работы алгоритма и при необходимости вернуться на предыдущие шаги.

# 3 АНАЛОГИ

Есть два наиболее известных сервиса обработки текста – это Главред и Grammarly.

Ссылка на первый сервис:

<https://glvrd.ru/>

Ссылка на второй сервис:

<https://www.grammarly.com/>

Оба сервиса предоставляют обширные возможности для работы с текстом. Главред помогает находить в русском тексте языковой мусор, рекламные и журналистские штампы, признаки плохого синтаксиса и т. п., Grammarly - писать на английском эффективнее и выдает рекомендации касательно правильности, четкости, увлекательности и тона сообщения.

Отличие сервисов заключается в алгоритмах и функциональности. Grammarly использует нейросетевые алгоритмы, и в нём, в отличие от русскоязычного сервиса, производится восстановление пунктуации.

Из этого можно сделать выводы, что нейросетевые технологии наиболее хророшо подходят для задачи восстановления пунктуации в тексте, и программа для выполнения этой задачи для русского языка актуальна. В особенности, для компании-заказчика, не желающего использовать для этой задачи сторонние сервисы.

# 4 ФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ

Точки использовали древние греки и римляне для разделения предложений. Также они разделяли текст на абзацы. “Пунктуация древних имела, главным образом, в виду ораторские требования — произнесение речи и её декламацию” — С. К. Булич.

# 5 НЕФУНКЦИОНАЛЬНЫЕ ТРЕБОВАНИЯ

Точки использовали древние греки и римляне для разделения предложений. Также они разделяли текст на абзацы. “Пунктуация древних имела, главным образом, в виду ораторские требования — произнесение речи и её декламацию” — С. К. Булич.

# 6 ХАРАКТЕРИСТИКА ВЫБРАННЫХ ПРОГРАММНЫХ СРЕД И СРЕДСТВ

PyTorch – это современная среда (фреймворк) машинного обучения с открытым исходным кодом, которую разрабатывает преимущественно группа искусственного интеллекта Facebook для решения ряда задач методами искусственного интеллекта, в том числе обработки естественного языка. Среда представлена набором библиотек, содержащих различные математические функции, основные из которых в машинном обучении – это определение графа вычислений, его дифференциации и вычисление. Из-за строгой типизации тензоров PyTorch является наиболее безопасным, чем например NumPy, выдавая ошибки при несоответствии типа тензора, а не преобразовывая его без уведомлений в другой. PyTorch при помощи специальной архитектуры CUDA позволяет управлять переносом вычислений с CPU на GPU (то есть с обычного процессора на графический), что очень хорошо сказывается на производительности. В целом, фреймворк PyTorch некоторым кажется даже слишком низкоуровневым, но свобода для кастомизации не мешает ему иметь такие удобные для асинхронной работы с данными (и не только) высокоуровневые модули как torch.nn и torch.optim, torch.utils.data.

Язык, который позволяет работать с фреймворком – это Python,

# СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. С. К. Булич, Энциклопедический словарь Брокгауза и Ефрона (1907 год)
2. [Альд Мануций — Википедия (wikipedia.org)](https://ru.wikipedia.org/wiki/%D0%90%D0%BB%D1%8C%D0%B4_%D0%9C%D0%B0%D0%BD%D1%83%D1%86%D0%B8%D0%B9)