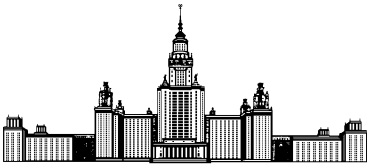
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 417 ГРУППЫ

«Построение поисковой системы для интернет магазина»

Выполнил:

студент 4 курса 417 группы

Федоров Илья Сергеевич

Научный руководитель:

д.ф-м.н., профессор

Дьяконов Александр Геннадьевич

Заведующий кафедрой

Математических Методов

Прогнозирования, академик РАН _____ Ю. И. Журавлёв

К защите допускаю

«_____» _____ 2021 г.

К защите рекомендую

«_____» _____ 2021 г.

Москва, 2021

Содержание

Аннотация

Todo

1 Введение

Современный технологический прогресс неразрывно связан с быстрым доступом к информации. Ежедневно поисковые системы обрабатывают миллиарды запросов от колоссального количества людей по всему миру, а их базы данных исчисляются десятками миллионов терабайт. И хотя поисковые системы, безусловно, стоит считать одним из главных достижений человечества за последние десятилетия, задача получения данных по запросу возникает и в другиях приложениях меньших масштабах, к примеру, задача выдача подходящих под запрос пользователя товаров в интернет-магазине.

Основной целью данной работы является исследование подходов к построению поисковой системы для абстрактного интернет-магазина с использованием современных технологий глубокого обучения и нейронных сетей. Безусловно для её решения, существуют (и активно применяются на практике) методы, никак не использующие машинное обучение, однако рост интереса исследователей и бизнеса к глубокому обучению, а также активное развитие вычислительной техники, позволяющие обучать глубокие нейросети, способствуют развитию применения технологий интеллектуальной обработки естественного языка в данной области.

В исследовании рассмотрены некоторые из существующих подходов к приложению технологий глубокого обучения к задаче построения поисковой системы для интернет-магазина, описаны особенности сбора данных в контесте решаемой задачи, предложена архитектура поисковой системы, основанная на BERT, а также представлены результаты экспериментов и значения метрик для предложенной модели.

2 Постановка задачи

Существует множество подходов к построению поисковой системы для интернет-магазина. Во многом они зависят от способа представления товаров в базе данных. К примеру, они могут описываться большим количеством признаков (такими как название, бренд, страна-производитель, размеры, цвет и т.д.), а могут лишь одной текстовой строкой - описанием, в котором, как предполагается, содержится вся информация о товаре.

Основным случаем, рассмотренным в данной работе, является модель данных, в которой товары в базе магазина описываются тремя признаками: описание товара, название товара и название бренда. Примеры представлены в таблице 1.

№	Описание товара	Название товара	Бренд
1	Смартфон Apple iPhone 5S 16Gb Space Gray (ME432RU/A)	Смартфон	Apple
2	Варенье Вкусвилл брусника - апельсин, 310 г	Варенье	Вкусвилл
3	Женские кроссовки CALVIN KLEIN JEANS	Кроссовки	CALVIN KLEIN JEANS
4	Гриль Bosch TFB3323V	Гриль	Bosch
5	TELEFUNKEN / Портативная колонка TF-1234B	Портативная колонка	TELEFUNKEN

Таблица 1: Пример данных в базе интернет-магазина

Данная модель является некоторым промежуточным вариантом между тем случаем, когда товар описывается большим количеством различных признаков, и тем, когда товар описывается лишь одной текстовой строкой (то, что в рассматриваемой модели называется описанием товара). Мотивация выбора такой модели будет описана в следующем подразделе «Особенности сбора данных».

Будем считать, что интернет-магазин имеет доступ к базе данных с описаной выше моделью, а также некоторый сайт, который позволяет получать текстовые запросы от клиентов (пользователей). Пользователь задает свой запрос в текстовом виде на естественном языке. В ответ поисковая система должна предоставить некоторый список товаров из базы, которые являются наиболее релевантными запросу пользователя. Мы предполагаем, что пользователь вводит запрос, содержащий в себе слова, характеризующие товар. Безусловно, идеальным запросов являлось бы точно описание товара из базы данных магазина, тогда мы бы смогли произвести явный поиск по существующим наименованиям и выдать точный результат. Однако на практике пользователь никогда не будет знать, как именно нужный ему товар представлен в базе, а также довольно часто клиента интересует не конкретный товар, а некоторый каталог релевантных его запросу товаров. К примеру, его может интересовать чайник (без уточнения бренда) или все продукты компании Apple (без уточнения названия товара). Таким образом, будем предполагать, что запрос пользователя будет состоять из трех компонент, каждая из которых может как присутствовать, так и отсутствовать: название товара, название бренда и уточняющая информация. При-

меры запросов с соответствующим разделением слов по указанным трем компонента представлены в таблице 2.

№	Запрос	Название товара	Бренд	Уточняющая информация
1	чехол для iphone 6	чехол	(нет)	для iphone 6
2	смартфоны xiaomi	смартфоны	xiaomi	(нет)
3	чайник электрический маленький	чайник	(нет)	электрический маленький
4	apple iphone 7	(нет)	apple	iphone 7
5	футболки мужские с принтом	футболки	(нет)	мужские с принтом

Таблица 2: Примеры запросов и соответствующих им смысловых частей

Исходя из описанной выше модели данных в базе магазина и предполагаемой стуктуры запроса пользователя, поставим задачу построения поисковой системы следующим образом. В запросе пользователя необходимо выделялить следующие сущности: название товара, название бренда и уточняющую информацию. После выделения этих компонент, соответствующим образом ведется поиск по базе товаров магазина. К примеру, если выделен и бренд B, и название товара I, то пользователю выдается некоторая выборка товаров, имеющих бренд B и название товара I (это легко сделать, поскольку мы предположили, что данные признаки уже выделены в базе данных интернет-магазина). Если же выделено, к примеру, только название товара I, то стоит показать пользователю выборку товаров с названием I, вне зависимости от бренда. Данная задача давно известна в области обработки естественного языка (natural language processing, NLP) как распознавание именованных сущностей (named entity recognition, NER).

Отдельного внимания заслуживает уточняющая информация. Поскольку данная компонента может изменяться в слишком широких масштабах (так как для разных категорий товаров подробности описания могут сильно отличаться, к примеру для чайников важен объем, а для планшетов – размер экрана), попытки привести её в унифицированный вид могут быть достаточно затруднительными (подробнее об этом будет сказано в следующем подразделе). В связи с этим данная информация должна учитываться в поисковой выдаче, которяа была отфильтрована названием товара и бренда, некоторым особым образом.

3 Данные

3.1 Особенности сбора данных

Проведение некоммерческих исследований в рассматриваемой в данной работе области затрудняется в связи с высокой коммерциализацией. Необходимые для постоения поисковых систем данные могут быть получены единственным образом: путем выгрузки данных из баз реальных интернет-магазинов. Чаще всего компании не стремятся делиться подобной информацией, поскольку она может быть использована конкурентами. Из постановки задачи следует, что для построения поисковой системы необходимы как минимум два набора данных: база товаров, а также примеры запросов пользователей. Опишем затруднения, которые возникли во время сбора указанных датасетов.

База товаров. На первый взгляд может показаться, что собрать такие данные довольно просто, ведь все интернет-магазины имеют каталоги товаров, а для каждого товара указаны как минимум название, описание, категория и чаще всего бренд. Однако на практике подавляющее большинство компаний стараются защитать свои сайты от автоматического сбора данных (скрапинга, парсинга). Безусловно, такие защиты можно обойти, но при этом они достаточно сильно затрудняют процесс получения необходимых датасетов. Кроме того, разные категории товаров описываются значительно отличающимися друг от друга наборами признаков. К примеру, для одежды важны такие параметры как размер, цвет, материал, бренд и страна-производитель, а для смартфона – мощность процессора, объем памяти, размеры экрана, бренд, модель и т.д. Таким образом, при попытке сбора максимально детализированной информации о товарах (то есть когда товар описывается большим количеством признаков), гипотетический исследователь может столкнуться с тем, что в собранных им данных количество признаков невероятно велико, и для получения хоть сколько-то репрезентативной выборки необходимо спарсить практически все товары с сайта магазина, что в условиях некоммерческой разработки просто невозможно (из-за описанных выше проблем с защитой сайтов). Именно этот факт мотивирует предложенную в предыдущем разделе модель данных для базы товаров. Однако, если бы построение поисковой системы ведет уже существующий интернет-магазин с

большим количеством детализированных примеров, модель с более подробным описанием каждого товара была бы возможна (и, скорей всего, более успешна).

Запросы пользователей. Если сведения о товарах, хоть и с трудом, но можно получить из открытых источников, то текстовые запросы клиентов в поиковую строку интернет-магазина являются полностью закрытой для сторонних лиц информацией. Автор данной работы прговел в поисках открытых данных подобного рода более двух недель, и ничего похожего найдено не было. Однако был найден способ добывать текстовые запросы, наиболее похожие на необходимые. Довольно часто пользователи ищут товары не непосредственно на сайтах интернет-магазинах, а вводят свой запрос в поисковые системы вроде Google или Яндекса. При этом, если взять запросы, по которым совершаются переходы на крупные интернет-магазины (или их аггрегаторы), какие как Ozon, Яндекс.Маркет и т.д., то окажется, что почти все такие запросы представляют собой как раз те самые текстовые строки, которые нам нужны (быть может, с некоторыми дополнительными словами, к примеру "купить"). Однако существуют сервисы (к примеру, ahrefs.com), которые позволяют анализировать и предсказывать переходы на сайты с поисковых систем. При этом подобные сервисы часто продают подобные данные. Именно таким образом можно собрать некоторую базу поисковых запросов в интернет-магазинах, не обращаясь к компаниям напрямую.

3.2 Описание собранных наборов данных

База названий товаров. Часть данных была получена путем сбора названий, категорий и брендов с сайтов крупных интернет-магазинов Ozon, BERU, Яндекс.Маркет, Юлмарт. С указанных площадок было собрано порядка 450.000 наименований. Отметим, что для построения основной компоненты поисковой системы (а именно глубокой нейронной сети для задачи NER) будут использовать лишь текстовые строки, содержащие всю информацию о товаре (то, что выше было названо описанием товара). Значительная часть товаров принадлежит разделу «Электроника». Пример данных из данной части выборки – в таблице 3.

№	Описание товара
1	Смартфон Sony Xperia Z2 D6503 Purple
2	Чайник Bosch TWK7808, золотистый
3	a4tech / Мышь Bloody P80 Pro
4	Игровая консоль PlayStation 5, белый

Таблица 3: Примеры товаров с сайтов с Ozon, BERU, Яндекс.Маркет, Юлмарт

Помимо собранных данных из открытых источников, были использованы данные, предоставленные на соревновании DataFusion, проходившем на площадке boosters.pro с 27 января 2021 года по 27 марта 2021 года. В задаче требовалось определить бренд товара по его текстовому описанию в чеке. Всего в датасете были представлены 3.1 млн подобных текстовых строк. Основная часть наименований является названиями продуктов питания, однако встречаются товары и из других разделов. Некоторые примеры из этого набора данных представлены в таблице 4.

№	Описание товара
1	20 БРЮКИ МУЖ / J Lin / 72 МС 1
2	"Русский Аппетит "Суп борщ 50 гр .
3	Йогурт Даниссимо фантазия 105 г хрустящие шарики
4	Штора рулонная "Декор белый 57 * 160 см

Таблица 4: Примеры товаров из даннных с соревнования DataFusion

Запросы пользователей. Изначально, запросы собирались с помощью метода описанного в предыдущем подразделе. Таким образом было собрано порядка 50.000 запросов среднего качества. Однако, к счастью, автору работы удалось связаться с тим лидом интернет-магазина KazanExpress (https://kazanexpress.ru) Юрием Гаврилиным, который на безвозмедной основе передал для проведения некоммерческих исследований датасет запросов пользователей размером 3.7 млн записей. Автор приносит благодарность компании KazanExpress и, в частности, Юрию Гаврилину, за оказанную поддержку. Примеры запросов – в таблице 5.

№	Запрос
1	ботинки рабочие
2	чехол для iphone
3	redmi note 9
4	сменная головка для бритвы philips
5	капсулы для кофемашин bosch tassimo

Таблица 5: Примеры запросов в поисковую систему KazanExpress

Для решения задачи распознавания именованных сущностей на представленных запросах необходима разметка: для каждого слова в обучающем датасете нужно указать, является ли слово брендом, названием товара или не относится к этим классам. Как было отмечено, в открытом доступ подобных данных нет, поэтому для выполнения разметки автор обратился в компанию, предоставляющую подобные услуги. Поскольку данное исследование является некоммерческим проектом, бюджет был существенно ограничен, поэтому итоговый размер обучающего набор составил 10000 сэмплов. Автор выражает благодарность компании LabelMe за льготные условия при выполнении данной задачи. Пример размеченных данных - в таблице 6. После каждого слова в скобках указана метка класса. Обозначения: O - отсутствие класса, I - название товара, B - название бренда.

№	Запрос
1	четырехгранный(O) ключ(I)
2	подвеска(I) с(O) цепочкой(O)
3	кроссовки(I) найк(B)
4	natura(B) siberica(B)
5	xiaomi(B) mi(O) очки(I)

Таблица 6: Примеры размеченных запросов

4 Обзор существующих работ

Приведем краткий обзор некоторых работ, связанных с данным исследованием.

4.1 Подход на основе рекуррентных нейросетей

Наиболее близкой по задаче и методу решения является статья [1], выпущенная командой сотрудников компании The Home Depot (https://www.homedepot.com/), которая занимается продажей товаров для дома, ремонта и строительства. Авторы рассматривают в точности такую же постановку задачи, как и в данной работе (классификация слов в запросе пользователя на название товаров, бренд и прочие слова). В рамках статьи они представили цельное (end-to-end) решение. Было предложено, как оптимальным образом организовать создание обучающих датасетов, была обучена нейросеть для распознавания именованных сущностей, а также представлены некоторые показатели метрик качества. Рассмотрим подробнее каждый из этих этапов.

Подготовка данных. С учетом того факта, что в качестве модели классификации будет использовать глубокая нейросеть, авторы статьи выдвигают три требования к обучающему датасету.

1. Большое количество данных
2. Высокое качество разметки
3. Широкое покрытие меток классов (названий товар и брендов)

Идеальным решением для такого перечня требований стал бы большой набор данных, размеченных вручную. Однако разметка текстовых данных (особенно с учетом того факта, что иногда потребуется дополнительно проводить поиск: является ли некоторое словыо брендом или неизвестным разметчику товаров) является достаточно сложной и дорогостоящей процедурой. В связи с этим, авторы предлагают составить обучающих датасет из трех компонент:

1. Размеченные автоматическими эвристическими методами данные
2. Размеченные вручную запросы
3. Запросы, состоящие только из бренда или названия товара

В первом компоненте для автоматической разметки используются детерминированные алгоритмы, основанные на сопоставлении в запросах пользователей слов с названиями брендов и товаров. Вторая часть, как и следует из названия, состоит из размеченных человеком данных, однако, как уже было отмечено, этот процесс трудозатратен, поэтому данная компонента сравнительно невелика в размерах. Третья часть призвана отвечать третьему требованию, а именно пополнять «кругозор» модели редко встречающимися товарами и брендами. Подробнее о том, как именно будут использоваться три указанные поднабора данных, будет сказано ниже.

Модель. В ходе экспериментов авторы тестировали различные нейросетевые архитектуры, в основе которых лежат рекуррентные нейросети (recurrent neural networks, RNN). Ещё с конца прошлого века было известно, что классические RNN обладают рядом недостатков: при их обучении стохастические градиенты часто, проходя через слои сети, начинают быстро затухать (в связи с чем модель «забывает» контекст очередного входа довольно быстро), а иногда, наоборот, «взрываются», когда модуль градиента экспоненциально увеличивается на каждом шаге алгоритма обратного распространения. В связи с обозначенными проблемами, на практике чаще всего применяются модификации RNN: LSTM (Long-Short Term Memory) [4] и GRU (Gated Recurrent Unit) [2]. В задачах, в которых заранее целиком известна входная последовательность (например, NER, но не предсказание временных рядов), рекуррентные модели также часто модифицируют до их двунаправленных версией: BiLSTM (Bidirectional LSTM) и BiGRU (Bidirectional GRU). Основным отличием является наличие в архитектуре двух нейросетей, в одну из которых входная последовательность поступает в прямом порядке, а в другую — в обратном. Это способствует улавливанию модели двухстороннего контекста очередного элемента входа.

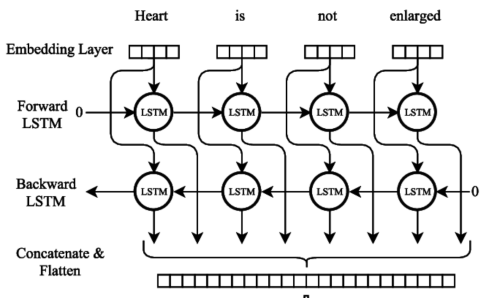


Рис. 1: BiLSTM

Для построения финальной модели, авторы рассматриваемой статьи комбинируют указанные выше двунаправленные архитектуры. А именно, для составления векторных представлений слов (эмбеддингов) в предложении используется посимвольная BiLSTM (рис. 2), а для решения задачи распознавания именованных сущностей BiGRU с CRF[5] (Conditional Random Field) слоем (рис. 3), на вход которой подается последовательность токенов, полученных из слов после прогонки через первую нейросеть.

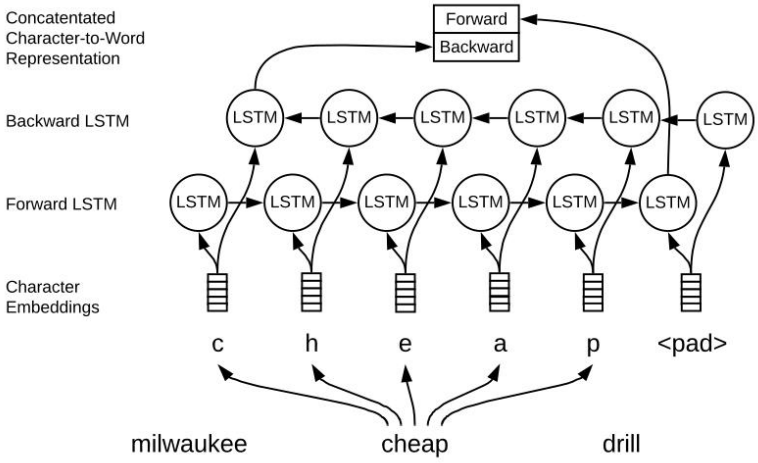


Рис. 2: Посимвольная BiLSTM нейросеть для получения эмбеддингов

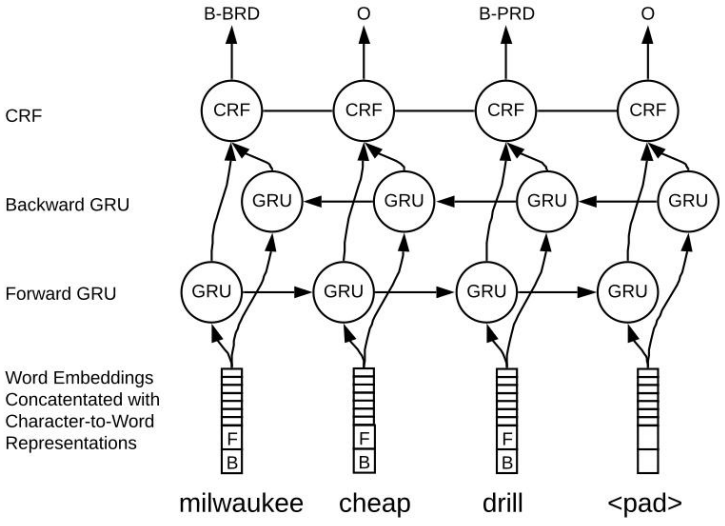


Рис. 3: BiGRU-CRF для NER

Итеративный процесс обучения. После того как были описаны входные датасеты и модель, мы можем рассмотреть предлагаемый авторами статьи процесс обучения итоговой системы. Все её компоненты представлены на рисунке 4.

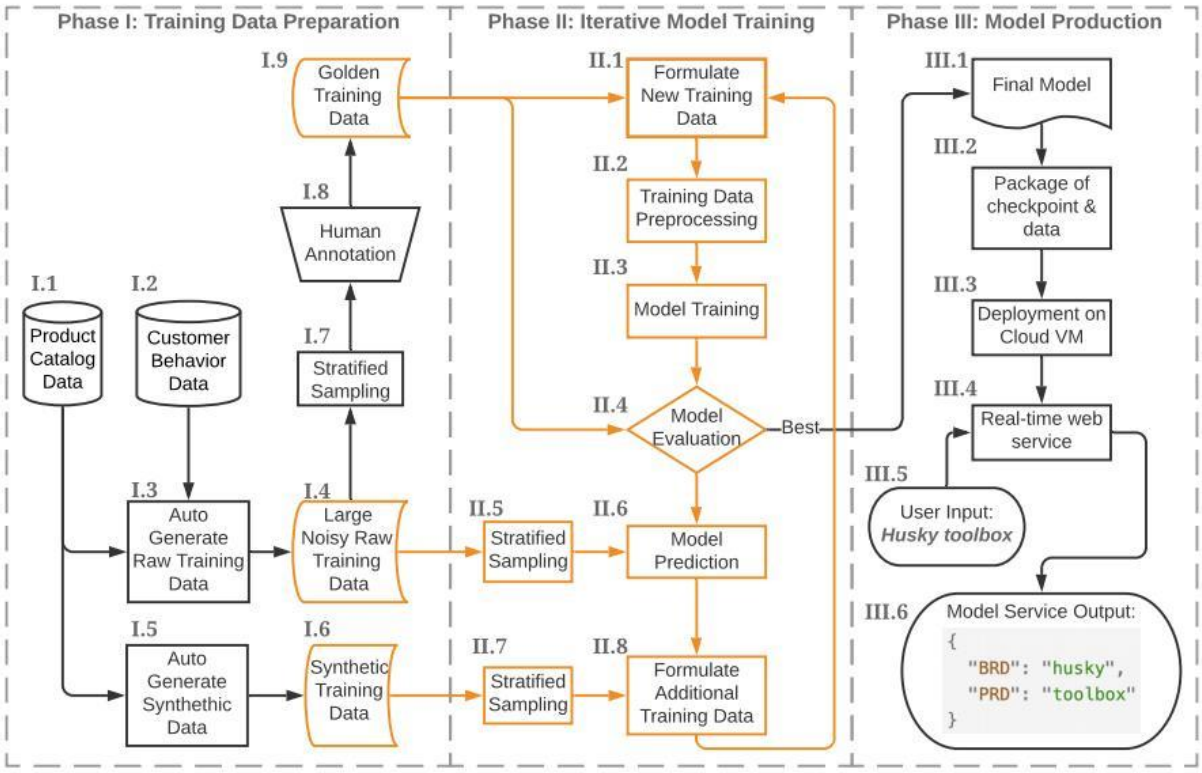


Рис. 4: Левый столбец — датасеты. Центральный — цикл обучения. Правый — особенности развертки модели в бизнес.

Первая итерация начинается с вручную размеченного датасета («золотого» датасета, так как данные размечать дорого) (I.9). 15% от этих данных случайным образом откладываются в качестве тестовой выборки; остальные данные случайным образом разделяются на обучающие (90% оставшихся данных) и валидационные («dev») данные (10%). Затем данные для обучения предобрабатываются (II.2). На следующем этапе (II.3) модель обучается до тех пор, пока F1-мера качества не перестанет улучшаться на валидационной выборке, либо пока не будет достигнуто максимальное число итераций (эпох). Далее выполняется оценка качества на тестовой выборке (II.4). Если выполнен некоторый критерий остановки (достигнуто максимальное число итераций, или модель достигла необходимого качества), то процесс завершается. Иначе алгоритм переходит на шаги расширения обучающего датасета (II.6 и II.7).

На этапе II.6 из составленной с помощью эвристических алгоритмов компоненты (I.4) данных случайным образом сэплируется некоторая подвыборка. Текущая модель делает предсказание меток классов. Если метки, полученный эвристическим алгоритмом, а также метки модели совпадают, то такой сэмпл добавляется в обучающий датасет. На этапе II.7 к нему случайным образом дополнительно добавляются сэмплы, состоящие только из бренда или названия товара.

Результаты. На рис. 5 изображена таблица, представленная авторами в статье. Как было отмечено выше, лучшие результаты показала архитектура, использующая 2 нейросети: посимвольная BiLSTM для построения векторных представлений слов и BiGRU-CRF для непосредственного распознавания именованных сущностей. В качестве метрики используется F1-мера (среднее гармоническое между точностью и полнотой классификации слов в запросе, усредненное по всем запросам в соответствующем наборе данных).

models	char emb.	dev F1	test F1
BiLSTM	No	85.77	85.05
	Yes	86.99	86.23
BiLSTM-CRF	No	87.69	86.72
	Yes	88.57	88.44
BiGRU	No	85.42	85.57
	Yes	86.53	87.09
BiGRU-CRF	No	87.12	87.04
	Yes	**88.71**	**88.82**

Рис. 5: Результаты авторов статьи

4.2 BERT

Ключевым отличием данной работы от исследования, описанного в предыдущем разделе, является использование более современной нейросетевой архитектуры, которая называется BERT (Bidirectional Encoder Representations from Transformers). Архитектура была предложена[3] в 2018 году и стала активно применяться во многих задачах обработки естественного языка, для решения которых используются нейросети. Поскольку данная модель лежит в основе предлагаемого в данной работе решения, кратко опишем основные её принципы работы.

Механизм внимания. Кек123

5 Предложенная архитектура

todo

6 Вычислительные эксперименты

Цель данного раздела: продемонстрировать, что предложенная теория работает на практике; показать границы её применимости; рассказать о новых экспериментальных фактах.

Чисто теоретические работы могут вообще не содержать раздела экспериментов (не работает, ну и не надо — зато теория красивая). Кстати, теоретики имеют право не догадываться, где, кому и когда их теории пригодятся.

6.1 Исходные данные и условия эксперимента

Описывается прикладная задача, параметры анализируемых данных (например, сколько объектов, сколько признаков, каких они типов), параметры эксперимента (например, как производился скользящий контроль).

6.2 Результаты эксперимента

Результаты экспериментов представляются в виде таблиц и графиков. Объясняется точный смысл всех обозначений на графиках, строк и столбцов в таблицах.

6.3 Обсуждение и выводы

Приводятся выводы: в какой степени результаты экспериментов согласуются с теорией? Достигнут ли желаемый результат? Обнаружены ли какие-либо факты, не нашедшие объяснения, и которые нельзя списать на «грязный» эксперимент?

Обсуждаются основные отличия предложенных методов от известных ранее. В чем их преимущества? Каковы границы их применимости? Какие проблемы удалось решить, а какие остались открытыми? Какие возникли новые постановки задач?

7 Заключение

В квалификационных работах последний раздел нужен для того, чтобы конспективно перечислить основные результаты, полученные лично автором.

Результатами, в частности, являются:

- Предложен новый подход к...
- Разработан новый метод..., позволяющий...
- Доказан ряд теорем, подтверждающих (опровергающих), что...
- Проведены вычислительные эксперименты..., которые подтвердили / опровергли / привели к новым постановкам задач.

Цель данного раздела: доказать квалификацию автора. Даже беглого взгляда на заключение должно быть достаточно, чтобы стало ясно: автору удалось решить актуальную, трудную, ранее не решённую задачу, предложенные автором решения обоснованы и проверены.

Иногда в Заключении приводится список направлений дальнейших исследований.

Список литературы

[1] Xiang Cheng, Mitchell Bowden, Bhushan Ramesh Bhange, Priyanka Goyal, Thomas Packer, and Faizan Javed. An end-to-end solution for named entity recognition in ecommerce search. 2020.

[2] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. 2014.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[5] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. 2015.