Федеральное государственное автономное образовательное
учреждение высшего образования
Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

Выпускная квалификационная работа
на тему

Распознавание именованных сущностей для языков с малыми ресурсами

Выполнила студентка группы БПМИ151, 4 курса,
Закирова Ксения Игоревна
Научный руководитель:
Доцент, кандидат технических наук,
Артемова Екатерина Леонидовна

Москва, 2020

Содержание

1 Введение

Распознавание именованных сущностей (Named entity recognition, NER) это одна из задач обработки естественного языка; задача обнаружения и классификации слов в тексте на несколько заранее определённых категорий, таких как, например, люди, места, организации и т.д. Распознавание именованных сущностей имеет множество применений, используется в автоматическом разделении на категории текстов, рекомендательных системах, системах извлечения информации. Как задача распознавание именованных сущностей была сформулирована ещё в прошлом веке, однако широкое распространение получила только в последнем десятилетии. Развитие глубоких нейронных сетей дало значительный толчок развитию обработке естественных языков, и, как следствие, задаче распознавания именованных сущностей. Были изобретены более эффективные и точные модели, которые показывают хорошие результаты. Однако существуют и серьёзные проблемы, связанные с данной задачей. Во-первых, упомянутые выше модели с хорошими результатами существуют только для широко распространённых языков, для которых имеются размеченные корпусы, а языки, которые не входят в «топ-10» по числу носителей, оказываются за бортом. Во-вторых, у компаний и исследователей нет причины вкладываться в задачу распознавания именованных сущностей для языков с малыми ресурами, так как, скорее всего, это не сможет принести большой выгоды в дальнейшем из-за сравнительно небольшого числа носителей (как и было сказано ранее). В то же время у меня есть причина: татарский язык является родным языком для меня, и я стараюсь сохранять его и продвигать его значимость, в том числе и с помощью такой работы.

Помимо патриотических мотивов существуют и мотивы прагматические: развитие распознавание именованных сущностей для татарского языка может быть использовано для всей кыпчакской группы тюркской ветви языков (татарский, башкирский, карачаево-балкарский, казахский, киргизский и др.). Это связано с тем, что языки тюркской ветви достаточно похожи между собой, как грамматически, так и лексически, как следствие, решение задачи для одного языка скорее всего будет иметь неплохие шансы и для других языков данной группы. К сожалению, это не сработает для турецкого языка, во-первых, потому что он относится к огузской группе, во-вторых (и это главная причина): там используется другой алфавит. Тюркская ветвь включает себя языки с различной письменностью, что усложняет возможность экстраполяции модели на «похожие» языки.

Также на тему конкретно татарского языка существует работа [**?**] иссле-

дователей из Академии наук Республики Татарстан. Далее я более подробно рассмотрю их работу в своем исследовании.

TODO Введение. В нем дается описание предметной области, актуальность и значимость работы, цель и задачи работы, неформальная и формальная постановка задачи, основной результат, структура работы

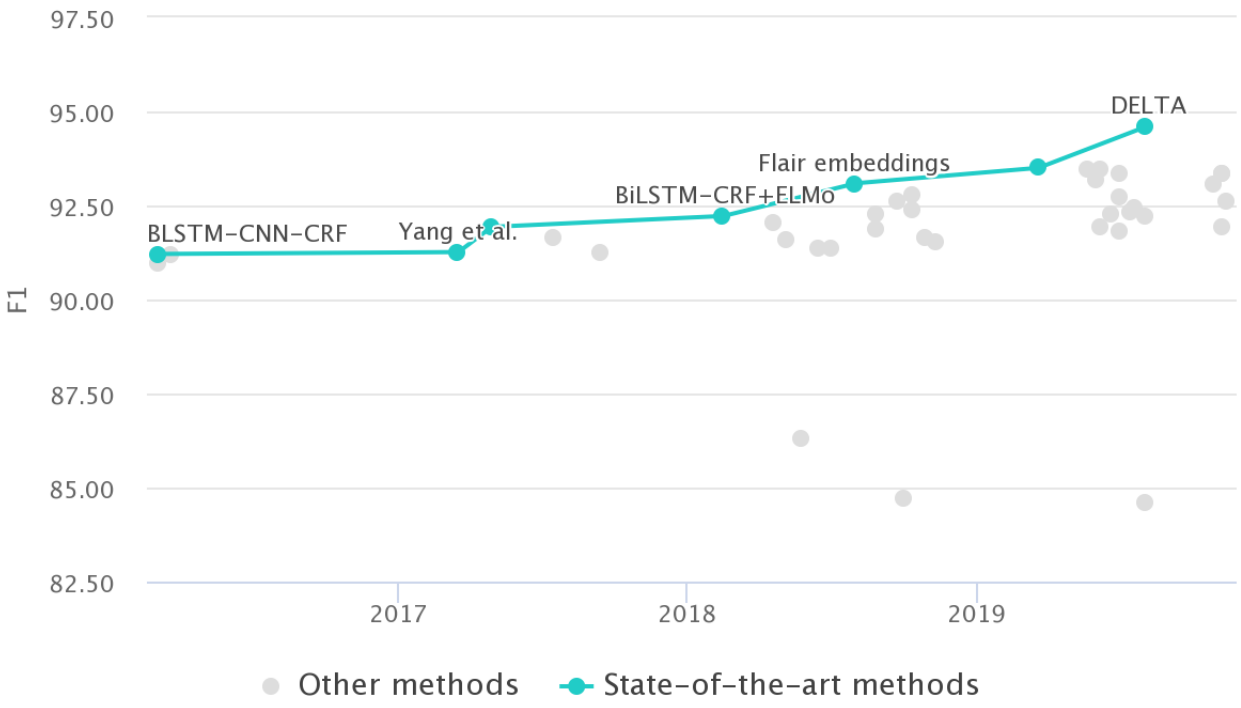
2 Обзор литературы

2.1 Стандартные подходы к распознаванию именованных сущностей

Тут про модели, которые люди используют: BERT, CRF, LSTM и прочие.

На текущий момент лучшими моделями на классическом датасете CoNLL 2003 по оценке сайта paperswithcode.com является Delta [?] (модель BERT), также высокие места занимают такие модели как CNN [?], GCDT [?], I-DARTS + Flair [?], LSTM-CRF [?].

Рис. 2.1: Лучшие модели для задачи распознавания именованных сущностей на датасете CoNLL 2003



Для более глубокого погружения в тему рекомендуется ознакомится со всеми моделями, приведенными выше; я же проведу краткий обзор по тем моделям, которые были использованы в работе.

2.1.1 Low-Resource Named Entity Recognition with Cross-Lingual, Character-Level Neural Conditional Random Fields

[?]

Статья от Ryan Cotterell и Kevin Duh, где представлена модель Conditional Random Fields (CRF) и представлены способы её улучшения с помощью таких надстроек, как обучение модели на языках с бОльшими ресурсами, а потом применение её к языку из того же семейства, но с меньшими ресурсами В данной работе рассматривались семья индоевропейских, ветви: романская, германская, славянская, индоарийская; и семья автронезийских, ветвь: филиппинская.

2.1.2 A Neural Layered Model for Nested Named Entity Recognition

[?]

Статья от Meizhi Ju, Makoto Miwa и Sophia Ananiadou. В данной работе представляется модель Layered-BiLSTM-CRF, которую я использовала в своей работе. Исследователи представляют модель, которая работает с «наслоенными» именованными сущностями, т.е. когда одна именованная сущность частично или полностью входит в другую именованную сущность. Используются последовательно идущие плоские слои, слой состоит из BiLSTM и поверх него один слой CRF.

2.1.3 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[?]

Данная статья от Google AI Language, которая на текущий момент является лучшей текущей моделью практически на всех популярных бенчмарках обработки естественного языка. Она включает в себя очень много возможностей для различных задач обработки естественных языков, которые на данный момент пользуются большой популярностью. Для конкретно моей задачи очень хорошо подходит претренированная модель bert-base-multilingual-cased, в обучающие данные которой входила также и Википедия на татарском языке.

Несмотря на популярность BERT, даже дообучать данную модель довольно сложно из-за её размеров. Спасибо Высшей Школе Экономики за возможность использования кластера для обучения моделей для выпускной квалификационной работы.

BERT — это очень большая модель, которая, однако, не решает задачи распознавания именованных сущностей сама по себе, поэтому я воспользовалась библиотекой transformers от huggingface [?]

2.1.4 Huggingface Transformers

Данная библиотека предоставляет возможность использовать различные лучшие на данный момент модели (не только BERT, но и многие другие) для решения различных задач обработки естественного языка, в том числе и распознования именованных сущностей, что и является моей задачей. Их репозиторий на github содержит множество примеров для удобного использования их библиотеки.

2.2 Стандартные подходы к разметке данных

Стандартным (самой распространённым) форматом разметки для корпусов текста для задачи распознавания именованных сущностей является разметка IOB (сокр. от Inside–outside–beginning). Она была представлена в работе Text Chunking using Transformation-Based Learning [?]. Данный формат имеет три префикса:

- B префикс перед тегом указывает, что тег находится в начале чанка (в нашем случае именованной сущности)
- I префикс перед тегом указывает, что тег находится в продолжении чанка.
- O префикс указывает, что данное слово не относится ни к какому чанку.

Теги могут быть различными; устанавливаются на усмотрение исследователя, примеры тегов: PER (персона), LOC (географический объект), ORG (организация), TIM (время) и другие. Разметка в тексте выглядит следующим образом:

B-PER	I-PER	O	O	B-LOC	I-LOC	O
Иван	Петров	проживает	в	Российской	Федерации	.

Что-то я не знаю, что ещё сюда написать.

2.3 Работы, связанные с распознаванием именованных сущностей в малоресурсных языках

2.3.1 Datasets and Baselines for Named Entity Recognition in Armenian Texts

Тема моей работы очень близка к теме работы данных исследователей, за исключением языка: у них, как понятно из названия, армянский язык, который так же относится к языкам малой языковой группы.

В отличие от моего случая, где существует релевантная работа, поднимавшая раньше тему моей работы, Т. Гукасян, Г. Давтян, К. Аветисян и И. Андрианов стали, можно сказать, первопроходцами в своей области, поскольку никто не делал подобных работ для армянского языка. У них не было подобранного и размеченного корпуса текста, поэтому, помимо распознавания именованных сущностей, они занимались также и сбором и разметкой данных. Их модель включала в себя CRF, которую я использую и в своей работе, и рекомендую как хорошую модель для языков с малыми ресурсами.

В своей работе исследователи не использовали BERT, поскольку это относительно новая модель, а статья вышла в конце 2018 года.

2.4 Работы, связанные с распознаванием именованных сущностей в татарском языке

При поиске корпусов на татарском языке я нашла корпус Туган Тел — работу Невзоровой и др. [**?**]. К сожалению, других релевантных данных найдено не было.

2.4.1 Developing Corpus Management System: Architecture of System and Database

Туган Тел – это корпус текстов на татарском языке, разработанный Институтом прикладной семиотики Академии наук Республики Татарстан. Корпус предназначен для широкого круга пользователей: лингвистов, специалистов в татарском языке, преподавателей татарского и всем тем, кому может понадобиться набор текстов на татарском языке. Основными функциями корпуса являются: поиск по словоформе, лемме (лексеме), набору морфологических параметров. Существует система «корпус-менеджер», которая поддерживает данные функции. На данный момент существует проект разработки электронного корпуса, который также включает в себя автоматическую разметку корпуса. Корпус включает в себя татарские тексты различных жанров, такие как художественная литература, тексты СМИ, тексты официальных документов, учебная литература, научные публикации и др. Каждый документ имеет метаописание, включающее в себя автора и его пол, выходные данные, дату создания, жанр, части, главы и др. Тексты, включенные в корпус, снабжены автоматической морфологической разметкой, которая включает в себя информацию о части речи и грамматической характеристики словоформы. Морфологическая разметка текстов корпуса выполняется автоматически с использованием модуля двухуровневого морфологического анализа татарского языка, реализованного в программном инструментарии PC-KIMMO, с чем связан ряд проблем в использовании данного корпуса, о которых я

скажу в основной части работы. На декабрь 2019 года в корпусе 194 млн. словоформ.

Проведя обзор литературы я нашла только одну статью, релевантную к моей теме.

2.4.2 Named Entity Recognition in Tatar: Corpus-Based Algorithm

Самая близкая к моей работе это статья «Named Entity Recognition in Tatar: Corpus-Based Algorithm» от О. Невзоровой, Д. Мухамедшина и А. Галиевой, Академия наук Республики Татарстан. В статье они предлагают алгоритм разметки корпусов, используя в качестве примера корпус «Туган Тел» [**?**], использовав следующие категории: книги, рестораны, фильмы, журналы, компании, аэропорты, корпорации, языки, колледжи, университеты, школы, магазины, музеи и больницы.

1. **Использованные данные**

   Исследователи использовали корпус Туган тел [**?**], о котором я говорила выше.

2. **Разбор алгоритма, предложенного в статье:**

   Представленный алгоритм основан на идее сравнения частотности n-грамм. Сравнение происходит на всём объёме корпуса, что увеличивает точность результата, заявляют авторы статьи. Алгоритм является итеративным, причём количество итераций определяется пользователем.

   Нулевым шагом алгоритма является выборка по поисковому запросу. Запрос может представлять собой форму слова, лемму, фразу или поиск по морфологическим параметром. Выборка представляет собой набор биграмм и их количество вхождений в текст. В биграмме одно слово является запросом, а второе ищется по корпусу с (опционально) морфологическими параметрами. Далее полученный список из запроса просматривается глазами и из него убирается мусор. Полученная «чистая» выборка используется для первого шага алгоритма.

   Полученный список биграмм ищется в корпусе и к нему добавляется третье слово, которое стоит с ним рядом в тексте; добавляться слово может слева или справа, данный параметр выбирается пользователем. Полученный список триграмм отсортировывается по частоте вхождений в корпус и в выборке остаются только самые частотные (например, первые 95%, в статье этот параметр обычно был равен 80%). Порог отсечения (в статье он называется «индекс покрытия», «covering index») более частотных вхождений также выбирается пользователем. Урезанный по

порогу список триграмм используется как входные данные для второй итерации алгоритма: каждая триграмма ищется по корпусу как фраза и, аналогично первой итерации, составляются 4-граммы и их частоты. Точно так же выбираются самые частотные 4-граммы (четвертое слово может добавляться справа или слева), список обрезается по пороговому значению и, при желании, алгоритм продолжается дальше, используя на вход уже список 4-грамм.

Таким образом алгоритм использует n-граммы для поиска $(n+1)$-грамм, некоторые из которых будут отсечены порогом, а остальные использованы в следующем шаге алгоритма.

3. **Окончание алгоритма:**

   Существует такое понятие как «точность сравнения» («accuracy of matching») P, которое задаётся пользователем в процентах. Если частота n-граммы меньше P от количества найденных $(n+1)$-грамм, то алгоритм прекращает увеличивать длину именованной сущности, иначе алгоритм переходит на следующую итерацию. Таким образом, в финальный результат входят самые стабильные n-граммы разной длины, включая результаты поиска изначального поискового запроса.

   Стоит отметить, что все сущности, выделенные на нулевом шаге алгоритма, так или иначе считаются именованными сущностями; вопрос только в том, сколько слов справа или слева к этой именованной сущности добавится. Если алгоритм перешёл от n-грамме к $n+1$-грамме, то n-грамма не входит в финальный результат.

   Запрос извлечения именованных сущностей представляет собой кортеж (1), где Q_1 и Q_2 — запрос в корпус-менеджер Туган Тел[**?**], L, R это, соответственно, порог ограничения итераций добавления слов слева и справа, C — порог отсечения частотности на каждой итерации (covering index), P — порог для принятия решения о включении фразы в итоговый список именованных сущностей (accuracy of matching).

$$Q = (Q_1, Q_2, L, R, C, P)$$

4. **Эксперименты:**

   Исследователи перечисляют довольно много категорий, над которыми они экспериментировали, но результаты они показали на словах «министерство», «улица», «язык», «ресторан» и «корпорация».

Также в данной статье очень интересный способ оценки результатов. Стандартные accuracy, precision и recall (и производная от них F-score) в статье не упоминается, но оценивание полученных результатов производится. Происходит это следующим образом: вручную просматриваются все полученные n-граммы и классифицируются: на именованные сущности, «требует дополнительной очистки, тогда станет именованной сущностью», «требует расширения, тогда станет именованной сущностью», «это именованная сущность, но требует другой тег», «это именованная сущность, но требует дополнительной очистки и другой тег» и некорректные, см. таблицу 2.1. Данное оценивание не позволяет мне сравниваться с результатами Невзоровой и др., так как в моей работе поставлена другая задача.

TODO номер таблицы?

Class of named entity	Correct	Require filtering	Require expansion	Correct names of subclasses	Names of subclasses that require filtering	Incorrect	Total
Names of ministries	100%	0%	0%	0%	0%	0%	50
Street names	72%	12%	0%	0%	0%	16%	600
Language names	53.5%	0%	0%	0%	0%	46.5%	471 (2310)
Restaurant names	37.7%	18.3%	0%	13%	15.9%	15.1%	285
Corporation names	45.7%	19.6%	10.9%	21.7%	0%	2.2%	138

Таблица 2.1: Таблица 3 из статьи [**?**]

2.4.3 Выводы

В области распознавания именованных сущностей написано много статей и изобретено много моделей, показывающих хорошие результаты на распространённых языках. Существуют так же работы по теме распознавания именованных сущностей для малоресурсных языков. Академия наук Республики

Татарстан начала работу в данном направлении для татарского языка; я же, воспользовавшись их результатами, размечу корпус текстов на татарском языке, применю существующие модели к имеющимся данным и сравнюсь с результатами алгоритма Невзоровой и др.

3 Методология

Целью работы было получить размеченный корпус и обученную модель, распознающую именованные сущности. После обзора литературы были намечены задачи и работа была предварительно разделена на несколько этапов.

1. Получение и разметка данных
2. Обучение и тюнинг моделей
3. Сравнение результатов

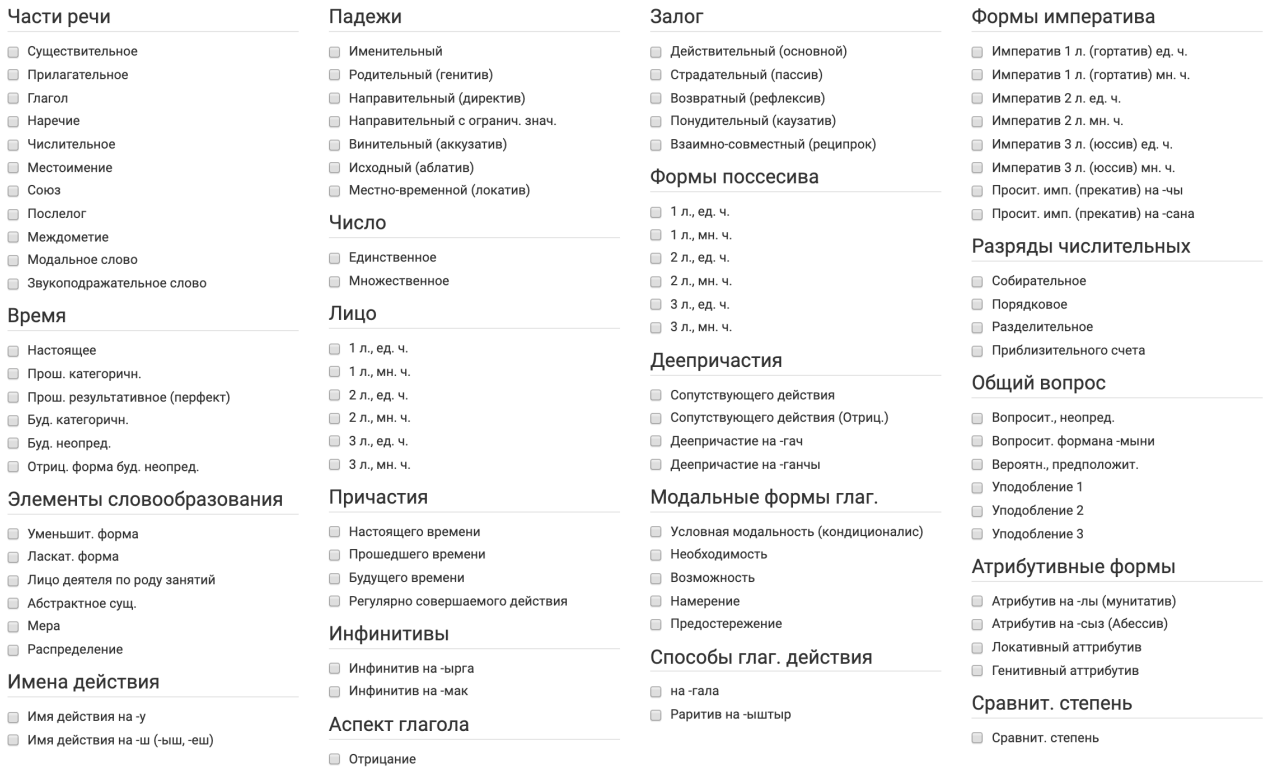
Но в течение работы по нескольким причинам были внесены корректировки. Во-первых, как я упоминала ранее в обзоре литературы, с представленными результатами в статье Невзоровой невозможно сравниваться, поскольку цели моей и их работ различаются. Во-вторых, качество полученных данных оказалось не лучшим из возможных, а алгоритм Невзоровой, разработанный как раз для разметки данных, мог бы улучшить имеющийся корпус, используемый для обучения моделей. Как следствие, было принято решение воспроизвести алгоритм из статьи Невзоровой насколько это возможно и воспользоваться полученными результатами.

1. Получение и разметка данных
2. Обучение и тюнинг моделей
3. Воспроизведение статьи Невзоровой
4. Разметка данных с помощью алгоритма Невзоровой
5. Обучение и тюнинг моделей
6. Сравнение результатов

3.1 Получение и разметка данных

Обзор литературы показал, что существует корпус татарских текстов Туган Тел[?]. Данный копрус имеет также свою систему «корпус-менеджер»,

Рис. 3.1: Параметры на сайте tugantel.tatar для поиска по корпусу



которая представлена в виде сайта. На этом сайте можно искать по словоформе или лемме с огромным количеством параметров [3.1], однако возможности просто скачать весь корпус не оказалось. Я предполагаю, что у Академии наук Республики Татарстан есть API для исполнения запросов на большом количестве данных и в каком-то более удобном формате, чем запрос на сайте, но у меня доступа к такому ресурсу нет.

Я связалась с Невзоровой по указанной в статье электронной почте, чтобы узнать подробности об их работе и попросить о сотрудничестве. Невзорова ответила на моё письмо и предоставила мне доступ к корпусу.

Корпус представляет из себя *.zip* файл, состоящий из 7557 *.txt* файлов, в общей сложности весом 1 183 023 978 Б. Как уже упоминалось ранее, корпус Туган Тел автоматически размечен с помощью программного инструментарии PC-KIMMO. Разметка выглядит следующим образом: 3.1 3.2

Первое слово в каждом файле распознано как Error, все русские слова не распознаны (а их в татарском языке некоторое ненулевое количество, так как происходит какое-то достаточное количество заимствований). К сожалению для меня, очень часто русскоязычные слова оказывались как раз именованными сущностями, такими как, например, названия улиц, но распознаны они были как Error, что очень печально.

*TODO: написать % Error от общего количества слов и привести пару

Рис. 3.2: Пример случайного предложения из корпуса Туган Тел

```
Аның
аны+PN+POSS_2SG(ың)+Nom;аның+PN;ул+PN+GEN(ның);
дөньяга
дөнья+N+Sg+DIR(ГА);
күз
күз+N+Sg+Nom;
карашы
караш+N+Sg+POSS_3(СЫ)+Nom;
хаман
хаман+Adv;
үзгәрми
үзгәр+V+NEG(мА)+PRES(Й);
.
Type1
```

Аның	аны+PN+POSS_2SG(Ың)+Nom;аның+PN;ул+PN+GEN(нЫң);
дөньяга	дөнья+N+Sg+DIR(ГА);
күз	күз+N+Sg+Nom;
карашы	караш+N+Sg+POSS_3(СЫ)+Nom;
хаман	хаман+Adv;
үзгәрми	үзгәр+V+NEG(мА)+PRES(Й);
.	Type1

Перевод: Его мировоззрение постоянно меняется.

Таблица 3.1: Пример случайного предложения из корпуса Туган Тел

примеров*

В этом файле огромная куча всяких тегов, про которые без бутылки не разберешься (и даже гугление этого парсера, блин, не помогает). Но эмпирическим методом было выяснено, что атрибут PROP — это как раз та самая именованная сущность, что нам нужна. На основании этого все остальные атрибуты были выброшены, а PROP заменен на B-PER, так как используемая модель использовала IOB метод разметки (TODO: написать про IOB метод разметки).

Всего в текстах 30 753 824 слов, из них 534 514 это автоматически размеченные именованные сущности, что составляет 1,738% от всех слов.

TODO: привести примеры слов с атрибутом PROP.

TODO: подсчитать, раз в сколько предложений в среднем встречается именованная сущность

Также в качестве корпуса текста есть татарская википедия. На данный момент она содержит 89 252 статей, причём некоторые из них сгенерированы автоматически, что ухудшает качество текстов как корпуса для обучения, так как некоторые фразы становятся частотными не из-за того, что они действительно часто используются в языке, а из-за множества сгенерированных статей. *TODO вставить пример про бассейны*

Проблема с татарской википедией также была в смеси латиницы и кириллицы *TODO экскурс в историю по поводу того, как мы до такой жизни дошли*, так что пришлось ещё и из латиницы в кириллицу переводить.

3.2 Обучение и тюнинг моделей

3.2.1 BiLSTM-CRF

Была использована модель BiLSTM-CRF, которая норм заработала. Она использует разметку IOB, так что данные пришлось немного подкорректировать и добавить данную разметку. Весь морфологический разбор, кроме разметки именованных сущностей, никак не используется. Из-за того, что у меня нет мощностей для вычислений, приходилось обучаться не на всей выборке, а только на части. Модель показала очень хороший результат.

Category	Precision	Recall	F-score	Predicts	Golds	Correct
PER	99.768	90.727	95.033	2589	2847	2583

TODO: описание модели

3.2.2 BERT

BERT есть для татарского языка, так что осталось его только запустить, что я ещё не сделала, но планирую вот уже на этой неделе. TODO: описать BERT.

4 Воспроизведение статьи Невзоровой

Была воспроизведена статья Невзоровой, на министерствах действительно показала хорошие результаты, но стало очевидно, что это полуручная

история, потому что мусор пришлось выкидывать в ручном режиме. Ну и не удалось воспроизвести запросы в Туган Тел, а поиск был возможен только по слову (фразе). С помощью этого результата хочется разметить википедию, на википедии обучиться, а потом попытаться протестировать на Туган Тел и сравнить результаты.

5 Сравнение результатов

В процессе.

6 Заключение

Проведена большая хорошая работа, получены хорошие результаты, статья Невзоровой должна была называться не так пафосно, но они тоже молодцы, хорошую работу сделали.

Можно будет сотрудничать с Академией наук Республики Татарстан и дальше двигать направление распознавания именованных сущностей, пробовать новые модели не только распознавания, но также и разметки данных, поскольку с каждым годом корпус Туган тел становится объемнее. Использовать в качестве признаков морфологические параметры и не только. Направлений для работы много и это хорошее поле для дальнейших исследований.