TODO: титульник!!!

Оглавление

0.1 Введение

Распознавание именованных сущностей (Named entity recognition, NER) это одна из задач обработки естественного языка; задача детектирования и классификации имен в тексте на несколько заранее определённых категорий, таких как, например, люди, места, организации и т.д. NER имеет множество применений в прикладных задачах, используется в автоматическом разделении на категории текстов, рекомендательных системах, системах извлечения информации. Как идея распознавание именованных сущностей была изобретена ещё в прошлом веке, однако широкое распространение и большую скорость развития получила только в последнем десятилетии. Быстроразвивающиеся глубокие нейронные сети дали значительный толчок развитию обработке естественных языков, и, как следствие, NER. 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 языками существуют только для широко распространённых языков, для которых имеются размеченные корпуса. Во-вторых, большинство из существующих решений требуют большого количества размеченных текстов. В-третьих, исследователи не тратят время на нераспространённые языки, потому что это не самая важная задача, которая, скорее всего, не сможет принести большой выгоды в дальнейшем из-за сравнительно небольшого числа носителей.

Однако развитие распознавание именованных сущностей для татарского языка может быть использовано для всей группы тюркских языков, таких как алтайский, башкирский, чувашский, карачаево-балканский. Это связано с тем, что все языки тюркской группы 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. *написать что-нибудь про то, что язык надо сохранять и вообще все языки важны, особенно язык на 5 миллионов человек*

Сейчас тема распознавания именованных сущностей для языков с малыми ресурсами непопулярна, однако на тему конкретно татарского языка существует работа [ссылка] исследователей из Академии наук Республики Татарстан. В дальнейшем я более подробно рассмотрю их работу в своем исследовании.

Одна из главных проблем NER для языков с малыми ресурсами это отсутствие размеченных корпусов; как следствие задача разметки является одной из подзадач моей работы.

В работе использовалась модель BiLSTM-CRF от [ссылка].

Черновик, вырезать в дальнейшем:

Данная работа представляет собой распознавание именованных сущностей в языке с малыми ресурсами (конкретно, татарским языком). Распознавание именованных сущностей может использоваться во многих прикладных задачах, таких как таргетинг, рекомендация для новостной ленты, приложения для почты. На текущий момент это первая работа по татарскому языку (подробнее об этом в обзоре литературы), поэтому её актуальность не вызывает сомнений: она может быть использована в индустрии для внедрения современных технологий в приложения для носителей языка. Несмотря на то, что татарский язык считается языком с малыми ресурсами, его активно используют более 5 миллионов человек (TODO ссылка на источник), что показывает важность поднятой мной темы.

0.2 Обзор литературы

На данный момент существует одна релевантная моему исследованию статья про работу конкретно в татарском языке, однако она не содержит в себе использование методов современного машинного обучения. Также есть работы на темы других языков с малыми ресурсами и работы о моделях, которые были полезны в моей работе.

0.2.1 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 рассказывают, как разметили корпус «Туган Тел»[ссылка], использовав следующие категории: книги, рестораны, фильмы, журналы, компании, аэропорты, корпорации, языки, колледжи, университеты, школы, магазины, музеи и больницы. Несмотря на наличие в названии распознавания именованных сущностей, они скорее использовали полуручной метод разметки.

TODO сделай везде доллары, где есть математика!!!

TODO Тут будет описание их статьи.

TODO Куда вставить описание Туган Тел?

Использованные данные

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, чем и занималась команда Невзоровой. 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.

TODO Нужен ли этот абзац или нафиг его надо?

В качестве релевантных статей Невзорова at al указывают LingPipe[ссылка], команда которой решает похожую задачу в английском языке (TODO проверить, так ли это, и о чём вообще статья), Annie[ссылка], Afner[ссылка], ссылаются также на марковские цепи, решающие деревья и CRF, которые потом не используют (в то время как я в этой работе использую). В общем, много хороших разных ссылочек, которые надо изучить подробнее, чтобы что-нибудь про них написать. Или вырезать это всё в целом.

Разбор алгоритма, предложенного в статье:

Представленный алгоритм основан на идее сравнения n-грамм. Сравнение происходит на всём объёме корпуса, что увеличивает точность результата, заявляют авторы статьи. Алгоритм является итеративным, причём количество итераций определяется пользователем (что показывает, что их алгоритм является в некоторой степени полуручным.

Первым шагом алгоритма включает в себя выборку 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, в то время как второе слово может добавляться слева или справа, данный параметр выбирается пользователем. Полученный список биграмм 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 биграмм используется как входные данные для второй итерации алгоритма: каждая биграмма ищется по корпусу как фраза и, аналогично первому шагу, составляются триграммы и их частоты. Точно так же выбираются самые частотные триграммы (третье слово может добавляться справа или слева), список обрезается по пороговому значению и, при желании, алгоритм продолжается дальше, используя на вход уже список триграмм.

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

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 n-грамма считается именованной сущностью, иначе алгоритм переходит на следующую итерацию. Таким образом, в финальный результат входят самые стабильные n-граммы разной длины, включая результаты поиска изначального поискового запроса.

Запрос извлечения именованных сущностей представляет собой кортеж (1), где Q_1 и Q_2 никак не объясняются, 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). В качестве примера они снова ссылаются на формулу (1) (скорее всего, имелась в виду формула (2) из примера).

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

Эксперименты:

Тут, конечно, всё хитро: выставляются, естественно, только те результаты, где всё получилось хорошо, а где получилось не слишком хорошо — об этом ничего не сказано. Исследователи перечисляют довольно много категорий, над которыми они экспериментировали, но результаты они показали на словах «министерство», «улица», «язык», «ресторан» и «корпорация». Одной из очевидных дополнительных тем являлись бы «реки», но Невзорова at al. на реках экспериментировать не стали.

Также в данной статье очень интересный способ оценки результатов. Стандартные accuracy, precision и recall (и производная от них F-score) в статье не упоминается, к сожалению, но по тексту можно вычленить нечто на них похожее.

TODO Допиши авторов!!!

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

Что-нибудь тут напишу о том, что статья хорошая, использовала я похожую архитектуру, но не их.

0.2.3 A Neural Layered Model for Nested Named Entity Recognition

Хорошая статья, из которой я решила использовать модель.

0.2.4 Datasets and Baselines for Named Entity Recognition in Armenian Texts

Очень вдохновляющая статья (вообще говоря, магистерская работа), которая, по факту, и стала решающей при выборе темы. 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. У меня, к счастью, такая возможность есть, поэтому я ей воспользовалась.

0.3 Основная часть

Была проделана большая и сложная работа, которая стоила мне года жизни, попыток суицида, кучи нервных клеток и денег на антидепрессанты.

Первым делом я связалась с Невзоровой по указанной электронной почте, чтобы узнать подробности и возможное сотрудничество. Она с большим энтузиазмом восприняла моё письмо и изъявила желание почитать мою работу впоследствии. Господи, надеюсь до этого не дойдёт.

Данная работа была разделена на три части:

1. Получение и разметка данных
2. Обучение и тюнинг моделей
3. Воспроизведение статьи Невзоровой
4. Сравнение результатов

Проблемы, как это часто бывает, возникали на каждом этапе.

0.3.1 Получение и разметка данных

Корпус Туган Тел, который был использован в статье Невзоровой al at, устроен очень хитро. С одной стороны, у них есть сайт, где можно искать по словоформе или лемме с огромным количеством параметров (TODO вставить скриншот), однако возможности просто скачать весь корпус не оказалось. Честно сказать, я до сих пор не очень понимаю, можно ли распространять этот корпус, сейчас пришло в голову, что надо бы это уточнить (TODO: уточнить по поводу распространения корпуса). Таааак вот. Корпус. Доступ к корпусу мне любезно предоставила Невзорова, за что ей огромная благодарность.

Судя по тому, что было в статье NER in Tatar, существует внутреннее IDE для выполнения автоматических запросов, которые сложно воспроизвести из-за отсутствия доступа (ну и доступа к любому коду статьи).

Теперь поговорим подробнее про корпус и про то, как он устроен.

Это .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 и, как водится для автоматической разметки, она далеко неидеальная.

TODO: показать пример разметки

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 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.

0.3.2 Обучение и тюнинг моделей

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

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.

0.4 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.

0.5 Сравнение результатов

В процессе.

0.6 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.