chaii - Hindi and Tamil Question Answering

https://www.kaggle.com/c/chaii-hindi-and-tamil-question-answering/code?competitionId=30060&sortBy=voteCount

Краткое описание задачи:

Это задача категории NLU -- Natural Language Understanding, где от нас требуется предсказать ответ на вопрос к статье на Википедии. Важность задачи состоит в том, что данные представлены на Тамильском языке и Хинди, которые являются непопулярными языками в интернете, что делает сложным поиск ответов на вопросы на этих языках. Соревнование представлено Google Research India.

Допустимо использование предобученных моделей; доступ к интернету недопустим; модель должна обучаться не более 5 часов как на GPU, так и на CPU. Соревнование предоставляет ноутбук слабого бейзлайна.

Участие принципиально осложняется нулевыми знаниями участников соревнования как Хинди, так и Тамильского (однако участники внешне отличают один язык от другого, если это может хоть как-то помочь).

Датасет:

На вход подается:

- `id` - уникальный идентификатор;
- `context` - текст примера на Хинди и Тамильском, где содержится ответ;
- `question` - вопрос на Хинди и Тамильском;
- `answer_text` (train only) - ответ на вопрос; то, что мы пытаемся предсказать;
- `answer_start` (train only) - стартовый символ ответа в `context` (определяется с помощью совпадения подстрок при подготовке данных);
- `language` - язык вопроса: Хинди или Тамильский.

Целевая метка - `answer_text` - строка символов с ответом на соответствующий вопрос.

Все файлы обязаны быть в UTF-8.

Вопросы нетривиальны и неоднозначны, а значит ответы не могут быть получены напрямую поиском по странице. Данные в `answer_start` вычислены приблизительно и с некоторым сдвигом. Участники могут найти собственный индекс начала ответа, могут игнорировать данный столбец или использовать его как есть - предсказать требуется только `answer_text`, поэтому способ использования данного поля зависит только от участников.

Оценка решения:

Метрикой качества в соревновании является *word-level Jaccard score* (подробнее), и итоговая метрика:

$$score = \frac{1}{n}\sum_{i=1}^{n} jaccard(gt_i, dt_i)$$

Где

- *n* - количество документов,
- *jaccard* - функция справа,
- gt_i - i-ое истинное значение,
- dt_i - i-ое предсказание.

```
def jaccard(s1, s2):
    a = set(s1.lower().split())
    b = set(s2.lower().split())
    c = a.intersection(b)
    return float(len(c)) / (len(a) + len(b) - len(c))
```

Предварительное исследование:

Лингвистическое исследование

Любое лингвистическое исследование (на предмет опечаток, неточностей и ошибок) невозможно по причине того, что участники совершенно не знакомы ни с одним из языков датасета, а потому придётся довериться авторам соревнования.

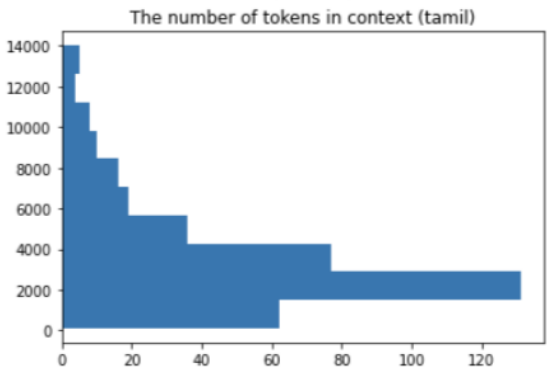
Вместе с тем, был найден один ответ в train_set, оканчивающийся на запятую и один ответ, начинающийся на точку, что вероятнее всего является опечатками.

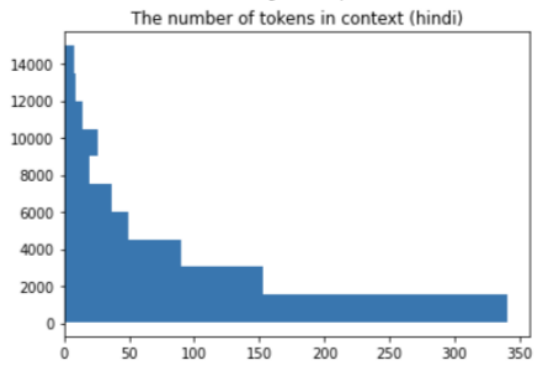
Основные характеристики датасета

1. Train set:
    a. Всего 1114 уникальных строк данных;
    b. Не содрежит пропусков;
    c. Уникальных фрагментов текста (context) 924 строки (из них 623 на Хинди и 301 на Тамильском);
    d. 1104 уникальных вопроса;
    e. 990 уникальных ответов (target, он же answer_text);
    f. Ответ на вопрос находится преимущественно в начале фрагмента текста;
    g. Распределение по языкам 2:1 - 67% пар вопрос-ответ на Хинди, 33% на Тамильском.
2. Test set (открытый):
    a. 5 пар контекст-вопрос, где каждый контекст и каждый вопрос уникальны;
    b. Соотношение языков 3:2 - 60% вопросов на Хинди, 40% вопросов на Тамильском.

Другие наблюдения

a. Во время исследования символов в контекстах, были выявлены не только символы хинди и тамильского языков, а также английского, русского, китайского, греческого и др.
b. Количество вопросов, на которые даются числовые ответы (арабскими цифрами) для языка Хинди - 68, для Тамильского - 65.
c. Строгой зависимости длины ответа в символах от длины вопроса не выявлено, однако частичная корреляция прослеживается.
d. Гистограммы распределения количества токенов в контекстах для обоих языков (вертикальная ось - количество токенов, горизонтальная - контекстов):





Обзор похожих решений и задач:

Несколькими главными критериями по выбору модели являлись:

- Предобученная модель (поскольку не обладаем достаточными данными на хинди и тамильском);
- Мультиязычность модели (из-за обилия сторонних символов);
- Совместимость **и** с Тамильским **и** с Хинди.

mBERT

Был выпущен вместе с BERT, поддерживает 104 языка. Подход очень прост: по сути, это просто BERT, обученный на текстах на многих языках. В частности, он был обучен на материалах Википедии с общим для всех языков словарным запасом. Для борьбы с дисбалансом содержания Википедии, например, английская Википедия имеет в 120 раз больше статей, чем исландская Википедия, малые языки были выбраны в избытке, а большие - в недостатке.

XLM

XLM (Lample and Conneau, 2019) - это модель на основе трансформатора, которая, как и BERT, обучается с целью моделирования языка по маске (MLM). Кроме того, XLM обучается с целью Translation Language Modeling (TLM) в попытке заставить модель изучать схожие представления для разных языков. TLM довольно проста: введите одно и то же предложение на двух разных языках и маскирует лексемы, как обычно. Чтобы предсказать замаскированную лексему, модель может использовать лексемы из другого языка.

XLM обучается как с помощью MLM, так и TLM, причем MLM - на данных из Википедии на 15 языках XNLI (***Cross-lingual Natural Language Inference (XNLI) dataset***), а TLM - на нескольких различных наборах данных в зависимости от языка. Для TLM требуется набор данных параллельных предложений, который может быть трудно получить.

XLM-R (XLM-RoBERTa)

Самой последней многоязычной моделью является XLM-R (Conneau et al., 2019), где R означает RoBERTa (Liu et al., 2019). Исходя из названия, естественно было бы предположить, что это XLM с RoBERTa вместо BERT, но это было бы неправильно.

Вместо этого XLM-R делает шаг назад от XLM, отказываясь от цели TLM, и просто обучает RoBERTa на огромном многоязычном наборе данных огромного масштаба. Неразмеченный текст на 100 языках извлекается из набора данных CommonCrawl, общий объем которого составляет 2,5 ТБ. Он обучается в стиле RoBERTa, то есть только с использованием цели MLM. Фактически, единственным заметным отличием от RoBERTa является размер словаря: 250 тысяч лексем по сравнению с 50 000 лексем RoBERTa. Это делает модель значительно более крупной: 550 миллионов параметров по сравнению с 355 миллионами у RoBERTa.

Если не учитывать разницу в масштабе, то основное различие между XLM и XLM-R заключается в том, что XLM-R полностью самоконтролируется, в то время как XLM требует параллельных примеров, которые трудно получить в достаточном масштабе.

Сравнительный обзор:

(Источники: XNLI (Conneau et al., 2018) и MLQA (Lewis et al., 2018))

Model	**Average**	**English**	**Hindi**
roberta.large.mnli	77.8	91.3	70.9
xlmr.large.v0	82.4	88.7	79.8
mBert	57.7	77.7	43.8

Допуская обобщения, XLM-R сильнее mBERT на 13.8% accuracy на датасете XNLI, на 12.3 F1-score на датасете MLQA и на 2.1% F1-score на датасете NER. **XLM-R выступает исключительно хорошо на низкоресурсных языках** (+11.8% accuracy на Суахили и +9.2% на Урду в сравнении с XLM).

Датасеты моделей

SQuAD

SQuAD (*Stanford Question Answering Dataset)* -- это набор данных по пониманию прочитанного, состоящая из вопросов, заданных толпой пользователей по набору статей Википедии, где ответом на каждый вопрос является сегмент текста, или фрагмент, из соответствующего отрывка для чтения. Или вопрос может быть без ответа.
В первой версии датасета 87.599 + 10.570 записей, в то время как во второй 130.319 + 11.873 записей.

XQuAD

XQuAD (Cross-lingual Question Answering Dataset) - это эталонный набор данных для оценки эффективности межъязыковых ответов на вопросы. Набор данных состоит из подмножества 240 абзацев и 1190 пар "вопрос-ответ" из набора разработки SQuAD v1.1 (Rajpurkar et al., 2016) вместе с их профессиональными переводами на десять языков: испанский, немецкий, греческий, русский, турецкий, арабский, вьетнамский, тайский, китайский и хинди. Таким образом, набор данных является полностью параллельным на 11 языках.
В частности, 98.714 записей на хинди.

MLQA

MLQA (Multilingual Question Answering Dataset) - это эталонный набор данных для оценки эффективности многоязычных ответов на вопросы. Набор данных состоит из 7 языков: арабский, немецкий, испанский, английский, хинди, вьетнамский, китайский.
В частности, 5.425 записей на хинди.

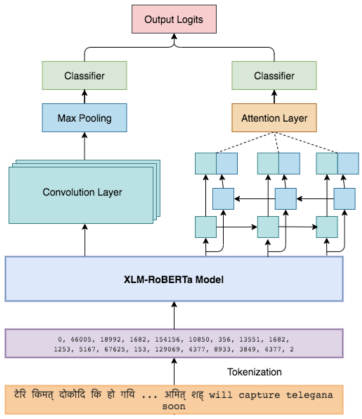
XNLI

XNLI - подмножество из нескольких тысяч примеров из MNLI, которые были переведены на 14 различных языков (некоторые с низким уровнем ресурсов). Как и в случае с MNLI, целью является предсказание текстовой эвентуальности (подразумевает ли предложение A/противоречит/не подразумевает предложение B) и задача классификации (даны два предложения, предскажите одну из трех меток).
7.500 записей.

Разработка модели

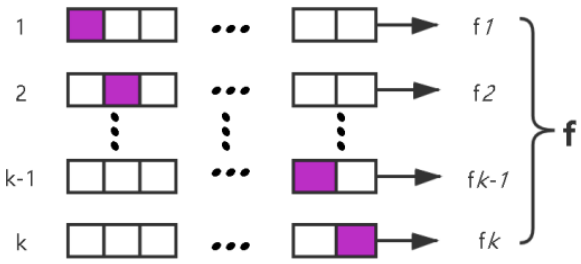
Сначала нами был выбран mBERT, как инструмент для первой попытки работы с мультиязычными моделями, чтобы достичь установленного *baseline*.
Однако после некоторого исследования и анализа, мы решили в дальнейшем перенести работу на XLM-R, как модели, показавшей себя лучше всего на тестах на Хинди, а также на малочисленных языках (к коим относится Тамильский).

А конкретнее, модель xlm-roberta-large-squad-v2, пред обученную на датасете squad-v2 (https://huggingface.co/deepset/xlm-roberta-large-squad2) пред обученный токенизатор.
Она использует механизм внимания, pool layer, архитектура подобна, но в нашем случае предсказывались индексы начала и конца ответа на вопрос. Другие гиперпарметры: `optimizer_type = 'AdamW', learning_rate = 1e-5.`



и

Чтобы улучшить предсказательную способность модели, мы использовали **k-fold ensemble method**, идея похожа на k-fold cross validation. Исходные данные случайным образом делятся на k частей (в нашем случае выбрано k = 5), (k-1)-ое подмножество используется для обучения, оставшееся подмножества используются в качестве набора для проверки, а затем повторяется k раз. Наконец, результаты накапливаются и усредняются для получения окончательного результата. Таким образом, рассчитывались индексы начала и конца ответа на вопрос в контексте и заполнялся результат.

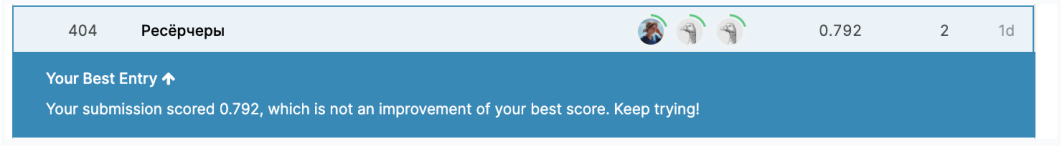


Результаты на keggle

a. Submission

Submission and Description	Status	Public Score	Use for Final Score
notebook9cb63e048b Version 3 (version 3/3) a day ago by Timofey_Sagitov_184 Notebook notebook9cb63e048b | Version 3	Succeeded	0.792	☐
notebook9cb63e048b Version 3 (version 3/3) a day ago by Timofey_Sagitov_184 Notebook notebook9cb63e048b | Version 3	Succeeded	0.792	☐

b. Public Leaderboard



(baseline соревнования достигнут)