Разметка текстовых трансляций и новостей

Заплатин Алексей

27 июня 2022

Abstract

Данная работа посвящена разметки данных текстовых трансляций и новстей к ним. Будут рассмотрены три модели для классификации, которые будут выступать в роле разметчиков данных. Две из рассмотренных моделей будут применены и получены работоспособные модели. Код работы можно найти здесь: `https://github.com/alekz99/summ-sports-broadcasts/tree/data_markup`.

1 Введение

Суммаризация – это автоматическое создание краткого содержания исходного текста. Выделяют два типа суммаризации: экстрактная и абстрактная. Алгоритмы для абстрактного резюмирования позволяют получать результат, который наиболее близок к ручному переводу. Главный недостаток алгоритмов заключается в необходимости разметки данных.

В данной работе будут рассмотрены два набора данных. Первый набор содержит текстовые трансляции спортивных матчей на русском языке. Второй набор содержит новости к этим матчам. Новости могут содержать любую информацию: коэффициенты букмекеров, покупки игроков, исходы матчей и др., поэтому перед использованием абстрактных алгоритмов необходимо разметить данные.

Для всех новостей предполагается реализовать несколько видов классификации:

1. релевантность, отделение новостей, которые не относятся к трансляции;
2. временной промежуток, новости, которые были до начала трансляции;
3. обобщающие новости, которые посвящены всей трансляции целиком;
4. реплика из трансляции, которой соответствует новость.

В данной работе были реализованы две первые классификации.

2 Рассматриваемые модели для разметки данных

Для разметки данных были рассмотрены три модели:

1. одиночный BERT;
2. двойной BERT;
3. сиамский BERT.

Архитектура одиночного BERT представлена на рис. 1. BERT получает на вход один текст, который является результатом объединения двух текстовых источников. BERT ожидает ввода данных в определенном формате. Слова конвертируются в токены, которые есть в словаре. Слова, выходящие за рамки словаря, обрабатываются методом WordPiece. Для обозначения начала и конца последовательностей из токенов используются специальные токены начала и конца последовательности [Cerliani, 2020].

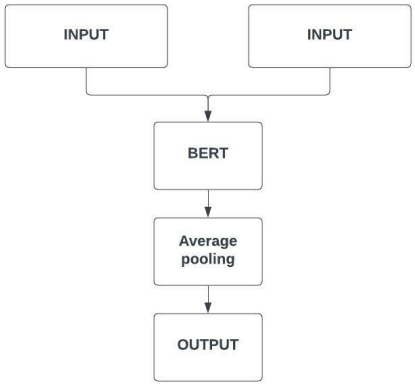

Figure 1: Модель одиночный BERT

Для каждого объекта из набора данных есть три матрицы: токен, маска и ид токенов из словаря. Для одиночного BERT будет только один набор матриц, так как в токенизатор передаются две текстовые последовательности, объединенные с токеном конца последовательности.

В архитектуре двойного BERT, представленного на рис. 2, используются два BERT, каждый из которых обучается на разных входных последовательностях. Первый получает текстовую трансляцию, второй – новость. При кодировании входных данных получается два кортежа матриц, по одному для каждого BERT. Окончательные скрытые состояния для каждого BERT усредняются, затем объединяются в один слой и проходят через Dense слой [Cerliani, 2020].

Архитектура сиамского BERT представлена на рис. 3. В данной модели два разных источника данных передаются в один и тот BERT. Входные матрицы совпадают с двойным BERT. Окончательные скрытые состояния

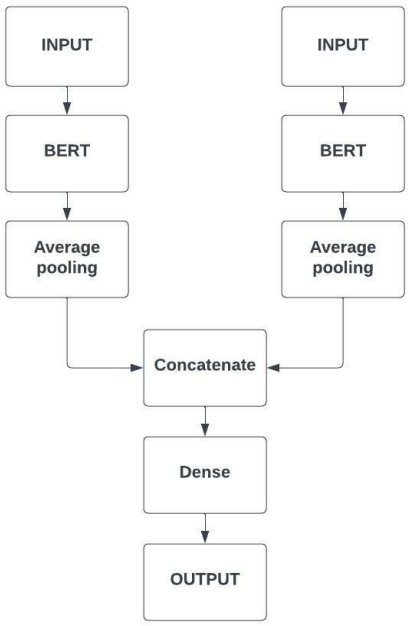

Figure 2: Модель двойной BERT

для обеих входных последовательностей усредняются, затем объединяются в один слой и проходят через Dense слой [Cerliani, 2020].

Опираясь на источник [Cerliani, 2020] были выбраны две первые модели для реализации и анализа полученных результатов.

3 Эксперименты

3.1 Бинарная классификация времени новости

Суть классификации заключается в том, что необходимо выделить новости, которые были до матча. Такие новости будут иметь значение признака 0, если новость была во время матча или после, то 1. Был выбран данный тип классификации, так как для данного типа разметки можно получить автоматическую классификацию средствами Python. Так как разметку можно получить автоматически, без использования модели, данный эксперимент направлен на оценку эффективности моделей.

Было проведено два эксперимента. В первом эксперименте использовалась модель одиночного BERT. На вход подавались объединенные первые 100 токенов трансляции и первые 100 токенов новости. Для обучения использовался весь набор данных, обучение длилось 3 эпохи. Во втором эксперименте

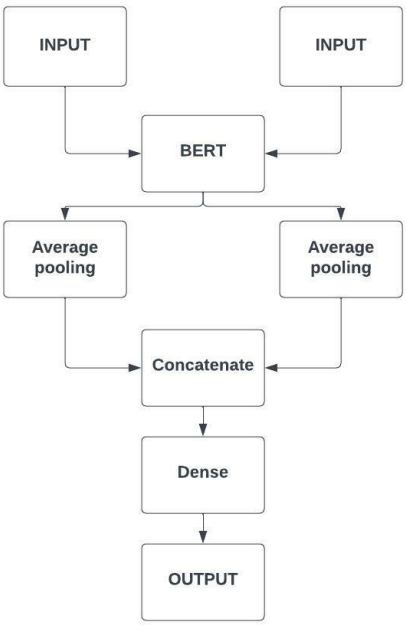

Figure 3: Модель сиамский BERT

использовалась модель двойного BERT. Текстовые трансляции делились на части по 300 токенов, а новости обрезались до 300 токенов. На вход первого BERT подавался кусок трансляции, на вход второго кусок новости. Обучение происходило на небольшой части , 5000 пар, так как обучение на всем наборе данных, 92000 пар, для одной эпохи составляло 10 часов. После предсказания для каждой новости было получено несколько откликов модели, по которым определялось значение целевого признака. Был реализован принцип: если хоть один отклик был равен 1, то это значило, что новость была во время трансляции или после.

3.2 Бинарная классификация релевантных новостей

Суть классификации заключается в определении новостей, которые релевантны трансляции. Например, коэффициенты букмекеров не являются релевантными новостями для трансляции. Целевая переменная равна 0, если новость не релевантна. Каждая трансляция делится на куски по 300 токенов, каждая новость обрезается до 300 токенов. На вход первого BERT подавался кусок трансляции, на вход второго кусок новости. Значение целевой переменной определялось с помощью среднего значения нескольких откликов для каждой новости.

Набор данных был поделен 5 частей, где размер каждого куска представлен

1 набор данных	2 набор данных	3 набор данных	4 набор данных	5 набор данных
929 пар	920 пар	911 пар	902 пары	89348 пар

Table 1: Деление набора данных

в табл. 1. Это было необходимо для активного обучения модели, алгоритм которого представлен на рис. 4. На первом этапе размечается первый набор данных вручную, после на нем обучается модель и сохраняются веса. С помощью обученной модели получаются предсказания для второго набора данных, валидируются и оценивается качество модели. Если достигнуто необходимое качество, то получают предсказания для оставшегося набора данных. Если нет, то первый и второй набор данных становится тренировочным, а третий – тестовым и процесс повторяется, пока не будет достигнуто необходимое качество [NewTechAudit, 2021]. В данном эксперименте после второй итерации было достигнуто необходимое качество.

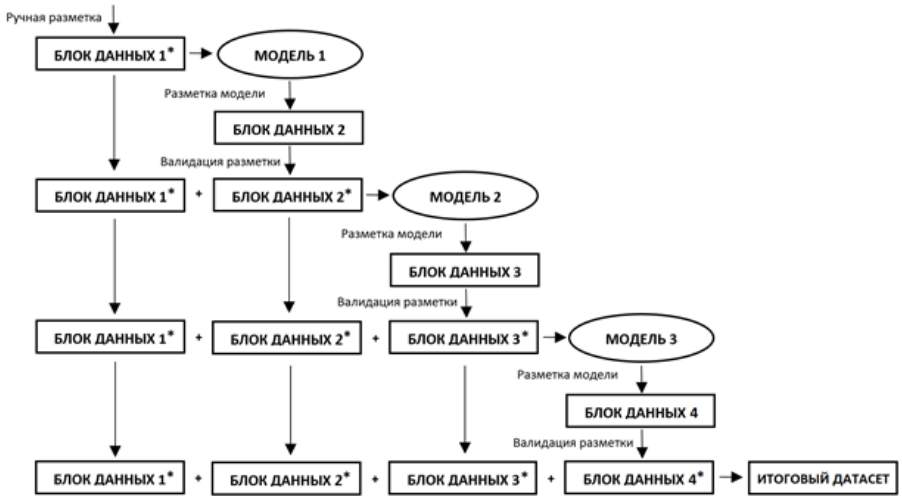

Figure 4: Алгоритм активного обучения

Также для решения проблем с оперативной памятью в Colab при получении предсказаний для всего набора данных было выполнено его деление на две части, где дальше каждая из частей делилась по 10.000 пар, токенизировалась и поступала на вход для получения откликов. В конце все куски были объединены в итоговый набор данных.

4 Результаты

Для временной классификации результаты представлены в табл. 2. Точность модели составляет 89%, где 76% новостей были во время матча, либо после.

	Полученная точность	Всего новостей во время матча или после
Эксперимент 1	89%	76%
Эксперимент 2	74%	71%

Table 2: Качество моделей для первой классификации

	Первая итерация	Вторая итерация
Эксперимент 3	94%	97%

Table 3: Качество модели по итерациям для второй классификации

Во втором эксперименте на выбранном кусе набора данных была получена точность 74%, где 71% был во время матча или позже.

Результаты классификации о релевантности представлены в табл. 3. Так как была достигнута точность 97%, было принято решение закончить активное обучение на двух итерациях. Модель научилась качественно определять нерелевантные новости. Примеры таких новостей приведены в табл. 4.

5 Выводы

В данной работе были рассмотрены 3 модели для классификации с использованием двух входных последовательностей. Были применены 2 модели для решения двух задач классификации. Также было применено активное обучение. Были получены работоспособные модели, которые могут решать данные задачи.

Дальнейшей целью является реализация других типов классификации и повышение качества уже реализованных моделей.

Причина нерелевантности
Информация о нескольких матчах
Новости, связанные с болельщиками
Турнирная сетка/содержание огромного html кода(оставшиеся после предобработки)
Переносы матчей
Турнирная сетка
Покупка/обсуждение игроков вне матча
Ожидания букмекеров
Билеты/популярность трансляции
Краткая сводка без имен собственных
Назначение судей/иногда разбирательство по судейству

Table 4: Примеры нерелевантных новостей

References

[Cerliani, 2020] Cerliani, M. (2020). Siamese and dual bert for multi text classification. https://towardsdatascience.com/siamese-and-dual-bert-for-multi-text-classification-c6552d435533.

[NewTechAudit, 2021] NewTechAudit (2021). Active learning для разметки своими руками. https://habr.com/ru/post/580448/.