Реализация автоматического реферирования содержимого веб-страницы

Подготовил: Дробов Денис, МО-191

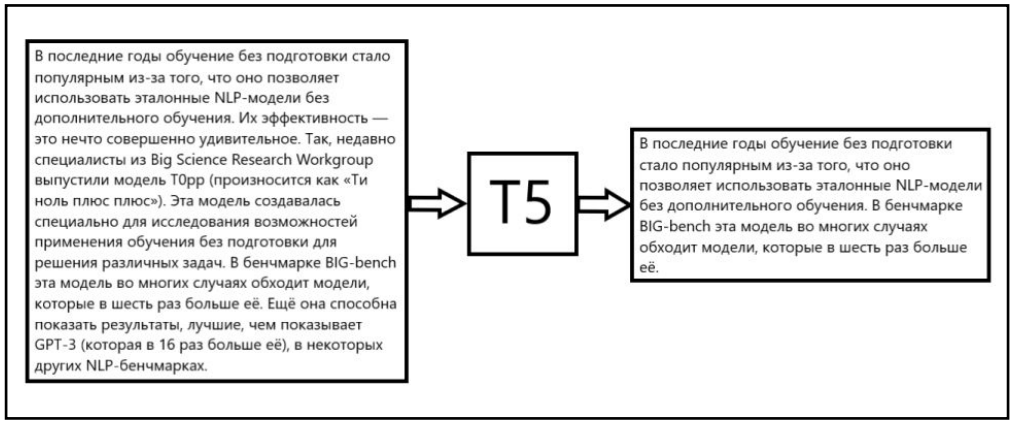
Идея и использование реферирования

Идея реферирования текста – сокращение времени его чтения за счет предоставления более краткой формы реферируемого текста

Использование реферирования текста в квалификационной работе – описание текстового содержимого веб-страницы

Трансформер преобразования из текста в текст

Text-to-text transfer transformer (T5) – нейронная сеть, основанная на архитектуре Transformer. Модель T5 очень эффективна для решения различных задач обработки естественного языка и используется во многих приложениях, включая поиск информации, генерацию текста и многое другое.



Фундаментальные для T5 математические концепции

– идея обучения с учителем;

– метод регуляризации;

– функция потерь перекрестной энтропии;

– механизм внимания

Обучение с учителем

Обучение с учителем в модели T5 предполагает использование пары входной и выходной текстовых последовательностей. Во время обучения модель пытается предсказать выходную последовательность на основе входной последовательности, используя одну из концепций – функцию потерь.

Метод регуляризации (сокращение весов)

Идея сокращения весов заключается в уменьшении параметров с наибольшими значениями, в процессе которого их нормы минимизируются с помощью коэффициента регуляризации. То есть регуляризация добавляет к оптимизационной функции модели штрафную функцию.

$$L_2 = \sum_i (y_i - y(t_i))^2 + \lambda \sum_i a_i^2.$$

Функция потерь перекрестной энтропии

Функция потерь перекрестной энтропии используется для оценки вероятности генерации правильного ответа.

Функция потерь перекрестной энтропии вычисляется путем суммирования отрицательных логарифмов вероятностей правильных ответов.

$$L = -\frac{1}{N}\left[\sum_{j=1}^{N}\left[t_j log(p_j) + (1-t_j) log(1-p_j)\right]\right].$$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}.$$

Механизм внимания

Проблемой при построении эмбеддинга, с которой борется механизм внимания, является возможный многозначный перевод слова. Для этого механизм реконструирует эмбеддинг каждого слова, добавляя к нему векторы слов, схожих по контексту, с определенными весами, в данном случае, эти весы будут являться скалярными произведениями векторов слов в предложении.

$$v'_{nail} = w'_1 v_1 + w'_2 v_2 + w'_3 v_{nail} + \ \ldots + w'_i v_i.$$

Анализ качества работы модели с помощью метрик BLEU и ROUGE

BLEU измеряет сходство между текстом, полученным машиной, и эталонными фразами с использованием n-грамм, которые представляют собой непрерывные последовательности из n слов. BLEU Score вычисляет точность n-грамм в машинно-сгенерированном переводе путем сравнения их с эталонными фразами.

$$BLEU = BP \cdot e^{\sum_{n=1}^{N} w_n log p_n}.$$

ROUGE выполняет обратную функцию – фокусируется на том, сколько n-грамм из эталонных фраз появляется в обработанных машиной данных.

Показатели метрик

BLEU – 0,177

ROUGE-1 – 0,622

ROUGE-2 – 0,381

ROUGE-L – 0,438

Сравнение показателей метрик разных моделей (GPT и BERT)

Особенность GPT – обучение на больших текстовых данных; при обучении каждое слово прогнозируется предыдущим

Особенность BERT – обучение на больших текстовых данных; использование контекста с обеих сторон слова для лучшего отражения значения в конкретном контексте

	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
T5	**0,177**	**0,622**	**0,381**	0,438
GPT	0,169	0,549	0,324	**0,444**
BERT	0,155	0,498	0,263	0,331