Baseline модель.

Суммаризация новостей - достаточно сложная задача для больших моделей в условиях сильной ограниченности вычислительных мощностей. Запустить тот же трансформер у нас, к сожалению, не вышло из-за реально большой длины входных в модель последовательностей. (в среднем 4 - 6 тыс. символов в оригинальных текстах).

Именно поэтому было принято решение реализовать более простую модель, а именно Encoder-Decoder RNN (а точнее LSTM) без механизма Attention. В дальнейшем планируем добавить Attention и, возможно, усложнить модель. Использованные параметры:

- batch_size = 32
- embedding_dim = 128
- hidden_size = 128

Реализована модель при помощи PyTorch, используя официальную документацию, open-source код и код из наших домашних заданий на ПМИ. Предварительно весь датасет токенизирован при помощи BPE (vocab_size = 100k). Все предварительные тесты были обучены на 20к объектах. Однако при обучении были выявлены некорректные объекты в выборке, пришлось дополнительно повторно обработать датасеты.

Обзор литературы. Метрики.

METEOR.

Основная цель метрики — улучшить корреляцию с человеческими оценками качества суммаризации или перевода по сравнению с существующими метриками, такими как BLEU и NIST. Основные особенности метрики:

1. METEOR основана на сопоставлении юниграмм (отдельных слов) между машинным текстом и эталонными текстами.
2. Учитываются не только точные совпадения, но и морфологические варианты и синонимы.
3. Метрика вычисляет оценку на основе точности (precision), полноты (recall) и меры фрагментации, которая учитывает порядок слов.
4. Оценка METEOR включает штраф за фрагментацию, который уменьшает итоговую оценку, если совпадающие слова расположены в неправильном порядке.

Формулы:

$$METEOR = Fmean * (1 - Penalty)$$

Где $Fmean$ – гармоническое среднее, вычисляющееся по следующей формуле:

$$Fmean = \frac{10 * P * R}{R + 9 * P}$$

Где P – precision, R – recall:

$$P = \frac{\text{кол} - \text{во совпадающих юниграмм}}{\text{число юниграмм в машинном тексте}}$$

$$R = \frac{\text{кол} - \text{во совпадающих юниграмм}}{\text{число юниграмм в эталонном тексте}}$$

Penalty – это штраф за разброс слов в машинном тексте. Формула:

$$Penalty = 0.5 * (\frac{chunks}{\text{кол} - \text{во совпадающих юниграмм}})^3$$

Где chunks – это число юниграмм машинного текста, которые находятся в какой-либо группе юниграмм из эталонного текста.

ROUGE.

Метрика ROUGE сравнивает автоматически сгенерированные суммаризации с эталонными (человеческими) суммаризациями, используя различные методы подсчета совпадений, такие как n-граммы, последовательности слов и пары слов.

Для оценки нашего бейзлайна мы использовали метрику Rouge-L. Она использует длину наибольшей общей подпоследовательности (LCS) между машинной и эталонной суммаризациями. Этот метод учитывает порядок слов и позволяет оценить, насколько хорошо машинная суммаризация сохраняет структуру эталонной.

Формула:

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

$$ROUGE - L = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

Где *LCS(X, Y)* – длина наибольшей общей подпоследовательности между машинной суммаризацией X и эталонной Y, n – длина машинной суммаризации, m – длина эталонной суммаризации, β – параметр, контролирующий вес R относительно P.

BLEU.

Метрика BLEU позволяет быстро и недорого оценивать тексты, сравнивая их с эталонными человеческими текстами. Основная идея заключается в том, чтобы измерять близость машинного перевода к профессиональному человеческому переводу с помощью числового показателя.

Формула:

$$BLEU = BP * exp(\sum_{n=1}^{N} w_n \log p_n)$$

Где BP(Brevity Penalty) – штраф за краткость машинного текста. Формула:

$$BP = \begin{cases} 1, & c > r \\ e^{1-r/c}, & c \leq r \end{cases}$$

Где c – длина машинного текста, r – длина эталонного текста.

w_n – это вес для n-грамм, обычно он одинаков для всех n-грамм и равен 1/N, N – максимальная длина n-грамм.

p_n – это модифицированный precision для n-грамм. Формула:

$$p_n = \frac{\sum Count_{clip}(ngram)}{\sum Count\ (ngram)}$$

Где сумма производится по всем n-граммам в машинном переводе,

$Count_{clip}$(n-gram) – это кол-во совпадений n-грамм между машинным и эталонным текстами, ограниченное максимальным количеством таких n-грамм в эталонных текстах.

Count(n-gram) - общее количество n-грамм в машинном тексте.

Perplexity.

Perplexity — это метрика, используемая для оценки качества языковых моделей. Она измеряет, насколько хорошо языковая модель предсказывает последовательность слов. Чем ниже её значение, тем лучше модель предсказывает текст.

Формула:

$$Perplexity(W) = \exp\left(-\frac{1}{N}\log P(W)\right)$$

Где N – число слов в последовательности W, P(W) – вероятность для последовательности слов W. Формула:

$$P(W) = \prod_{i=1}^{N} P(w_i|w_1, \dots, w_{i-1})$$

Ссылки

METEOR - https://aclanthology.org/W05-0909.pdf

BLEU - https://arxiv.org/abs/1601.00248

Rouge - https://aclanthology.org/W04-1013.pdf