BLEU4

BLEU4 - ключевая метрика этого соревнования, считается она следующим образом:

$$\text{BLEU4} = \text{BP} \cdot \text{sinergy}$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$\text{sinergy} = \exp\left(\sum_{n=1}^{4} w_n \log p_n\right)$$

То, что я назвал синергией - среднее геметрическое точности пересечений н-грамм.

Также есть сглаживание нулевых попаданий перед геометрическим средним, что сильно не меняет сути.

После релизации beam-search я был довольно разочарован в значении этой метрики на данных, и теперь хочу это исправить с помощью модификаций выбора наилучшего бима.

Моя идея - выбрать с помощью вероятностей токенов и длинн в последовательности оптимальный с точки зрения матожидания бим, а также выбросить из него ненужные токены.

$$\text{total n-grams} = \sum_{n=0}^{4} \binom{len}{n}$$

новый токен добавляет:

$$\sum_{n=1}^{4} \binom{l}{n} - \sum_{n=1}^{4} \binom{l-1}{n} = \sum_{n=1}^{4} \binom{l}{n} - \binom{l-1}{n} = \sum_{n=1}^{4} \frac{l!}{n! \cdot (l-n)!} - \frac{(l-1)!}{n! \cdot (l-n-1)!} =$$
$$= \sum_{n=1}^{4} l \frac{(l-1)!}{n! \cdot (l-n-1)! \cdot (l-n)} - \frac{(l-1)!}{n! \cdot (l-n-1)!} = \sum_{n=1}^{4} \frac{l \cdot (l-1)! - (l-1)! \cdot (l-n)}{n! \cdot (l-n-1)! \cdot (l-n)} =$$
$$= \sum_{n=1}^{4} \frac{(l-1)! \cdot n}{n! \cdot (l-n)!} = \sum_{n=1}^{4} \frac{n}{n!} \cdot \frac{(l-1)!}{(l-n)!} = \frac{1}{1} \cdot \frac{(l-1)!}{(l-1)!} + \frac{2}{2} \cdot \frac{(l-1)!}{(l-2)!} + \frac{3}{6} \cdot \frac{(l-1)!}{(l-3)!} + \frac{4}{24} \cdot \frac{(l-1)!}{(l-4)!} =$$
$$1 + l - 1 + \frac{1}{2}(l-1)(l-2) + \frac{1}{6}(l-1) \cdot (l-2) \cdot (l-3) = \frac{6l + (l-1) \cdot (l-2) \cdot (3+l-3)}{6} = \frac{6l + l \cdot (l-1) \cdot (l-2)}{6} =$$
$$= \frac{l \cdot (6 + (l-1) \cdot (l-2))}{6}$$

Новых н-грам, которые уйдут в в штраф через sinergy.

Для упрощения будет максимизировать $\ln(\text{bleu4}) = \text{penalty} + \frac{1}{w_n} \cdot \sum \ln p_n$

penalty будем штрафовать отдельно по длине бимов, остается оценить вклад каждого токена в логарифм presision.

$$\sum \ln p_n = \sum_{n=1}^{4} \ln \text{correct}_n - \sum_{n=1}^{4} \ln \text{total}_n$$

Нам нужно сделать предположение, в какую сумму токен даст больший вклад, если мы добавим его в перевод, вторая сумма, кстати, тоже является функцией от длины.

У нас есть оценка на логарифм вероятности того, что токен впишется в существующий контекст, так как при тренировке мы максимизируем кросс-энтропию. Эту ценку мы можем получить из логитов модели.

Вообще говоря, кажется, что вероятности пересечения для каждой длины можно хорошо оценить двунаправленной Masked Language моделью и в бимсерче редактировать бим скоры на ходу при помощи оценки вклада логарифмов вероятностей в BLEU4, но в этой ДЗ к сожалению запрещены ансамбли.

Остается только поставить порог на то, начиная с какого значения логарифма вероятности мы будем брать токен в последовательность. (параметр border в beam-search).

Вот к какому решению я пришел в коде:

```
if border > 0.0:
    trg_seq = trg_seq.masked_fill(trg_seq_scores < np.log(border), pad_idx)
```