Алгоритмы восстановления пробелов в тексте

2. Вероятностные подходы

Вероятностные подходы основаны на оценке встречаемости каждого слова корпуса. В данном случае выбирается наиболее вероятная подпоследовательность слов, то есть:

$$s^* = \text{argmax}_{s \in S} P(s), \tag{1}$$

где S – множество всех возможных подпоследовательностей слов.

2.1. Вероятностная языковая модель

Наилучшая последовательность слов определяется формулой (1). Тогда при n словах в последовательности s вероятность разбиения:

$$P(s) = \prod_{k=1}^{n} P(w_k), \tag{2}$$

где $P(w_k)$ – вероятность встретить в тексте слово w_k.

Рассмотрим строку «the men dine here». Для оценки вероятностей воспользуемся открытым словарем: http://norvig.com/ngrams/. Данный словарь содержит более триллиона слов. В используемом корпусе приведенные слова встретились:

the – 23135851162 раза;
men – 174058407 раз;
dine – 2012279 раз;
here – 639711198 раз.
Всего слов в корпусе: 1024908267229. Тогда

$$\begin{aligned} P(\text{«the men dine here»}) &= P(\text{«the»})\, P(\text{«men»})\, P(\text{«dine»})\, P(\text{«here»}) = \\ &= 4.698 \times 10^{-15}. \end{aligned} \tag{3}$$

Поскольку $P(w) \in [0; 1]$, то при перемножении вероятностей может возникнуть проблема арифметического переполнения снизу. Чтобы этого избежать часто прибегают к логарифмированию, тогда (2) примет вид:

$$\log P(s) = \log \prod_{k=1}^{n} P(w_k) = \sum_{k=1}^{n} \log P(w_k). \tag{4}$$

Другой возможной проблемой является отсутствие встретившегося слова в словаре. Оценка вероятности встретить данное слово равно нулю. Чтобы не возникало такой ситуации, применяют аддитивное сглаживание (сглаживание Лапласа), которое состоит в искусственном добавлении единицы к встречаемости каждого слова

$$P(w_i) = \frac{\nu_i + 1}{\sum_{i' \in V} (\nu_{i'} + 1)} = \frac{\nu_i + 1}{|V| + \sum_{i' \in V} \nu_{i'}}, \tag{5}$$

где ν_i – сколько раз встретилось слово w_i, V – словарь.

2.2. Биграммы

Предыдущий метод рассматривает каждое слово в отдельности, однако не учитывает соседние слова, которые также могут помочь правильно разбить текст. Чтобы учесть сочетания слов рассматриваются биграммы – сочетания по два слова и на основе вероятностей встречаемости сочетаний слов восстановить пробелы.

Строка «the men dine here» с учетом начала <B> и конца <E> предложения может быть разбита на 5 биграмм:

1. <B> the
2. the men
3. men dine
4. dine here
5. here <E>

Формула (2) тогда примет вид:

$$P(s) = \prod_{k=1}^{n} P(w_k \mid w_{k-1}). \tag{6}$$