$P(w|h)$ denote the probability of a word w given some history h

suppose the history h is 'its water is so transparent that'

The probability that the next word is 'the'

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{\text{Count}(\text{its water is so transparent that})}{\text{Count}(\text{its water is so transparent})}$$

How to compute probability of entire sentences $P(w_1, w_2, \cdots, w_n)$

$$P(X_1 X_2 X_3 \cdots X_n) = P(X_1) P(X_2 | X_1) P(X_3 | X_{1:2}) \cdots P(X_n | X_{1:n-1})$$

$$= \prod_{k=1}^{n} P(X_k | X_{1:k-1})$$

hence

$$P(w_{1:n}) = P(w_1) P(w_2 | w_1) P(w_3 | w_{1:2}) \cdots P(w_n | w_{1:n-1})$$

$$= \prod_{k=1}^{n} P(w_k | w_{1:k-1})$$

The bigram model only consider probability of preceding word $P(w_n | w_{n-1})$

$$P(w_n | w_{1:n-1}) \approx P(w_n | w_{n-1})$$

For n-gram, we have $P(w_n | w_{1:n-1}) \approx P(w_n | w_{n-N+1:n-1})$

by bigram assumption, we have $P(w_{1:n}) = \prod_{k=1}^{n} P(w_k | w_{k-1})$

$$P(w_n | w_{n-1}) = \frac{C(w_{n-1} w_n)}{\sum_w C(w_{n-1} w)} = \frac{C(w_{n-1} w_n)}{C(w_{n-1})}$$

all bigram that share the same first word

For general case

$$P(w_n | w_{n-N+1:n-1}) = \frac{C(w_{n-N+1} \cdots w_{n-1} w_n)}{C(w_{n-N+1:n-1})}$$

Perplexity

the perplexity of a test set is the inverse probability of a test set, normalized by the number of given word

$$PP(W) = P(w_1 w_2 \cdots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \cdots w_N)}} \qquad W = w_1 \cdots w_n$$

$$= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

we want to minimize PP since minimize PP $\Leftrightarrow$ maximize test set probability

3.3 Sampling sentences from a language model

Choose random value between 0 and 1, find that point on probability line

Print the word whose interval include that value until we generate </s>

3.4 Generalization and zeros

unknown words are denoted as <UNK>

3.5 smoothing

3.5.1 Laplace smoothing

Add one to all n-gram counts

Given word w_i and its count c_i

$P(w_i) = \frac{c_i}{N}$ N is total number of word tokens

$P_{Laplace}(w_i) = \frac{c_i+1}{N+V}$

Instead of adjusting both numerator and denominator, we define adjusted count c^*

$c_i^* = (c_i+1)\frac{N}{N+V}$ $\quad P_i^* = \frac{c^*}{N}$

$$P_{Laplace}(w_n | w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{\sum_w (C(w_{n-1}w)+1)} = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$$

Add k smoothing

$$P^*_{Add\text{-}k}(w_n | w_{n-1}) = \frac{C(w_{n-1}w_n)+k}{C(w_{n-1})+kV}$$

3.8 Perplexity's relation to Entropy

Entropy is a measure of information. Given random variable X, entropy is defined as

$H(X) = -\sum_{x \in X} P(x)\log_2 P(x)$ Hence the entropy of a random variable over finite sequences is defined as

$$H(w_1 \cdots w_n) = -\sum_{w_{1:n} \in L} P(w_{1:n}) \log P(w_{1:n})$$

Hence the entropy rate is defined as

$$\frac{1}{n} H(w_{1:n}) = -\frac{1}{n} \sum_{w_{1:n} \in L} P(w_{1:n}) \log P(w_{1:n})$$

but to measure the true entropy of a language, we need to consider the sequence of infinite length

$$H(L) = \lim_{n \to \infty} \frac{1}{n} H(w_1, w_2 \cdots w_n)$$

$$= -\lim_{n \to \infty} \frac{1}{n} \sum_{w_{1:n} \in L} P(w_{1:n}) \log P(w_{1:n})$$