How to compute the probabilities?

Language Models are defined by the following objective :

1. Obective - compute the probability of a sentence or sequence of words. $P(W) = P(w1,w2,w3,...,wn)$
2. Related task - computing the probability of the upcoming word. $P(w4|w1,w2,w3)$

Probaility of the entire sentence : $P(W) = P(w1,w2,w3...wn)$ = Probability when all the words are happening together in sequence.

P(The, water, of, Walden, Pond, is, so, beautifully, blue) : The probability of when the sentence "The water of Walden Pond is so beautifully blue" will occur in a corpus.

There is a difference between the following -

1. **Conditional Probability $P(B|A)$** : Probability of B given A : Event B has happened in past. A is happening now. This is known as Conditional Probability.
2. **Joint Probability $P(A,B)$** : Joint probability of A and B. Or probability when both the events A and B are happening simultaneously.

$P(B|A) = P(A \bigcap B) / P(A)$ or, $P(B|A) = P(A,B) / P(A)$ or, $P(A,B) = P(B/A) \times P(A)$

> or, $P(A,B) = P(A) \times P(B|A)$

> ie. probability of two events A and B happening together (joint probability) is probability of A multiplied by probability of B when A has already happened.

3. **Extending it to multiple events we can write**

> $P(A, B, C, D) = P(A) \times P(B|A) \times P(C|A,B) \times P(D|A,B,C)$

To get the intuition, following is the chain of thought -

- first the event A happened. So, the probability is P(A) as nothing else has happened now.
- second the event B happened. Event A has already happened in the last step. so the probability of B, we need to compute $P(B|A)$, i.e., probability of B when A has already happened.
- now, the third event C happened. A and B has already happened. So, probability of C would be, $P(C|A,B)$, because A and B has already happened. So,
- now, the fourth event D happened. A, B and C has already happened by now. So, porbability of D when A, B and C has already happened is $P(D|A,B,C)$

4. P(blue|the water of walden pond is so beautifully) = C(the water of walden pond is so beautifully blue)/C(the water of walden pond is so beautifully)

5. Generalizing the above

$$ P(w_{1:n}) = P(w1) \times P(w2|w1) \times P(w3|w_{1:2}). . . P(w_n|w{1..n-1}) \ = \prod_{k=1}^n(P(w_k|w_{1:k-1})) $$

Eg : $P("The \ water \ of \ walden \ pond \ is \ so \ beautifully \ blue") = P(The) \times P(water|The) \times P(of|The \ water) \times P(walden|The \ water \ of) \times P(pond|The \ water \ of \ walden) \times P(is|The \

water \ of \ walden \ pond) \times P(so|The \ water \ of \ walden \ pond \ is) \times P(beautifully|The \ water \ of \ walden \ pond \ is \ so) \times P(blue|The \ water \ of \ walden \ pond \ is \ so \ beautifully)$

1. We will never see enough data for estimating all these probabilities. Hence **Markov assumption** is used to simplify the matter. It states the following -

Markov Assumption

P(blue | The water of Walden Pond is so beautifully) ≈ P(blue | beautifully) $P(w_n|w_{1:n-1}) ≈ P(w_n|w_{n-1})$ The approximation is known as *bi-gram assumption or 1st order markov assumption*. It is considering the probability of *$(n-1)^{th}$ word*, instead of probability of *n-1 words*, to determine the Probability of n^{th} word.

k = 2 : bi-gram model : probability of the current word depends on the previous 2-1 words. k = n : n-gram model : probability of the current word depends on the previous n-1 words. k = k : k-gram model : probability of the current word depends on the previous k-1 words.

$ P(w_{1:n}) ≈ ∏_{k=1}^n(P(w_k|w_{k-1})) $

Going more generally, **n-gram** would consider *n-1* words prior to the n^{th} word to determine the probability of the n^{th} word. bi-gram was considering 2-1 words prior.

Therefore bi-gram model, the probability of the n^{th} word would be determined by the $n-k^{th}$ word.

Therefore for an n-gram model,

$$ P(w_{1:n}) ≈ ∏_{k=1}^n(P(w_k|w_{k-(n-1)})) $$