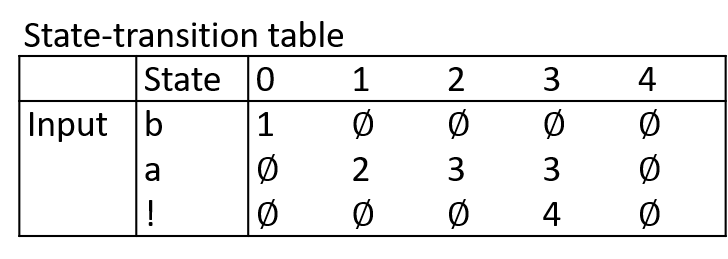
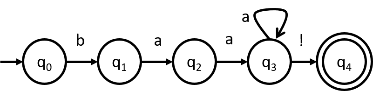
**Text Pre-processing**

* Regular expression

|  |  |  |  |
| --- | --- | --- | --- |
| \* | 0 or more | \{nts} | New line/tab/whitespace |
| + | 1 or more | [a-zA-Z0-9] | Range of characters |
| ? | 0 or 1 | ()[] | Group/any |
| {n, m} | N to m | [^ab] | Not a nor b (^ must be in [] and in front) |
| • | Any character | ^$ | Start/end |
| \{\*.?} | Escaped character | cat|dot | “cat” or “dog” (disjunction) |

* + Hierarchy: () -> counters (e.g. \*) -> start/end -> disjunction
* ****Finite state automata (FSA)

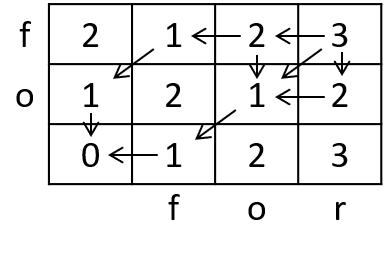
(e.g.) RE: /baa+!/



* Byte-Pair Encoding (BPE)

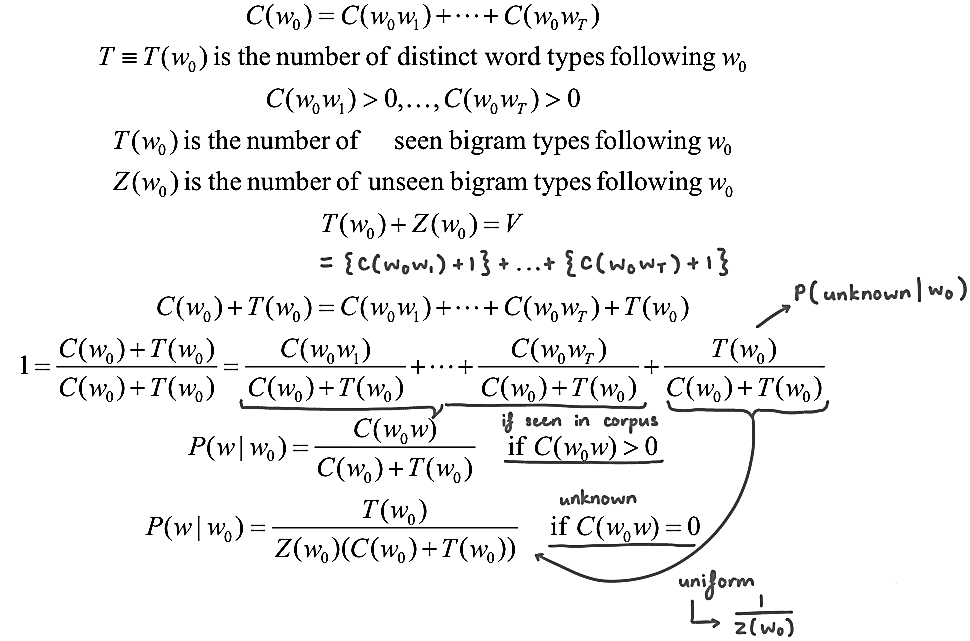
1. Vocab = individual characters + ‘\_\_’ end-of-word symbol
2. Find 2 symbols that are most frequently side-by-side (‘X’ ‘Y’) and add to vocab ‘XY’
3. Replace adjacent X Y with symbol ‘XY’

**Spelling Errors**

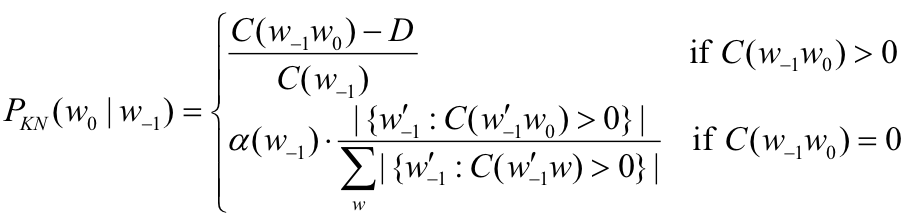
* Bayesian Classification
  + aka Maximum Likelihood Estimate (MLE)
  + Smoothing:
  + Smoothed Count:
  + Discount:
  + Laplace’s Law λ=1, Jeffreys-Perks law λ=0.5
* Minimum Edit Distance
  + Right to left (insert), up to down (delete), across (replace)

**N-Grams**

* + Assume , only dependent on k before
* Witten-Bell Smoothing



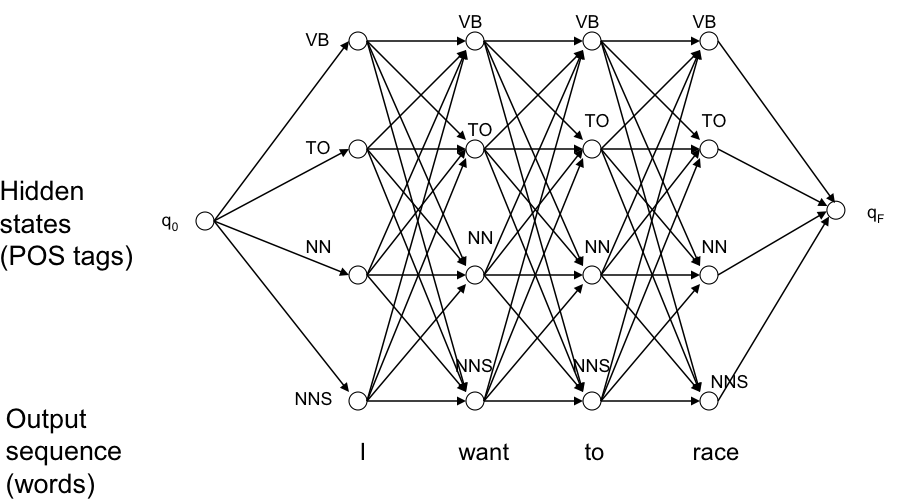
* Interpolation
* Backoff for Bigram
* Kneser-Ney Smoothing for Bigram



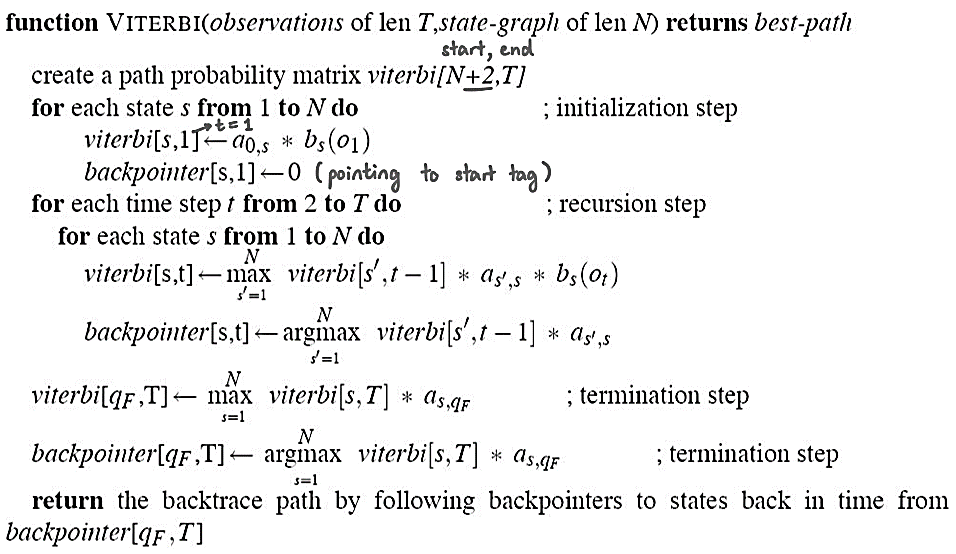
* Entropy of a sequence:
* Entropy of a language:
* Cross entropy:
  + , difference = accuracy of model
* Perplexity on test data, lower the better

**POS Tagging**

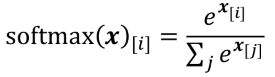
* Hidden Markov Model (HMM)



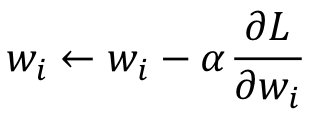
* Viterbi Algorithm
  + : max prob of all paths ending in state qj at time t

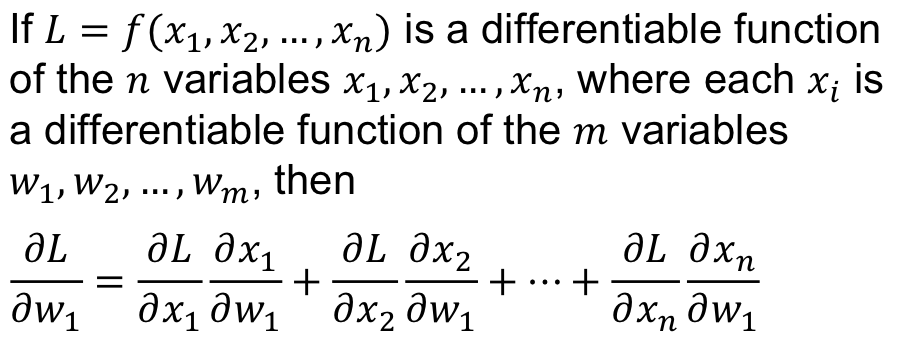


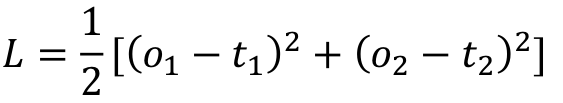
**Linear Models**

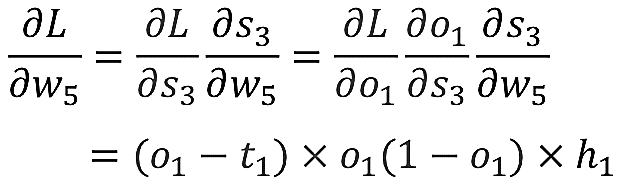
 

**Gradient Descent**

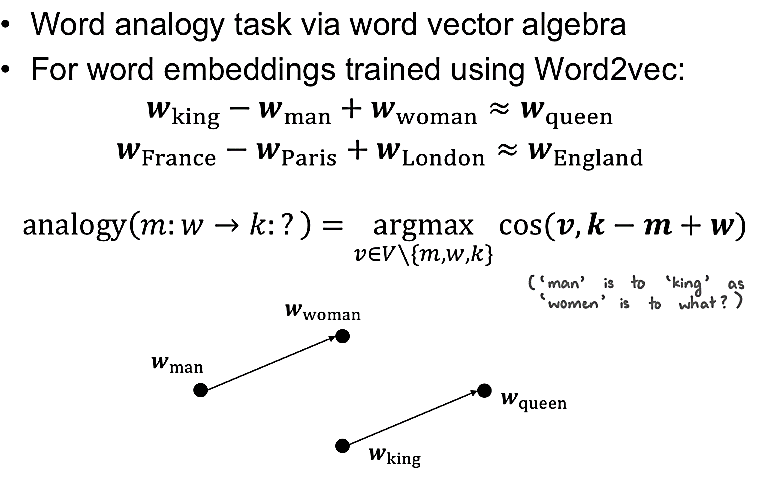




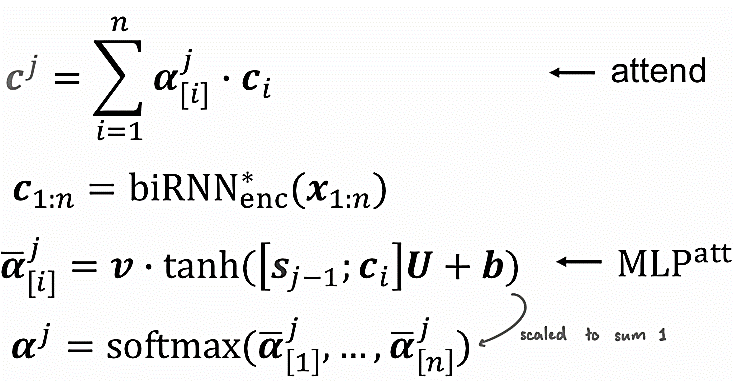


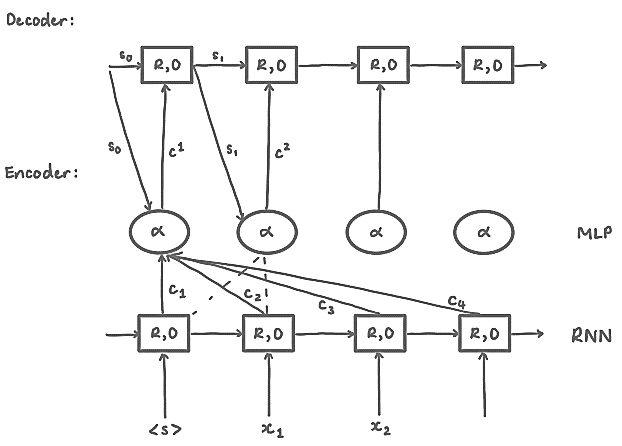


**Word Embeddings**

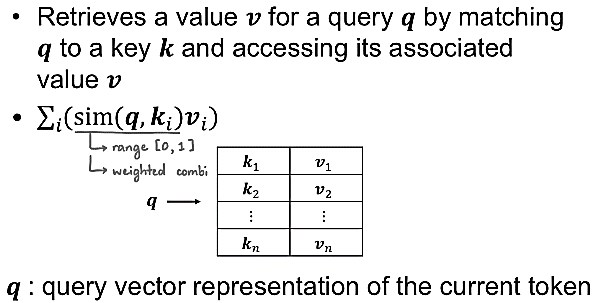


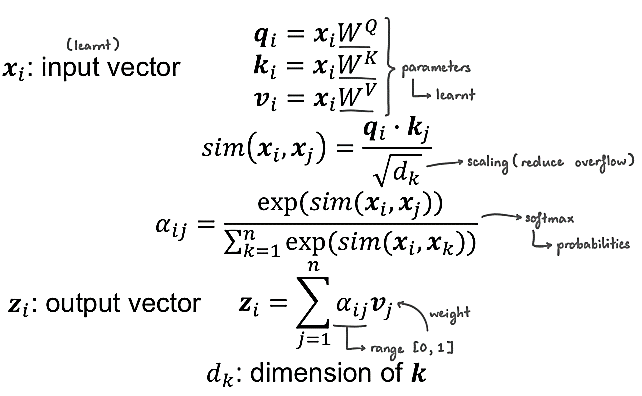
**Encoder-Decoder**

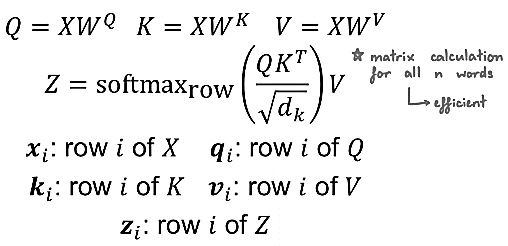




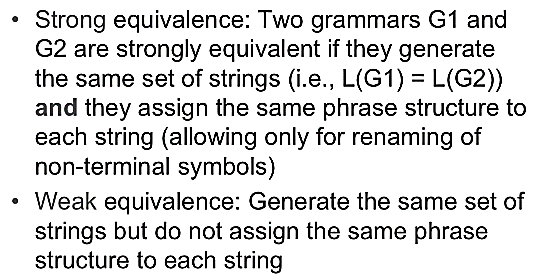
**Transformers & Self-Attention Net**

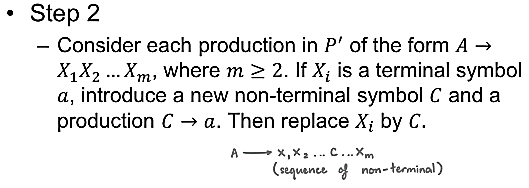
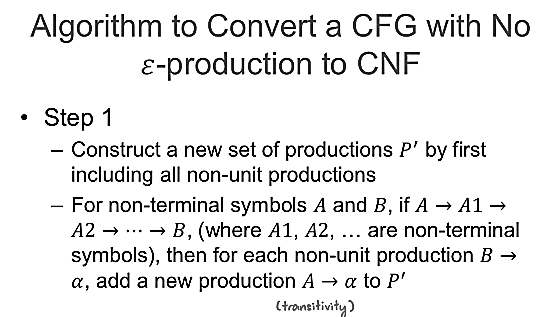


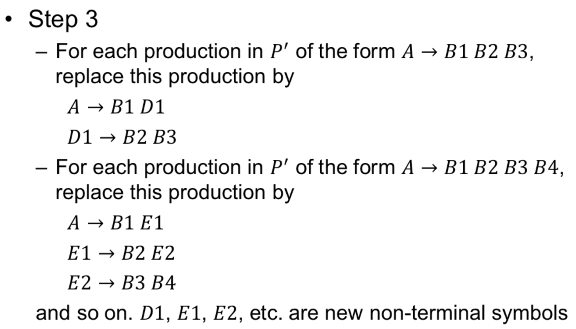


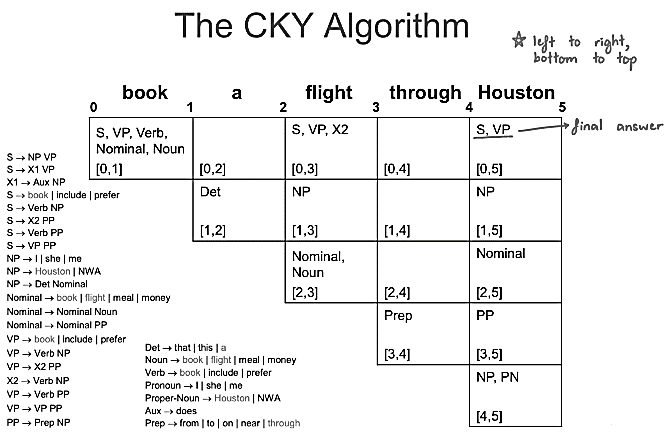


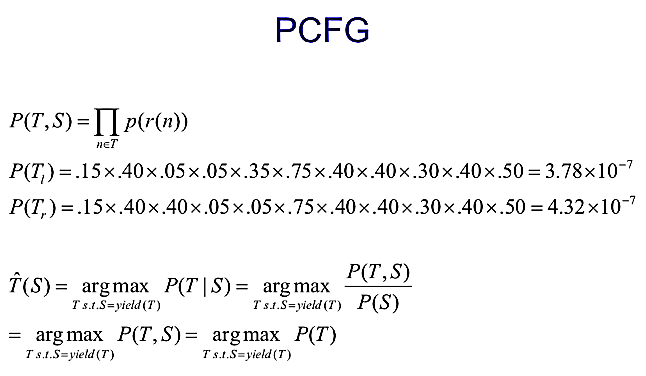
**CFG**

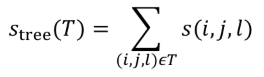


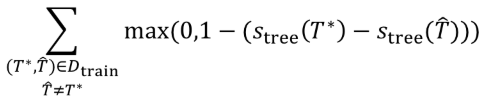












**FOL**

