LANGUAGE MODEL

The goal of a language model is to assign a probability to a sequence of words by means of a probability distribution. Formal grammars (e.g. regular, context free) give a hard
–binary‖ model of the legal sentences in a language. NLP is a probabilistic model of a language that gives a probability that a string is a member of a language or not. To specify a correct probability distribution, the probability of all sentences in a language must sum to 1.

Uses of Language Models

- Speech Recognition
- OCR & Handwriting Recognition
- Machine Translation
- Generation
- Context sensitive spelling correction.

A language model also supports predicting the completion of a sentence. Predictivetext input systems can guess what is been typed and provide choices on how to complete it.

N- Gram Word Models

- This model is considered over sequences of words, characters, syllables or other units.
- Estimate probability of each word given prior context.
- An N-gram model uses only N-1 words of prior context.
    - ✓ Unigram: P(phone)
    - ✓ Bigram: P(phone | cell)
    - ✓ Trigram: P(phone | your cell)
- The Markov assumption is the presumption that the future behavior of a dynamical system only depends on its recent history. In particular, in a Kth-Order Markov Model, next state only depends on the k most recent states, therefore an N – gram model is a (N-1) – order Markov model.

N-gram Character Models

- One of the simplest language models: $P(c_1^N)$

- Language identification: given the text determine which language it is written in.
- Build a trigram character model of each candidate language: $P(c_i \mid c_{i-2i-1}, l)$
- Train and Test Corpora
  - ✓ A language model must be trained on a large corpus of text to estimate goodparameter values.
  - ✓ Model can be evaluated based on its ability to predict a high probability for adisjoint test corpus.
  - ✓ The training corpus should be representative of the actual application data.
  - ✓ To handle words in the test corpus that did not occur in the training data anexplicit symbol is used.
  - ✓ Symbol to represent unknown words (<UNK>)
  - ✓ **Perplexity** – Measure of how well a model ‒fits‖ the test data.

$$Perplexity(W_1^N) = \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

  - ✓ Smoothing - reassigns probability mass to unseen events.