Statistical Language Model

Technology Review
CS–410 Text Information Systems
Fall-2021

Pradosh Kumar Subudhi

University Of Illinois Urbana-Champaign
USA
pradosh2@illinois.edu

Statistical Language Model:
Language model is a probabilistic mechanism to determine the probability of a given sequence of words in a sentence. In other words, Language Model or LM in short is the development of probabilistic models that can predict the next word in a sequence given the words that precede it.

Language modeling has various technical application including but not limited to speech recognition machine translation, document classification and routing, optical character recognition, information retrieval, handwriting recognition, spelling correction etc. Speech recognition problem gave birth to Language modeling.

Common statistical language models are Unigram, Bigram, Trigram (n-gram models). In this document.

n-gram LMs are bases on Markov Assumption. The assumption that the probability of a word depends only on the previous word/s. With this understanding, n-gram looks back at n-1 words in the past (bigram – 1 word in the past and trigram 2 words in the past and so on).

Unigram LM: Unigram language model is a very simple model which evaluates each term independently. It doesn’t consider context or history. This LM is mostly used in Information Retrieval where a complex language model in not needed. The unigram is the foundation of a more specific model variant called the query likelihood model, which uses information retrieval to examine a pool of documents and match the most relevant one to a specific query.

Probability of sentence containing terms t_1, t_2, t_3 –

$$P(t_1t_2t_3) = P(t_1) * P(t_2) * P(t_3)$$

Bigram LM: Bigram LM uses approximates the probability of a word given all the previous words by using only the conditional probability of the preceding word.

Probability of sentence containing terms t_1, t_2, t_3 –

$$P(t_1t_2t_3) = P(t_1) * P(t_2|t_1) * P(t_3| t_2)$$

Trigram LM: Trigram LM uses approximates the probability of a word given all the previous words by using only the conditional probability of the 2 preceding words Trigram uses longer contiguous history.

Probability of sentence containing terms t_1, t_2, t_3, t_4 –

$$P(t_1t_2t_3t_4) = P(t_1) * P(t_2|t_1) * P(t_3|t_1t_2)$$

n-gram models are generally used for spelling correction, word breaking and text summarization.

Bidirectional LM: Unlike n-gram models, which analyze text in one direction (backwards), bidirectional models analyze text in both directions, backwards and forwards. These models can predict any word in a sentence or body of text by using every other word in the text. Examining text bidirectionally increases result accuracy. This type is often utilized in machine learning and speech generation applications. For example, Google uses a bidirectional model to process search queries.

Exponential LM: This type of statistical model evaluates text by using an equation which is a combination of n-grams and feature functions. Here the features and parameters of the desired results are already specified. The model is based on the principle of entropy, which states that probability distribution with the most entropy is the best choice. Exponential models have fewer statistical assumptions which mean the chances of having accurate results are more.

Continuous Space LM: In this type of statistical model, words are arranged as a non-linear combination of weights in a neural network. The process of assigning weight to a word is known as word embedding. This type of model proves helpful in scenarios where the data set of words continues to become large and include unique words.

NLP is an exciting and at the cutting edge of ML where practitioners strive to reduce the errors and improve the abilities of NLP. Language models are the base on which this technology rests, the better the language model the better the model trains and the more accurate the result.

References:

https://towardsdatascience.com/learning-nlp-language-models-with-real-data-cdff04c51c25
https://medium.com/unpackai/language-models-in-ai-70a318f43041
Stanford NLP
Jon Dehdari