Binod Pandey
bpand3@illinois.edu
CS 410 Text Information Systems
Technology Review

Statistical Language Model

Introduction

A language model is basically a probability distribution over words or word sequences. In practice, a language model gives the probability of a certain word sequence being valid. Validity in this context does not refer to grammatical validity.[1] Instead it is trying to validate whether these words sequences resemble the normal speaking and writing within the context where these words appear. Statistical language model is a prevalent probabilistic model for such natural language processing.[2] This model uses natural language such as characters, words, phrase, and paragraphs as inputs and provides the probability of other information that is relevant in those contexts. Statistical language models develop probabilistic models that help with predictions for the next word in the sequence. These models help businesses to solve different problems such as speech recognition, machine translation, optical character recognition, and information retrieval, to name a few. Depending upon the requirement different kinds of statistical language models are used to solve this problem. We will discuss a few of these statistical models below:

N-Gram:

You can think of an N-gram as the sequence of N words, by that notion, a two-gram (or bigram) is a two-word sequence of words like “please turn”, “turn your”, or” your homework”, and a three-gram (or trigram) is a three-word sequence of words like “please turn your”, or “turn your homework”.[4] The intuition of the n-gram model is that instead of computing the probability of a word given its entire history, we can approximate the history by just the last few words.[5]

Bidirectional

Unlike n-gram models, which analyze text in one direction, bidirectional models analyze text in both directions, backwards and forwards.[6] This model 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

Exponential statistical language model is also known as maximum entropy models.[6] This model is more complex than n-grams. The model evaluates text using an equation that combines feature functions and n-grams. It specifies features and parameters of the desired results. Unlike n-grams, it leaves analysis parameters more ambiguous and doesn't specify individual gram sizes. The model is based on the principle of entropy, which states that the probability distribution with the highest entropy is the best choice. In other words, the model with the most chaos, and least room for assumptions, is the most accurate. Exponential models are designed to maximize cross entropy, which minimizes the amount statistical assumptions that can be made. This enables users to better trust the results they get from these models.

Conclusion

Statistical language model is diverse, and it is combined with other models to make it more powerful for context-based prediction. The research in this field continues and another language model called neural network-based models helps ease the sparsity problems. These language models are the base for next generation human like intelligence systems.[1]

Citations:

1. Kapronczay, M. (2022, January 8). *A beginner's guide to language models*. https://towardsdatascience.com/. Retrieved November 6, 2022, from https://towardsdatascience.com/the-beginners-guide-to-language-models-aa47165b57f9Taylor,
2. K. (2022). *Here are the Top NLP Language Models That you need to know!* https://www.hitechnectar.com/. Retrieved November 6, 2022, from https://www.hitechnectar.com/blogs/here-are-the-top-nlp-language-models-that-you-need-to-know/
3. Dehdari, J. (2015, March 1). *A Short Overview of Statistical Language Models*. https://jon.dehdari.org. Retrieved November 6, 2022, from https://jon.dehdari.org/tutorials/lm_overview.pdf
4. Kapadia, S. (2019, March 26). *Language Models: N-Gram*. https://towardsdatascience.com. Retrieved November 6, 2022, from https://towardsdatascience.com/introduction-to-language-models-n-gram-e323081503d9
5. Jurafsky, D., & Martin, J. (2021, December 21). *N-gram Language Models*. https://web.stanford.edu. Retrieved November 6, 2022, from https://web.stanford.edu/~jurafsky/slp3/3.pdf
6. Lutkevich, B. (2020, March 1). *Language Modeling*. https://www.techtarget.com. Retrieved November 6, 2022, from https://www.techtarget.com/searchenterpriseai/definition/language-modeling