Parth Shukla

CS 410

6 November 2022

Statistic Language Models: Google Suggest

Among many different aspects of text data retrieval and analysis, there are a few different areas that emerging and existing technologies utilize without our knowledge. Of these, statistic language models are one of the most important. What statistical language models entails is the prediction of subsequent words in a sequence given the what words come before it. We see this in many facets of our daily lives such as with text prediction while texting on a phone. The biggest example of this is Google Suggest which gives suggestions of searches when a query is entered. The technology that Google uses in its products is very complex, but at the basis of its suggestion feature is the incorporation of statistical language models in order to properly autocomplete queries.

Before understanding how Google Suggest works, it is important to discuss the foundations of the model that runs behind it. In a research paper from Carnegie Mellon University titled *Two Decades Of Statistical Language Modeling: Where Do We Go From Here?,* author Ronald Rosenfeld explains that statistical language modeling is essentially “estimating the probability distribution of various linguistic units, such as words, sentences, and whole documents” (1). What this means is that most basically, a statistical language model incorporates the probability distribution of all possible sentences. This can then be utilized in order to create suggestion and autocomplete features that we see in Google Suggest. So, with a statistical language model, we are able to predict what word is most likely to occur next in a sequence, given what words come before it. One such model that implements this would be the n-gram model. This model incorporates a sequence of N words (thus N-gram) and calculates the probability that a word came after it. For example, it will find the P(w|h) such that the word w follows the history h. So, with a statistical language model, we are able to calculate, find, and assign the probability of a word occurrence depending upon the document collection that is given. This would thus give us a way to calculate the probability that a word occurs next, and make proper suggestoins on that.

Google Suggest is one of the many applications of statistical language models that we utilize in our daily lives. Google Suggest utilizes statistical language models in order to correctly suggest the next word in a query to help make the searching process much easier and quicker. There are a few factors that affect what Google Suggest is recommending such as search frequency, behavior, and location; each contributing to what is the most likely next word in a sequence as we saw in a basic n-gram model. With this standard model, there are relevant autocomplete features that are implemented based on the probability of what the next word would be. In collaboration with Google Suggest, there is also Google Instant. This is a feature that dynamically updates the Google search page based on autocompletion. So, if a user was typing “Google trans-”, the autocomplete would be “translate” and search results for Google Translate would appear as they are typing. This incorporates the statistical language model in order to effectively find and dynamically update the correct autocompletion and recommendation. Jon Dehdari in a lecture explains an example of how Google Suggest would work. With a language model, given a search query of “how to cook french…” the model would calculate the effective probabilities of p(how to cook french fries) versus p(how to cook french dictionary). This would then find what is the most likely autocompletion and suggest the correct one. This technology waas very innovative at the time, and has now become standard practice for many of the product we use on a daily basis such as Spotify, YouTube, and TikTok.

Overall standard language models are at the basis for many technologies we use. In their essence, they calculate hte probability of a word occurring next in a sequence given the words that precede it. This is very helpful for recommendation and autocomplete systems that we utilize. Among those, Google Suggest is very prominent and incorporates this technology to help its users. It is influenced by many different factors such as search history and frequency, and allows users to properly receive the recommendation for their query. Overall, within its complexity, Google Suggest incporrates the fundamental properties of statistical language models to properly autocomplete search queries.

Works Cited

Brownlee, Jason. “Gentle Introduction to Statistical Language Modeling and Neural Language Models.” *Machine Learning Mastery*, 31 Oct. 2017, machinelearningmastery.com/statistical-language-modeling-and-neural-language-models/. Accessed 1 Nov. 2022.

Dehdari, Jon. *A Short Overview of Statistical Language Models Invited Talk at the Workshop on Data Mining and Its Use and Usability for Linguistic Analysis*. 2015.

Kapadia, Shashank. “Language Models: N-Gram - towards Data Science.” *Medium*, Towards Data Science, 26 Mar. 2019, towardsdatascience.com/introduction-to-language-models-n-gram-e323081503d9. Accessed 1 Nov. 2022.

“What Is Google Suggest? Search Using Autocomplete.” *Ryte.com*, 2012, en.ryte.com/wiki/Google_Suggest#:~:text=Google%20Suggest%20or%20autocomplete%20is,as%20it%20is%20being%20entered. Accessed 1 Nov. 2022.