Omar Boffil
Oboffil2

Statistical language model Review

Professionals in the economic and environmental sciences must solve problems from the collection and analysis of data. In general, these data are taken from a sample from surveys or experiments; that is, the information they work with is partial. Therefore, they must have tools to help them make the best decision when faced with questions that have uncertain answers. Statistics provide the tools required to collect the data. It allows us to summarize and present the information in the sample, to later infer from its fundamental characteristics of the population from which it was extracted. Besides, it makes it possible to quantify the uncertainty associated with our responses or, in other words, the probability of making a mistake in the decision made.

Statistical language model documents reviewed describes that assigns a probability to a sequence of X words P(X1, X2, …. Xn) using a probability distribution. Having a way to estimate the likelihood of different sentences is useful in many natural language processing applications. Language modeling is used in speech recognition, machine translation, speech tagging, analysis, handwriting recognition, information retrieval, and other applications.

This document also describes that N-GRAMs are statistical models that predict the next word in the sentence using the previous n-1 words. These types of statistical models that use sequences of words are also called language models. For example, we have a sentence "I cannot read without reading _____", we can say that the next most likely word would be "glasses". N-GRAMS predicts the following word in the sequence using the conditional probability of the next word. The N-GRAM model is essential in speech and language processing.

The conditional probability of the next most likely word can be obtained using a large corpus managed Collection of Text or Speech Data. It's all about counting things (words) from the corpus. The goal is to find P(w | h), which is the probability of the next word in the sequence given a certain story h.

The N-GRAM model concept is that instead of calculating the probability of a word given its entire history, it shortens the narrative to a few previous words. When we use a single last word to predict the next word, it is called a Bi-GRAM model. For example, we have P (glasses | reading), the probability of the word "glasses" given the previous word "reading".

P (glasses | reading) = Count (reading glasses) / Count (reading)

N-GRAM models are very important when we have to identify words in a loud and ambiguous input. N-GRAM models are used in:

- Speech recognition
- Handwriting recognition
- Spell correction
- Translator machine
- many other applications

The model document shows how the n-grams can also be used to make approximate fits efficiently. By converting a sequence of elements into a set of n-grams, it can be entered into a vector space; in other words, represented as a histogram, thus allowing the sequence to be compared with other sequences efficiently. For example, suppose we convert text strings with only letters of the Spanish alphabet into 3-grams. In that case, we will get a vector space of 27 ^3 dimensions (the first dimension measures the number of occurrences of "aaa", the second of "aab", and so on for all possible 3-letter combinations). Using this representation, we lose information about the text string. For example, the strings "abcba" and "bcbab" will lead to exactly the same diagrams. However, it is empirically known that if two real text strings have a similar vector representation measured through the dot product, they are very likely similar. The document said the other metrics can also be applied to n-gram vectors mixed and sometimes produce better results. For example, the normal distribution can be used to compare documents, examining how many standard deviations of each n-gram differ from the mean over a large set of documents.

In conclusion, the model report reviewed focuses the study on the MLs, which are built based on bigrams or trigrams from a training corpus that is widely used in many fields: speech recognition, handwriting recognition, spell checking, machine translation, etc. They are modeled based on n-grams since they are easy to construct and are formulated as a probability distribution P(x), which reflects the frequency of occurrence of the chain x in a training corpus.