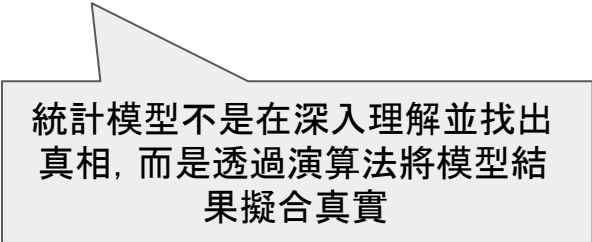


On Norvig's Blog

Chomsky's thought on statistical learning

It's true there's been a lot of work on trying to apply statistical models to various linguistic problems. I think there have been some successes, but a lot of failures. There is a notion of success ... which I think is novel in the history of science. It interprets success as approximating unanalyzed data.



統計模型不是在深入理解並找出真相，而是透過演算法將模型結果擬合真實

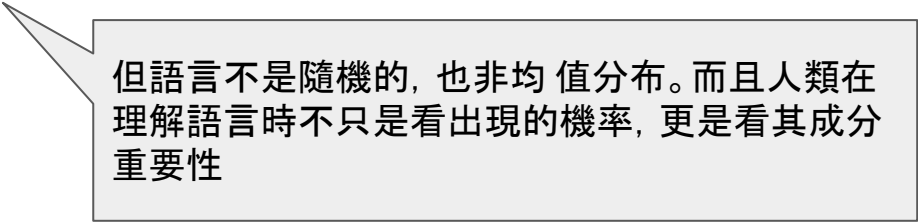
Norvig:

收集語料確實重要，但重點是怎麼使用那些語料

- B. Science is a combination of gathering facts and making theories; neither can progress on its own. In the history of science, the laborious accumulation of facts is the dominant mode, not a novelty. The science of understanding language is no different than other sciences in this respect.
- C. I agree that it can be difficult to make sense of a model containing billions of parameters. Certainly a human can't understand such a model by inspecting the values of each parameter individually. But one can gain insight by examining the *properties* of the model—where it succeeds and fails, how well it learns as a function of data, etc.

語言學的目的正是要了解語言的運作邏輯(例如有什麼principles and parameters)但統計模型的方法忽略了語言的運作核心，只是試圖模仿語言的表現。這使得統計模型的結果就算是對的，也不足以證明統計模型可以完整描述語言。就像是用錯公式卻剛好算出正確答案罷了。

everything as simple as possible, but no simpler. Many phenomena in science are stochastic, and the simplest model of them is a probabilistic model; I believe language is such a phenomenon and therefore that probabilistic models are our best tool for representing facts about language, for algorithmically processing language, and for understanding how humans process language.



但語言不是隨機的，也非均值分布。而且人類在理解語言時不只是看出現的機率，更是看其成分重要性

e.g., X-bar 就是一種數學模型

- **數學模型 (mathematical model)**

定義了變數之間的關係，這種關係可以是從輸入到輸出的函數形式（例如： $y = mx + b$ ），也可以是集合形式（例如：以下這些 (x, y) 數值對屬於此關係）。

- **機率模型 (probabilistic model)**

指定了隨機變數可能取值的機率分佈（例如： $P(x, y)$ ），而非嚴格決定性的關係（例如： $y = f(x)$ ）。

- **經過訓練的模型 (trained model)**

使用某種訓練或學習演算法，從可能的模型集合以及大量資料點（例如：許多 (x, y) 的配對）之中選擇出最佳模型。這通常透過統計推論的過程來完成，例如從資料中估計出模型參數（如上例中的 m 和 b ）。



A relevant probabilistic statistical model is the [ideal gas law](#), which describes the pressure P of a gas in terms of the the number of molecules, N , the temperature T , and Boltzmann's constant, K :

$$P = N k T / V.$$

The equation can be derived from first principles using the tools of statistical mechanics. It is an uncertain, incorrect model; the *true* model would have to describe the motions of individual gas molecules. This model ignores that complexity and *summarizes* our uncertainty about the location of individual molecules. Thus, even though it is statistical and probabilistic, even though it does not completely model reality, it does provide both good predictions and insight—insight that is not available from trying to understand the *true* movements of individual molecules.

兩者的研究對象本質不一樣。例如 空氣粒子在空間中的分佈是均值的，但語言使用不是

Clearly, it is inaccurate to say that statistical models (and probabilistic models) have achieved *limited* success; rather they have achieved an *overwhelmingly dominant* (although not exclusive) position.

這和「統計模型科學上是不是正確的」是兩回事。
只能說它在工程(商業)上成功

I said that statistical models are sometimes confused with probabilistic models; let's first consider the extent to which Chomsky's objections are actually about probabilistic models. In 1969 he famously [wrote](#):

But it must be recognized that the notion of "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

語言「表現」是無限的, $1/\infty = 0$

His main argument being that, under any interpretation known to him, the probability of a novel sentence must be zero, and since novel sentences are in fact generated all the time, there is a contradiction. The resolution of this contradiction is of course that it is not necessary to assign a probability of zero to a novel sentence; in fact, with current probabilistic models it is standard practice to do smoothing and assign a non-zero probability to novel occurrences. So this criticism is invalid, but

建模型的工程步驟
沒有回答語言本質的問題

Chomsky說(a)或(b)或句子中的任何一部份是1955年以前的英文使用者從未使用過的句子, (a)是符合文法的句子, 而 (b)不是。

In *Syntactic Structure*, Chomsky introduces a now-famous example that is another criticism of finite state probabilistic models:

Neither (a) 'colorless green ideas sleep furiously' nor (b) 'furiously sleep ideas green colorless', nor any of their parts, has ever occurred in the past linguistic experience of an English speaker. But (a) is grammatical, while (b) is not.



Chomsky appears to be correct that neither sentence appeared in the published literature before 1955. I'm not sure what he meant by "any of their parts" had occurred, for example:

Norvig 抓住 "any of their parts" 攻擊並表示有出現過。
→ 未能證明句子 (a) 與 (b) 確實未曾出現

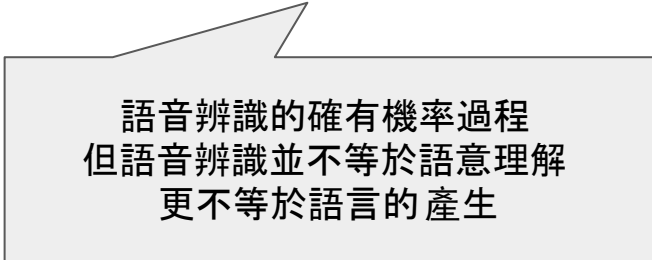
- "It is neutral green, **colorless green**, like the glaucous water lying in a cellar." [The Paris we remember](#), Elisabeth Finley Thomas (1942).
- "To specify those **green ideas** is hardly necessary, but you may observe Mr. [D. H.] Lawrence in the role of the satiated aesthete." [The New Republic: Volume 29](#) p. 184, William White (1922).
- "**Ideas sleep** in books." [Current Opinion: Volume 52](#), (1912).

Norvig用1800-1954的Google Book Corpus訓練模型，得出句子 (a)出現的機率為句子 (b)的10,000倍。

→經過實驗後仍未能證明句子 (a)與(b)確實未曾出現

But regardless of what is meant by "part," a statistically-trained finite state model *can* in fact distinguish between these two sentences. Pereira (2001) [showed](#) that such a model augmented with word categories and trained by expectation maximization on newspaper text, computes that (a) is 200,000 times more probable than (b). To prove that this was not the result of Chomsky's sentence itself sneaking into newspaper text, I repeated the experiment, using a much cruder model with Laplacian smoothing and no categories, [trained over the Google Book corpus from 1800 to 1954, and found that \(a\) is about 10,000 times more probable.](#) If we had a probabilistic model over trees as well as word sequences, we could perhaps do an even better job of computing degree of grammaticality.

From the beginning, Chomsky has focused on the *generative* side of language. From this side, it is reasonable to tell a non-probabilistic story: I *know* definitively the idea I want to express—I'm starting from a single semantic form—thus all I have to do is choose the words to say it; why can't that be a deterministic, categorical process? If Chomsky had focused on the other side, *interpretation*, as Claude Shannon did, he may have changed his tune. In interpretation (such as speech recognition) the listener receives a noisy, ambiguous signal and needs to decide which of many possible intended messages is most likely. Thus, it is obvious that this is inherently a probabilistic problem, as was recognized early on



語音辨識的確有機率過程
但語音辨識並不等於語意理解
更不等於語言的產生