Colorless green ideas learn furiously

Chomsky and the two cultures of statistical learning

Language recognition programs use massive databases of words, and statistical correlations between those words, to translate or to recognise speech. But correlation is not causation. Do these statistical data-dredgings give any insight into how language works? Or are they a mere big-number trick, useful but adding nothing to understanding? One who holds the latter view is the theorist of language Noam Chomsky. **Peter Norvig** disagrees.

Machines can now translate between French and German, or English and Chinese. Google's machine translation system handles over 4000 language pairs, and translates more text each day than all the world's professional translators do in a year. But at the Brains, Minds, and Machines symposium (see `http://mit150.mit.edu/symposia/brains-minds-machines`) held in 2011 when Noam Chomsky was asked about such systems he replied "*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*."

What did Chomsky mean, and is he right?

I take Chomsky's main points to be the following. Accurately modelling linguistic facts is just butterfly collecting; what matters in science (and specifically in linguistics) is the underlying principles. And statistical models do not approach those principles. They are incomprehensible; they provide no insight. They may provide an accurate simulation of language, but the simulation is done completely the wrong way. People do not decide what the third word of a sentence should be by consulting a probability table keyed on the previous two words, instead they use words ordered by grammar to express a thought that is in their minds. This is done without any probability or statistics. So why are these statistical modellers wasting their time on the wrong enterprise? Language must be innate, not based on statistics; a statistical model is irrelevant to any understanding of language.

Is Chomsky right? It is a long-standing debate. These are my answers.

First I believe "*some successes, but a lot of failures*" is a mischaracterization. Statistical language systems are used successfully by hundreds of millions of people every day, and have come to completely dominate the field of computational linguistics. For the first three decades, the field used techniques that were more in line with Chomsky's recommended approach, but starting in the mid-1980s, some researchers started to experiment with statistical models. Their success convinced over 90% of the researchers in the field to switch to statistical approaches.

Second, this is not "*novel in the history of science*," it is typical of science. Science has always been a combination of gathering facts and making theories; neither can progress on its own. If the accumulation of language facts is butterfly-collecting – well, entomologists need to collect butterflies before they can understand them. Open any issue of *Nature* or *Science* and you'll find the majority of articles are mostly about documenting data, not about a new theory. Wikipedia currently says that science "organizes knowledge in the form of testable explanations and predictions." Chomsky seems to want to remove predictions (which are the same thing as "*approximating unanalyzed data*") from the definition. But science needs both predictions and explanations.

> Language is a probabilistic phenomenon; so probabilistic models are best for understanding how humans process language

Third, there is explanatory value in a statistical model. Certainly it can be difficult to make sense of a model containing billions of parameters. A human cannot 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 the amount of data, and so on. Fourth, it is important that these new models are not only statistical, they are also probabilistic: they make probabilistic predictions, not categorical (true/false) ones. Many phenomena in science have random elements – radioactive decay is an obvious example – 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.

What is a statistical model?

We ought at this point to distinguish between a statistical model and a probabilistic one. A statistical model is a mathematical model which is modified or trained by the input of data points. Given a set of data points $\{(x_1,y_1),(x_2,y_2)\ldots\}$ a statistical model is a function $y=F(x)$ that predicts the value y for as-yet unanalyzed values of x. A probabilistic model is one where the function $G(x)$ computes a probability distribution, not just a single value. A model can be statistical or probabilistic or both or neither.

For example, a decade before Chomsky, Claude Shannon proposed probabilistic models of communication (see again `http://mit150.mit.edu/symposia/brains-minds-machines`) based on Markov chains of words[1]. If you have a vocabulary of 100 000 words and a second-order Markov model in which the probability of a word depends on the previous two words, then you need a quadrillion (10^{15}) probability values to specify the model. The only feasible way to learn these 10^{15} values is to gather statistics from data and introduce some smoothing method for the many cases where there are no data.

As another example consider the ideal gas laws, which describes the pressure P of a gas in terms of the number of molecules, N, the temperature T, and Boltzmann's constant, K:

$$P = N\,k\,T\,/\,V$$

This is a probabilistic model. It ignores the complexity of interactions between individual molecules, and *summarizes* our uncertainty about the molecules. Even though it is probabilistic, even though it does not completely model reality, it provides good predictions, and it also provides insight—insight that is not available from trying to understand the *true* movements of individual molecules. Other sciences that cover complex behavior, such as cognitive psychology, biology and particularly genetics, also rely heavily on statistical modeling and probabilistic explanation.

Now consider the non-statistical, non-probabilistic model of spelling expressed by the rule "I before E except after C". We'll call this a *logical* or *categorical* rule because it expresses a definitive yes/no conclusion, not a probability distribution. In contrast, a statistical, probabilistic model is given in the table below:

P(IE) = 0.0177, P(CIE) = 0.0014, P(*IE) = 0.163

P(EI) = 0.0046, P(CEI) = 0.0005, P(*EI) = 0.0041

This model comes from statistics on a corpus of a trillion words of English text[2]. P(IE) is the probability that a word sampled from this corpus contains the consecutive letters "IE". P(CIE) is the probability that a word contains the consecutive letters "CIE", and P(*IE) is the probability of any letter other than C followed by IE. The statistical data confirms that I before E is in fact more common than E before I, and that the dominance of IE lessens when following a C, but, contrary to the rule, CIE is still more common than CEI. (Examples of "CIE" words include "science", "society", "ancient" and "species". "Seize" is an example of E before I in the absence of C.) The disadvantage of the "I before E except after C" logical rule is that it is not very accurate. It predicts the correct spelling only 75% of the time (in fact the simpler rule "I before E" is more accurate, at 79%).

Strictly logical models have had difficulty with tasks like spelling correction, speech recognition, and machine translation. I spent about 14 years trying to get logical models to work on language tasks. Then I made the switch: I started to adopt probabilistic approaches trained with statistics. And I saw everyone around me making the same switch. (And I did not see anyone going in the other direction.) There is an *engineering* reason for the switch: statistical models have state-of-the-art performance, and most logical models perform worse.

But for the rest of this essay we will ignore engineering and concentrate on *scientific* reasons: I assert that probabilistic models better represent linguistic facts, and statistical techniques make it easier for us to make sense of those facts.

What does Chomsky not like about statistical models?

In 1969 Chomsky 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."[3] In *Syntactic Structures*, he introduces a now

legendary example: "*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.*"[4]

Chomsky asserts that since neither sentence has occurred before, a statistical model must assign both a probability of zero, and thus can't distinguish between them, but a syntactic model can. This claim is true for the very simplest statistical models, but not for models that can generalize away from individual words. Fernando Pereira (2001)[5] (`http://www.cis.upenn.edu/%7Epereira/papers/rsoc.pdf`) showed that a simple bigram model that has a set of word classes trained by expectation maximization on newspaper text computes that (a) is 200 000 times more probable than (b). Furthermore, probabilistic models are capable of delivering the judgement that even (a) is *extremely* improbable, when compared to, say, "Effective green products sell well". Chomsky's categorical model cannot make this distinction; it can only distinguish between grammatical and ungrammatical.

Another part of Chomsky's objection is "*we cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds*". (Note that modern models can be much larger than the 10^9 parameters that were contemplated in the 1960s.) But of course nobody is proposing that these parameters are learned one by one; the right way to do learning is to set large swaths of near-zero parameters simultaneously with a smoothing or regularisation procedure, and update the high-probability parameters continuously as observations come in. And no one is suggesting that Markov models by themselves are a serious model of human language performance. But I (and others) suggest that probabilistic, trained models are a better model of human language performance than are categorical, untrained models. And yes, it seems clear that an adult speaker of English does know billions of language facts – for example, that one says "big game" rather than "large game" when talking about an important football game. No grammatical law distinguishes the two, but speakers of English have come to an agreement that the first is natural and the second is not.

Sense versus grammar

Probabilistic models are better for non-categorical judgements, such as the likelihood of a sentence, or the degree to which it makes sense. But even if you do not care about sense and are only interested in the grammaticality of sentences, probabilistic models still do a better job at describing the linguistic facts.

The *mathematical* theory of formal languages defines a language as a set of sentences. That is, every sentence is either grammatical or ungrammatical; there is no need for probability in this framework. The linguist's job is to describe rules that make the distinction. For example, consider the notion of a *pro-drop* language, from Chomsky's Lectures on Government and Binding (1981). The idea is that in Spanish, we say "Tengo hambre" ("have hunger"), dropping the pronoun, rather than "I have hunger." Chomsky's theory is that there is a "pro-drop parameter" which is set to "true" in Spanish and "false" in English (because we say "I'm hungry" not "am hungry"). But natural languages are not like formal languages. As Edward Sapir said in 1921: "All grammars leak." Pro-drop is certainly more rare in English than in Spanish, but it does appear, for example, in the exclamation "Found it!" and in the sentence "Turns out to be false" (taken from Chomsky's remark at the MIT symposium). So English's use of pronouns cannot be described by a single categorical parameter.

Reality is messier than theory. Models should be as simple as possible but no simpler

Now let us consider what I think is Chomsky's main point of disagreement with statistical models: the tension between "accurate description" and "insight". This is an old distinction. The physicist Ernest Rutherford disdained mere description, saying: "All science is either physics or stamp collecting." Chomsky stands with him: "You can also collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles."

The two statistical cultures

I think the most relevant rebuttal of this came from the statistician Leo Breiman[7] (1928–2005). Alluding to C.P. Snow, he describes two cultures. First, the *data modelling culture* (to which, Breiman estimates, 98% of statisticians subscribe) holds that nature can be described as a black box that has a relatively simple underlying model which maps from input variables to output variables (with perhaps some random noise thrown in). It is the job of the statistician to wisely choose an underlying model that reflects the reality of nature, and then use statistical data to estimate the parameters of the model.

Second is the *algorithmic modelling culture* (subscribed to by 2% of statisticians and many researchers in biology, artificial intelligence, and other fields that deal with complex phenomena) which holds that nature's black box cannot necessarily be described by a simple model. Complex algorithmic approaches (such as support vector machines, boosted decision trees or deep belief networks) are used to estimate the function that maps from input to output. We expect that the input/output behaviour of the learned function will be a close approximation to reality, but we are not concerned with the *form* of the function that emerges. The form will be a complex messy model, not a simple formula like $P=NkT/V$.

Breiman explained his objections to the first culture, data modelling. Basically, the conclusions made by data modelling are about the model, not about nature. If the model does not emulate nature well, then the conclusions may be wrong. For example, linear regression is one of the most powerful tools in the statistician's toolbox. Therefore, many analyses start out with "Assume the data are generated by a linear model ..." and lack sufficient analysis of what happens if the data are not in fact generated that way. Breiman is inviting us to give up on the idea that we can uniquely model the true underlying *form* of nature's function from inputs to outputs. Instead he asks us to be satisfied with a function that accounts for the observed data well (even if the function has a complex expression).

It is the algorithmic modelling culture that Chomsky is objecting to most vigorously. It is not just that models are statistical (or probabilistic), it is that they produce a form that cannot be read as a pithy explanatory principle. Chomsky wants to know the parameters – by analogy, the slope and intercept of the line that emerges from a linear regression model – so that the messy data can be discarded.

The problem is that reality is messier than this theory. We can all agree that models should be as simple as possible but no simpler (even if we can't agree whether Einstein said it).

Chomsky prefers a data modelling approach where the real issue for a linguist is to decide which model to use, and then try to fit language data to that model. "Observed use of language ... may provide evidence ... but surely cannot constitute the subject-matter of linguistics, if this is to be a serious discipline" he wrote in Aspects of the Theory of Syntax, in 1965.

Reproductions of some Lascaux artworks in Lascaux II (Photo: Jack Versloot). Source: Wikimedia Commons

As in Plato's allegory of the cave, Chomsky thinks we should focus on the ideal, abstract forms that underlie language, not on the superficial manifestations of language that happen to be perceivable in the real world. But Chomsky, like Plato, has to tell us where these ideal forms come from. Chomsky's answer is that they are innate to the mind, part of human biological endowment.

It was reasonable for Plato to think that the ideal of, say, a horse, was more important than any individual horse we can perceive in the world. In 400 BC, species were thought to be eternal and unchanging. We now know that is not true; that the horses shown on another cave wall—in Lascaux—are now extinct, and that current horses continue to evolve slowly over time. Thus there is no such thing as a single ideal eternal "horse" form.

We also now know that language is like that as well: languages are complex, random *biological* processes that are subject to the whims of evolution and cultural change. What constitutes a language is not an eternal ideal form, represented by the settings of a small number of parameters, but rather is the contingent outcome of complex processes. Since they are contingent, it seems they can only be analysed with probabilistic models. Since people have to continually understand the uncertain, ambiguous, noisy speech of others, it seems they must be using something like probabilistic reasoning. Chomsky for some reason wants to avoid this, and therefore he must declare the actual facts of language use (performance) out of bounds and declare that true linguistics only exists in the mathematical realm (competence), where he can choose the data-modelling formalism he wants. This may be very interesting from a mathematical point of view, but it misses the point about what language is, and how it works.

Acknowledgements

Thanks are due to Ann Farmer, Fernando Pereira, Dan Jurafsky, Shalom Lappin, Hal Varian, and others for comments and suggestions. A fuller version of this essay is at `http://norvig.com/chomsky.html`.

References

1. Shannon, C. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423.
2. Norvig, P. (2008) Natural language corpus data: beautiful data. `http://norvig.com/ngrams/` (accessed June 13th, 2012)
3. Chomsky, N. (1969) Quine's empirical assumptions. In D. Davidson and J. Hintikka (eds), *Words and Objections*. Dordrecht: Reidel.
4. Chomsky, N. (1957) *Syntactic Structures*. The Hague: Mouton & Co.
5. Pereira, F. (2002) Formal grammar and information theory: together again? In B. Nevin and S. M. Johnson (eds), *The Legacy of Zellig Harris*. Amsterdam: Benjamins.
6. Abney, S.(1996) Statistical methods and linguistics. In J. L. Klavans and P. Resnik (eds), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA: MIT Press.
7. Breiman, L. (2001) Statistical modeling: the two cultures. *Statistical Science*, **16**(3), 199–231.

Peter Norvig is Director of Research at Google Inc. Previously he was head of Computational Sciences at NASA and a faculty member at USC and Berkeley.