Speech and Language Processing: Where Have We Been and Where Are We Going?

Kenneth Ward Church

AT&T Labs-Research
180 Park Ave
Florham Park, NJ 07932 USA
church@att.com

Abstract

Can we use the past to predict the future? Moore's Law is a great example: performance doubles and prices halve approximately every 18 months. This trend has held up well to the test of time and is expected to continue for some time. Similar arguments can be found in speech demonstrating consistent progress over decades. Unfortunately, there are also cases where history repeats itself, as well as major dislocations, fundamental changes that invalidate fundamental assumptions. What will happen, for example, when petabytes become a commodity? Can demand keep up with supply? How much text and speech would it take to match this supply? Priorities will change. Search will become more important than coding and dictation.

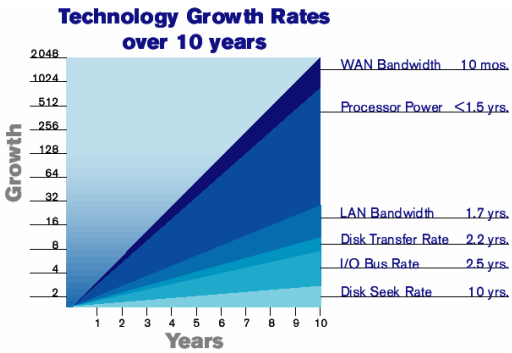

Figure 1: Moore's Law

1. Progress as a function of time

Moore's Law (Figure 1) uses past performance to predict future capability in a very convincing way. Figure 2 [1] demonstrates consistent progress in speech coding. Figure 3 [2] makes a similar argument for speech recognition; see [3] for predictions of where the field will go over the next 10 years.

We have grown accustom under Moore's Law to incredible improvements in disks, memory, CPU cycles, network bandwidth, etc. Performance doubles and prices halve every 18 months or so. The time constant varies considerably depending on physical limitations, market forces and other factors. Disk capacities, for example, have improved faster than disk seek times because of physical limitations. Market forces account for the dramatic improvements in PCs. PCs are not only cheaper than supercomputers, but ironically, they can be faster as well.

Speech coding has made remarkable progress over the years, though not at Moore's Law rates. Figure 2 shows significant quality improvements, especially at low bit rates ($\leq$ 8 kb/s) where there is more room for improvement. There is a quality ceiling imposed by telephone standards. At high bit rates ($\geq$ 8 kb/s), that ceiling was reached some time ago.

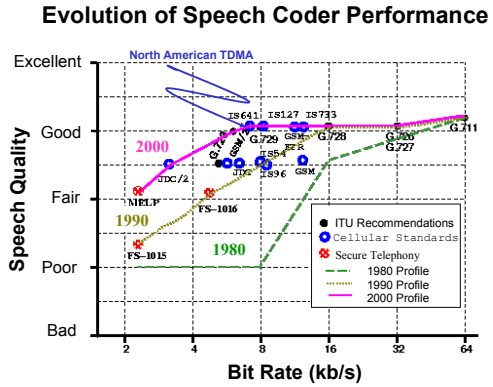

Figure 2: Advances in Speech Coding

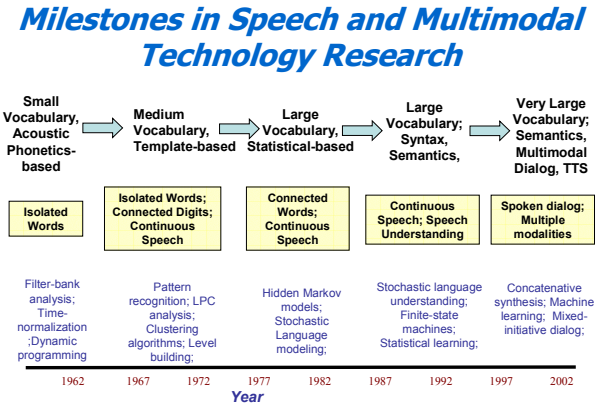

Figure 3: Advances in Speech Recognition

Extrapolating from Figure 2, we could hope for a factor of two improvement in space every decade or so, though we may already be running up against diminishing returns. In any case, speech rates (2 kb/s) are considerably higher than text rates (2 bits/character). Speech is likely to consume 100-1000 as much space as text for the foreseeable future.

Speech recognition has also made dramatic improvements, though these improvements are harder to quantify.

Figure 3 describes a transition from small vocabulary isolated word template matching to large vocabulary continuous speech recognition and interactive dialog systems. Figure 4 shows 15 years of consistent error rate reduction [4].

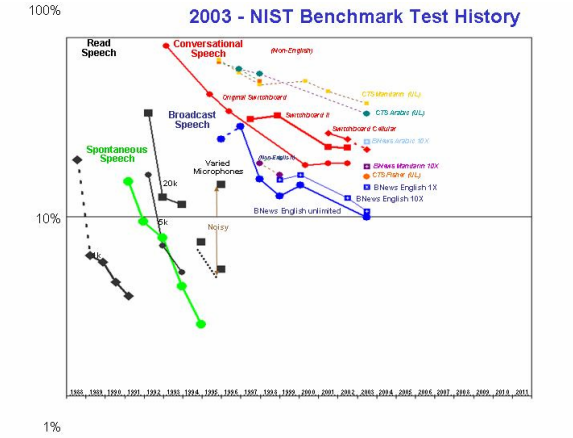

Figure 4: Challenge: demonstrate consistent progress

2. 1990s Revival of Empiricism: Progress (or Oscillating Fads)?

It is reassuring when there is consistent progress over decades, but there are also cases where history repeats itself. This is not necessarily a bad thing. It can be exciting and motivating to look at the world in a new way, or at least a way that hasn't been in vogue for a while.

Empiricism was at its peak in the 1950s, dominating a broad set of fields ranging from psychology (behaviorism) to electrical engineering (information theory). Zellig Harris [5], Chomsky's mentor, was advocating his distributional hypothesis. Firth [6] was advocating a similar position with his memorable line: *You shall know a word by the company it keeps.*

Interest in empiricism faded with Chomsky's criticism of ngram statistics in Syntactic Structures (1957) [7] and Minsky and Papert's criticism of neural networks in Perceptrons (1969) [8]. Rationalism (knowledge-based approaches) dominated the 1970s, especially in universities. A major objection to behaviorism (and empiricism in general) was that the methodology was so burdensome that it was getting in the way or progress. At the time, only the wealthiest industrial labs could afford data intensive methods.

As data intensive methods have become more affordable in the 1990s, thanks to advances in computing as well as data collection efforts such as the Linguistic Data Consortium (LDC) [9], these data intensive methods have become the method of choice. This process started in speech, but soon spread into natural language processing as well.

Historically, it took about 20 years to move from the empiricism of the 1950s to the rationalism of the 1970s and another 20 years to move back the other way to the empiricism of the 1990s. Based on this pattern, rationalism should be back in vogue in the 2010s. The empiricist methodology is becoming burdensome once again. Enhancing the representation can lead to larger performance gains in some cases than fine tuning of the learning procedure.

3. More Data is Better Data!

Mercer's *more data is better data* [10] was a pragmatic response to lengthy debates over balance. A lot of work went into creating samples of text, balanced corpora, to be representative of general English. The million word Brown corpus [11], the 20 million word Cobuild Corpous [12] and the 100 million word British National Corpus (BNC) [13] are some of the more influential examples of this tradition. Mercer's position no longer seems as shocking these days when Brill and others make more or less the same point. Many researchers are finding that performance continues to improve with corpus size. Probability estimates obtained from larger samples (e.g., private collections of billions of words, Google, AltaVista) seem to be better than those obtained from smaller balanced corpora.

"Better" can be defined in a number of ways ranging from language modeling to predicting psycholinguistic judgments. Rather than using smoothing techniques such as Katz-style back-off [14], it is preferable to simply collect more and more data when that is an option. Many researchers are finding that performance continues to improve as corpus sizes increase, over the full range of corpus sizes that they have been able to examine [15]. The rising tide of data will lift all boats.

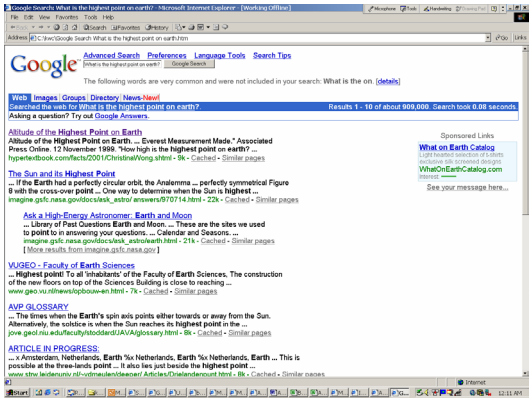

Figure 5: Google can answer questions like: "What is the highest point on Earth."

Using lots of data and not much more seems to work remarkably well in TREC [16] question answering tasks. Figure 5 above shows that Google can answer questions like "What is the highest point on Earth?" [17, 18]. Table 1 below illustrates Google's ability to find interesting sets of words. Given "cat" and "dog", labs1.google.com/sets returns a list of animals. Given a few countries, the page returns more countries, often from similar parts of the world. We used to try to do similar things a decade ago, but the results were not nearly as good, probably because we were working with relatively tiny corpora in the sub-billion-word range.

Of course, the web is hardly a balanced corpus. Norvig [18] made this point rather dramatically with the seed words, "cat" and "more," which caused labs1.google.com/sets to return a list of Unix commands! Unix commands and many other subjects, especially taboo subjects, are over represented on the web. There has been a lot of discussion in corpus linguistics on balance, but overall, the results mentioned above suggest that while balance is desirable, size is even more desirable.

Table 1: Google Sets (labs1.google.com/sets)

Cat	more	England	Japan
Dog	cat	France	China
Horse	ls	Germany	India
Fish	rm	Italy	Indonesia
Bird	mv	Ireland	Malaysia
Rabbit	cd	Spain	Korea
Cattle	cp	Scotland	Taiwan
Rat	mkdir	Belgium	Thailand
Livestock	man	Canada	Singapore
Mouse	tail	Austria	Australia
Human	pwd	Australia	Bangladesh

4. Bait and Switch

As large at the public Internet is, there are even larger opportunities. Changing copyright laws in various ways might unlock vast resources. Publishers like www.lexisnexis.com have impressive collections that may well surpass the public Internet. Private Intranets and telephone networks have even larger sources of linguistic data. Of course, most of the data on private Intranets cannot be distributed outside the Intranet, and most of the telephone traffic cannot even be recorded. But attitudes are changing. Voice mail took a while to catch on. It used to be considered rude to have an answering machine; now it is considered rude not to have one. It is hard to know how much speech could be recorded, but between answering machines and call centers [19], perhaps 10% can be recorded.

How large is large? According to [20], Google is nearly 1000 times larger than the British National Corpus (BNC); that is, Google has about 100 billion words compared to the BNC's 100 million. Telephone networks, of course, are much larger. According to [21], there are roughly 200 million telephone lines in America (Table 8.1), one for each person in the country. Each line is used about an hour a day (Table 11.2). If we assume that a second of speech corresponds to roughly a word, the American telephone network generates about 10 Google collections per day. It is hard to say how large the private Intranets are, but right now, at least, revenue-wise, the data networks are roughly comparable to the voice networks, though data is growing faster than voice.

In the past, recording all this data would have been prohibitively expensive. But thanks to Moore's Law, storage costs have been falling faster than transport for some time and will continue to do so. Even at current prices, if I am willing to pay for a long-distance telephone call (5 cents/minute), I might as well pay for the disk space to keep the speech online (0.5 cents/minute). Similar comments hold for web pages. Why flush a page if there is any chance that it might be requested again? Over time, web caches will look more like web crawlers. Go find the pages that I might ask for and keep them forever.

The proposed bait and switch strategy is to use the public Internet as the bait to develop and test and socialize new ways of extracting value from large linguistic repositories. The value to society, though, is when these solutions are applied to the private repositories that we care about. (No one cares about data that everyone can have, just as Grocho Marx doesn't want to be in a club that would have him.)

To a large extent this strategy encourages the research community to keep doing more of all the great stuff that we have been doing. There will be more interest in papers that not only report performance on currently available corpora, but also report how well the techniques port from one corpus to another. There will also be more interest in papers that report how well performance scales with corpus size. (Hopefully, performance improves with corpus size, though it doesn't always.) There have been many examples of such papers in the past, and hopefully there will be many more. It should be noted, of course, that all the data in the world will not solve all the world's problems. It would be useful to know when more data will help and when it is better to do something else (e.g., a revival of linguistic representation).

As for investments in infrastructure, in addition to traditional data collection efforts that are focused on public repositories, we ought to think about private repositories as well. Most of us, for example, do not keep voice mail for very long, though I have been using Scanmail [22] to copy my voice mail to my email, and like many people, I keep a lot of email online for a long time. Unfortunately, the tools for searching email archives and other private repositories are not as good as the tools for searching public repositories. We could make a huge difference in the size of private repositories by making it more convenient to capture private data, and by demonstrating that there is value in doing so.

5. Meeting Demand for Petabytes

Extrapolating from Moore's Law, petabytes are coming [23]. A petabyte costs $2,000,000 today, but in a decade, it will cost just $2,000. Can demand keep up with supply? If not, prices will collapse and there will be an industry meltdown.

How do we explain to a lay audience what a petabyte is and why they are all going to buy lots and lots of them? A petabyte is a huge: 10^{15} bytes. We used to call that a zillion. That is, 10^6 is a million, 10^9 is a billion, 10^{12} is a trillion and 10^{15} is a zillion. Not all that long ago, a zillion was synonymous with infinity, an unimaginably large number.

Will everyone have so much disk space and network bandwidth that they will no longer need to buy any more? What is everyone going to do with their own petabyte(s)? Are they going to fill them up with text? Speech? Video?

A petabyte per lifetime is about 18 megabytes per minute. It is hard to imagine how we could all produce or consume 18 megabytes of text per minute per capita forever. No one can type that fast. No one can dictate that fast. No one can read that fast. It is hard to imagine that everyone's personal library will be in the petabyte range.

Admittedly, the larger text repositories such as Google are approaching the petabyte range, but those repositories are shared by billions of people. If we amortize the demand for such shared resources over the beneficiaries, the per capita demand falls well short of the projected supply.

It is also hard to see how speech is going to create that much demand, since speech is only 100-1000 times larger than text, as mentioned above. A petabyte/lifetime can also be thought of as 317 telephone channels for 100 years per capita. It is hard to imagine why anyone would want to record that much speech.

DVD Video could solve the problem. DVD video rates are 1.8 gigabytes per hour which equates to 1.6 petabytes per

100 years (lifetime). Unfortunately, there are various video compression mechanisms that could reduce the demand considerably. In addition, there have been a number of attempts to sell Picture Phones in the past, with limited success. It is not clear that everyone will want everything to be recorded on DVD Video forever.

Bell and Gray [23] estimate how much storage it would take to archive a person's life. What is disturbing about these estimates, shown in Table 2, is how small they are. Even if everyone stored their entire life forever, there still would not be enough demand to keep up with the petabytes of supply that are coming.

Table 2: Digital Immortality: Storage Requirements

Data-types	Per day	Per Lifetime
email, papers, text	0.5 MB	15 GB
Photos	2 MB	150 GB
Speech	40 MB	1.2 TB
Music	60 MB	5.0 TB
video-lite (200 Kb/s)	1 GB	100 TB
DVD video (1.8 GB/hour)	20 GB	1 PB

6. Conclusions

Moore's Law provides a nice answer to where have we been and where are we going for many technologies. In the speech and language area, it is harder to quantify progress so elegantly in terms of price and performance as a function of time. Nevertheless, there are some compelling demonstrations of consistent progress over many years.

Sometimes progress moves in a consistent direction and sometimes history repeats itself and sometimes there are major dislocations, fundamental changes that invalidate fundamental assumptions. The oscillations between empiricism and rationalism were cited as an example of history repeating itself. The current cycle, the empiricist revival of the 1990s, was brought about by the availability of massive amounts of data and computing power, a major dislocation that caused many researchers to question the knowledge intensive approaches that were in vogue at the time.

More data is better data! The rising tide of web data will lift all boats. Performance continues to improve as more and more data becomes available. Using lots of data and not much else seems to work remarkably well for many applications such as language modeling TREC question answering and labs1.google.com/sets.

As large at the public Internet is, there are even larger opportunities. Private Intranets and telephone networks have much larger sources of linguistic data. Of course, most of the data on private Intranets and telephone networks cannot be distributed. A bait and switch strategy was proposed. The bait is the public Internet which is large, sexy and available. The public Intranet would be used to develop, test and socialize new ways to extract value from large linguistic repositories, but once solutions are identified they would be applied to private repositories. Switching successful applications to exclusive data sets creates value.

Where are we going? Moore's Law makes it very clear that petabytes are coming. The availability of massive amounts of storage will cause major dislocations. It will be possible for everyone to store lots and lots of text, speech and video. Priorities will change. Search will become a killer app, more important than compression (speech coding) and data entry (dictation). But even if everyone stored everything I can possibly think they might want to store, I still don't see how demand can keep up with supply.

7. References

CL: *Computational Linguistics*; ACL: aclweb.org; SIGIR: sigir.org

[1] Cox, R., personal communication, 2003.

[2] Rahim, M., personal communication 2003.

[3] Bernsen, N. (ed), "Speech-Related Technologies: Where will the field go in 10 years?" *ELSNET Roadmap Workshop*, 2000. Available from elsnet.dfki.de/task.php.

[4] Le, A., personal communication, 2003.

[5] Harris, Z. S., "Distributional structure," *Word*, 10, pp. 146-162, 1954. See also: www.dmi.columbia.edu/zellig.

[6] Firth, J., "A Synopsis of Linguistic Theory 1930-1955," 1957.

[7] Chomsky, N., *Syntactic Structures*, Mouton, The Hague, 1957.

[8] Minsky, M. and Papert, S., *Perceptrons*. MIT Press, Cambridge, MA, 1967.

[9] *Linguistic Data Consortium*, www.ldc.upenn.edu

[10] Church, K. and Mercer, R., "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," *CL* 19:1, pp. 1-24, 1993.

[11] Kucera, H. and Francis, W., *Computational Analysis of Present-Day American English*. Brown University Press, Providence, R.I., 1967.

[12] Sinclair J., *Looking up: An Account of the COBUILD Project in Lexical Computing*. Collins, Glasgow, 1987.

[13] *British National Corpus,* info.ox.ac.uk/bnc.

[14] Katz, S. M. "Estimation of probabilities from sparse data for the language model component of a speech recognizer," ASSP-35(3):400-401, 1987.

[15] Banko, M. and Brill, E., "Scaling to Very Very Large Corpora for Natural Language Disambiguation," *ACL-01*.

[16] *TREC*, http://trec.nist.gov/

[17] Dumais, S., Banko, M., Brill, M., Lin, J., and Ng, A., "Web Question Answering: Is More Always Better?" *SIGIR*, 2002.

[18] Norvig, P., "Better Web Search with and Without Computational Linguistics," *ACL-02*.

[19] Whiting, R., "Hidden Value in Customer Calls," *Information Week*, Aug 2002, see informationweek.com.

[20] Keller, F. and Lapata, M., "Using the Web to Obtain Frequencies for Unseen Bigrams," *CL* 29:3, 2003.

[21] *Statistical Trends in Telephony*, available from www.fcc.gov/Bureaus/Common_Carrier/Reports/FCC-State_Link/trends.html, 2001.

[22] Rosenberg, A., Hirschberg, J., *et al*., "Caller Identification for the SCANmail Voicemail Browser," *Eurospeech*, 2001.

[23] Gray, J. and Hey, T., "In Search of Petabytes," *High Performance Transaction Systems Workshop*, available from research.microsoft.com/~Gray/talks, 2001.

[24] Bell, G. and Gray, J. *Digital Immortality*, MSR-TR-2000-101, 2000. available from research.microsoft.com/~Gray.