What's Happened Since the First SIGDAT Meeting?

Kenneth Ward Church
AT&T Labs-Research
180 Park Ave
Florham Park, NJ 07932-0971
kwc@research.att.com

Abstract

The first workshop on Very Large Corpora was held just before the 1993 ACL meeting in Columbus, Ohio. The turnout was even greater than anyone could have predicted (or else we would have called the meeting a conference rather than a workshop). We knew that corpus-based language processing was a "hot area," but we didn't appreciate just how hot it would turn out to be.

The 1990s were witnessing a resurgence of interest in 1950s-style empirical and statistical methods of language analysis. Empiricism was at its peak in the 1950s, dominating a broad set of fields ranging from psychology (behaviorism) to electrical engineering (information theory). At that time, it was common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words. Firth, a leading figure in British linguistics during the 1950s, summarized the approach with the memorable line: "You shall know a word by the company it keeps." Regrettably, interest in empiricism faded in the late 1950s and early 1960s with a number of significant events including Chomsky's criticism of n-grams in *Syntactic Structures* (Chomsky, 1957) and Minsky and Papert's criticism of neural networks in *Perceptrons* (Minsky and Papert, 1969).

Perhaps the most immediate reason for this empirical renaissance is the availability of massive quantities of data: text is available like never before. Just ten years earlier, the one-million word Brown Corpus (Francis and Kucera, 1982) was considered large, but these days, everyone has access to the web. Experiments are routinely carried out on many gigabytes of text. Some researchers are even working with terabytes.

The big difference since the first SIGDAT meeting in 1993 is that large corpora are now having a big impact on ordinary users. Web search engines/portals are an obvious example. Managing gigabytes is not only the title of a popular book that recently came out with a second edition (Moffat, Bell and Witten, 1999), but it is something that ordinary users are beginning to take for granted. Recent progress in Information Retrieval and Digital Libraries is worth a fortune (according to the stockmarket). Speech Recognition and Machine Translation are also changing the world. If you walk into any software store these days, you will find a shelf full of speech recognition and machine translation products. And it is getting so you can't use the telephone these days without talking to a computer.

References

Chomsky, N. 1957. *Syntactic Structures*, The Hague: Mouton & Co.

Firth, J. 1957. A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis*, Philological Society, Oxford; reprinted in Palmer, F. (ed.), 1968, *Selected Papers of J. R. Firth*, Longman, Harlow.

Francis, W., and Kucera, H. 1982. *Frequency Analysis of English Usage,* Houghton Mifflin Company, Boston.

Minsky, M. and Papert, S. 1969. *Perceptrons; An Introduction to Computational Geometry*, MIT Press, Cambridge, MA.

Moffat, A, Bell, T., and Witten, I. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*, Academic Press/Morgan Kaufmann.