

From Human Language Technology to Human Language Science

Mark Liberman¹

(1) Linguistic Data Consortium. University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA, 19104 USA

myl@cis.upenn.edu

ABSTRACT

Thirty years ago, in order to get past roadblocks in Machine Translation and Automatic Speech Recognition, DARPA invented a new way to organize and manage technological R&D : a “common task” is defined by a formal quantitative evaluation metric and a body of shared training data, and researchers join an open competition to compare approaches. Over the past three decades, this method has produced steadily improving technologies, with many practical applications now possible. And Moore’s law has created a sort of digital shadow universe, which increasingly mirrors the real world in flows and stores of bits, while the same improvements in digital hardware and software make it increasingly easy to pull content out of the these rivers and oceans of information.

It’s natural to be excited about these technologies, where we can see an open road to rapid improvements beyond the current state of the art, and an explosion of near-term commercial applications. But there are some important opportunities in a less obvious direction. Several areas of scientific and humanistic research are being revolutionized by the application of Human Language Technology. At a minimum, orders of magnitude more data can be addressed with orders of magnitude less effort – but this change also transforms old theoretical questions, and poses new ones. And eventually, new modes of research organization and funding are likely to emerge.
