Lessons for Reproducible Science from DARPA’s Programs in Human Language Technology

Mark Liberman
University of Pennsylvania

http://ling.upenn.edu/~myl
The story begins in the 1960s…

… with two bad reviews by John Pierce, an executive at Bell Labs who invented the word “transistor” and supervised development of the first communications satellite.
In 1966, John Pierce chaired the Automatic Language Processing Advisory Committee (ALPAC) which produced a report to the National Academy of Sciences, *Language and Machines: Computers in Translation and Linguistics*

And in 1969,

he wrote a letter to the Journal of the Acoustical Society of America, “Whither Speech Recognition”
MT in 1966 was not very good, and ALPAC said diplomatically that

“The Committee cannot judge what the total annual expenditure for research and development toward improving translation should be. However, it should be spent hardheaded toward important, realistic, and relatively short-range goals.”

In fact, U.S. MT funding went essentially to zero for more than 20 years.

The committee felt that science should precede engineering in such cases:

“We see that the computer has opened up to linguists a host of challenges, partial insights, and potentialities. We believe these can be aptly compared with the challenges, problems, and insights of particle physics. Certainly, language is second to no phenomenon in importance. And the tools of computational linguistics are considerably less costly than the multibillion-volt accelerators of particle physics. The new linguistics presents an attractive as well as an extremely important challenge.”
John Pierce’ views about automatic speech recognition were similar to his opinions about MT.

And his 1969 letter to JASA was much less diplomatic than that 1966 N.A.S. committee report....
“Whither Speech Recognition?”

“… a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English.”

“Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve ‘the problem.’ The basis for this is either individual inspiration (the ‘mad inventor’ source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach). . . .”

“The typical recognizer … builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. **No simple, clear, sure knowledge is gained.** The work has been an experience, not an experiment.”
“We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn’t attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamor.”

“It is clear that glamor and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect.”
Fallout from these blasts

The first idea: **Try Artificial Intelligence . . .**

DARPA Speech Understanding Research Project (1972-75)
  Used classical AI to try to “understand what is being said with something of the facility of a native speaker”
  DARPA SUR was viewed as a failure; funding was cut off after three years

The second idea: **Give Up.**

**1975-1986: No U.S. research funding for MT or ASR**
Pierce was far from the only person with a jaundiced view of R&D investment in the area of human language technology.

By the mid 1980s, many informed American research managers were equally sceptical about the prospects.

At the same time, many people believed that HLT was needed and in principle was feasible.
Charles Wayne -- DARPA program manager – has an idea.

He’ll design a speech recognition research program that

– protects against “glamour and deceit”
  • because there is a well-defined, objective evaluation metric
  • applied by a neutral agent (NIST)
  • on shared data sets; and
– and ensures that “simple, clear, sure knowledge is gained”
  • because participants must reveal their methods
  • to the sponsor and to one another
  • at the time that the evaluation results are revealed

In 1986 America,

no other sort of ASR program
could have been gotten large-scale government funding.
“Common Task” structure

• A detailed “evaluation plan”
  – is developed in consultation with researchers
  – and published as the first step in the project.

• Automatic evaluation software
  – is written and maintained by NIST
  – and published at the start of the project.

• Shared data:
  – Training and “dev(elopment) test” data is published at start of project;
  – “eval(uation) test” data is withheld for periodic public evaluations
Not everyone liked it

Many Piercian engineers were skeptical:
  you can’t turn water into gasoline,
  no matter what you measure.

Many researchers were disgruntled:
  “It’s like being in first grade again --
  you’re told exactly what to do,
  and then you’re tested over and over .

But it worked.
Why did it work?

1. The obvious: it allowed funding to start
   \textit{(because the project was glamour-and-deceit-proof)}
   and to continue
   \textit{(because funders could measure progress over time)}

2. Less obvious: it allowed project-internal hill climbing
   \begin{itemize}
   \item because the evaluation metrics were automatic
   \item and the evaluation code was public
   \end{itemize}
   This obvious way of working was a new idea to many!
   \begin{itemize}
   \item … and researchers who had objected to be tested twice a year
   \item began testing themselves every hour…
   \end{itemize}

3. Even less obvious: it created a culture
   \begin{itemize}
   \item (because researchers shared methods and results
   \item on shared data with a common metric)
   \end{itemize}

   Participation in this culture became so valuable
   that many research groups joined without funding
What else it did

The *common task method* created a positive feedback loop.

When everyone’s program has to interpret the same ambiguous evidence, ambiguity resolution becomes a sort of gambling game, which rewards the use of statistical methods.

Given the nature of speech and language, statistical methods need the largest possible training set, which reinforces the value of shared data.

Iterated train-and-test cycles on this gambling game are addictive; they create “*simple, clear, sure knowledge*”, which motivates participation in the common-task culture.
The past 25 years

Variants of this method have been applied to many other problems:

- machine translation,
- speaker identification,
- language identification,
- parsing,
- sense disambiguation,
- information retrieval,
- information extraction,
- summarization,
- question answering,
- OCR,
- …, etc.

The general experience:

1. Error rates decline by a fixed percentage each year,
   to an asymptote
   which is defined by the quality of the data
   and the difficulty of the task.
2. Progress usually comes from many small improvements;
   a change of 1% can be a reason to break out the champagne.
   Thus the larger the community, the faster the progress.
3. Glamour and deceit have been avoided.

...and self-sustaining ignition has been achieved!
[CITATION] Darpa Timit: Acoustic-phonetic Continuous Speech Corps CD-ROM
JS Garofolo, LF Lamel, WM Fisher, JG Fiscus… - 1993 - citeulike.org
... CiteULike is a free online bibliography manager. Register and you can start organising your references online. Tags. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM ...
Cited by 578 - Related articles - Cached - Library Search

Speech database development at MIT: TIMIT and beyond
V Zue, S Seneff… - Speech Communication, 1990 - Elsevier
Automatic speech recognition by computers can provide the most natural and efficient method of communication between humans and computers. While in recent years high performance speech recognition systems are beginning to emerge from research institutions, scientists ...
Cited by 153 - Related articles - All 2 versions

[CITATION] Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database
JS Garofolo - 1988 - National Institute of Standards and ...
Cited by 132 - Related articles

MMI training for continuous phoneme recognition on the TIMIT database
S Kapadia, V Valtchev… - Acoustics, Speech, and ..., 2002 - ieeexplore.ieee.org
ABSTRACT This paper reports our experiences with a phoneme recognition system for the TIMIT database which uses multiple mixture continuous density monophone HMMs trained using MMI. A comprehensive set of results are presented comparing the ML and MMI training ...
Cited by 69 - Related articles - All 3 versions

[CITATION] Transcription and alignment of the timit database
S Seneff… - TIMIT CD-ROM Documentation, 1988
Cited by 60 - Related articles

Phonetic analyses of word and segment variation using the TIMIT corpus of American English
PA Keating, D Byrd, E Flemming… - Speech Communication, 1994 - Elsevier
This paper reports a set of studies of some phonetic characteristics of the American English represented in the TIMIT speech database. First we describe some relevant characteristics of TIMIT, and how we use the non-speech files on the TIMIT CD with a commercial database program. Two ...
Cited by 43 - Related articles - All 5 versions
The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms
DB Dean, S Sridharan, RJ Vogt... - ... of Interspeech 2010, 2010 - eprints.qut.edu.au
This is the accepted version of this article. To be published as: This is the author version published as: ... QUT Digital Repository: http://eprints.qut.edu.au/ ... Dean, David B. and Sridharan, Sridha and Vogt, Robert J. and Mason, Michael W. (2010) The QUT-NOISE-TIMIT corpus for the ...
Cited by 1

Robust speaking rate estimation using broad phonetic class recognition
J Yuan... - Acoustics Speech and Signal ..., 2010 - ieeexplore.ieee.org
... to predict possible regions where syllable nuclei can appear, and then a simple slope based peak counting algorithm was used to get the positions of the syllable nuclei [9]. Although these syllable detection algorithms work well on short and fluent speech, eg, TIMIT, it remains a ...
Related articles - All 2 versions

An analysis of sparseness and regularization in exemplar-based methods for speech classification
D Kanevsky, TN Sainath... - ... Conference of the ..., 2010 - sites.google.com
... The goal of this paper is to answer the above two questions, both through mathematically analyzing different sparseness methods and also comparing these approaches for phonetic classification in TIMIT. 1. Introduction ... Section 3 gives a brief description of the TIMIT corpus. ...
Cited by 1 - View as HTML

Noise-Robust Voice Activity Detector Based on Hidden Semi-Markov Models
X Liu, Y Liang, Y Lou, H Li... - 2010 - nsde.buaa.edu.cn
... Motivated by statistical observations and tests on TIMIT and the IEEE sentence database, we use Weibull distributions to model state durations approximately and estimate their parameters by maximum likelihood estimators. ...
Cited by 1 - Related articles - View as HTML - All 4 versions

Sparse Representation Features for Speech Recognition
TN Sainath, B Ramabhadran... - ... Annual Conference of ..., 2010 - wiki.inf.ed.ac.uk
... On the TIMIT corpus, we show that applying the SR features on top of our best discriminatively trained system allows for a 0.7% absolute reduction in phonetic error rate (PER), from 19.9% to 19.2%. In fact, after applying model adaptation we reduce the PER to 19.0%, the ...
Automatic Content Extraction (ACE) Evaluation

What is Automatic Content Extraction (ACE)?

The objective of the ACE program is to develop automatic content extraction technology to support automatic processing of human language in text form from a variety of sources (such as newswire, broadcast conversation, and weblogs). ACE technology R&D is aimed at supporting various classification filtering, and selection applications by extracting and representing language content (i.e., the meaning conveyed by the data). Thus the ACE program requires the development of technologies that automatically detect and characterize this meaning.

The ACE program will be carried out in several phases, beginning with EDT (Entity Detection and Tracking) Phase-1.

View the presentation describing the ACE program that was given at the TIDES program kick-off meeting.

Current and Recent ACE Activities

The most recent ACE evaluation was ACE08 and took place in May 2008.

Results of recent ACE Evaluations:

- NIST ACE08 Official Evaluation Results
- NIST ACE07 Official Evaluation results

ACE is becoming a track in the Text Analysis Conference (TAC) in 2009. Please explore the TAC website.

Find more information on past ACE evaluations by clicking a specific year in the tabs below.

[ 1999 ] [ 2000 ] [ 2001 ] [ 2002 ] [ 2003 ] [ 2004 ] [ 2005 ] [ 2007 ] [ 2008 ]
Automatic content extraction and named entity recognition based on Conditional Random Fields

X Guo, Y Du, X Lv... - Innovative Computing, Information ..., 2010 - ieeexplore.ieee.org

The task of entity relation extraction was first proposed in MUC-7 conference in 1998 [1]. The ACE conference in 1999 which is organized by NIST (National Information Service for Science and Technology) also continue the evaluation task of the...

An Evaluation of Technologies for Knowledge Base Population

P McNamee, HT Dang, H Simpson... - Proceedings of the ..., 2010 - lrec-conf.org

... NIST-organized community evaluation of Natural Language Processing technologies which began in 2008... The TAC-KBP evaluation builds on the work of the Automatic Content Extraction (ACE) (Doddington et al., 2004) and the Question Answering (QA) evaluations of the Text...

Text analysis and entity extraction in asymmetric threat response and prediction

E Chan, J Ginsburg, B Ten Eyck... - ... (ISI), 2010 IEEE ..., 2010 - ieeexplore.ieee.org

... NIST-organized community evaluation of Natural Language Processing technologies which began in 2008... The TAC-KBP evaluation builds on the work of the Automatic Content Extraction (ACE) (Doddington et al., 2004) and the Question Answering (QA) evaluations of the Text...

Transformation-based named entity extraction from spoken content for personal memory aid

JH Kim - Consumer Electronics, IEEE Transactions on, 2010 - ieeexplore.ieee.org

... news workshop [4] and the 2002 and 2003 Conferences on Natural Language Learning (CoNLL) [9][10], the Automatic Content Extraction (ACE) [11]... As the 1998 NIST Hub-4 evaluation used broadcast news data, each participant in this evaluation was required to handle various...

Annotating Event Chains for Carbon Sequestration Literature

H Ji, X Li, A Lucia... - Proc. LREC, 2010 - nlp.cs.qc.cuny.edu

Where we were

ANLP-1983
(First Conference on Applied Natural Language Processing)

34 Presentations:

None use a published data set.
None use a formal evaluation metric.

Two examples:

Wendy Lehnert and Steven Shwartz,
“EXPLORER: A Natural Language Processing System for Oil Exploration”.
Describes problem and system architecture; gives examples of queries and responses.
No way to evaluate performance or to compare to other systems/approaches.

Larry Reeker et al.,
“Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions”
Describes problem and system architecture; gives examples of inputs and outputs.
No way to evaluate performance or to compare to other systems/approaches.
Where we are

ACL-2010
(48th Annual Meeting of the Association for Computational Linguistics)

274 presentations –
   Nearly all use published data and published evaluation methods. (A few deal with new data-set creation and/or new evaluation metrics.)

Three examples:

Nils Reiter and Anette Frank, "Identifying Generic Noun Phrases". Authors are from Heidelberg University; use ACE-2 data.

Shih-Hsiang Lin and Berlin Chen, "A Risk Minimization Framework for Extractive Speech Summarization". Authors are from National Taiwan University; use Academia Sinica Broadcast News Corpus and the ROUGE metric (developed in DUC summarization track).

Laura Chiticariu et al., "An Algebraic Approach to Declarative Information Extraction". Authors are from IBM Research; use ACE NER metric, ACE data, ENRON corpus data.
And yet...

In the area of HLT, a form of “reproducible research” has been in place for more than 20 years.

This is based on shared data and shared evaluation metrics – but NOT shared code (in most cases…)

Nevertheless, results really are reproducible –

at least most of time –

and we usually find out pretty quickly when they’re not.

In some ways, this is a Good Thing – because we avoid replicating bugs.

But this depends on having an unusual source of motivation for doing the work needed to try a replication.
Science is different!

- Explanations, not applications
- etc.
- etc.
- etc.
- etc.
But not that different…

Sharing data and problems
  – lowers costs and barriers to entry
  – creates intellectual communities
  – speeds up replication and extension
  – guards against glamour and deceit
    (...as well as simple confusion)

This is true in many areas of science
as well as in engineering
Thank you!