Reproducible Computational Experiments

Mark Liberman
University of Pennsylvania

http://ling.upenn.edu/~myl

Reproducible (?) Replicable Computational Experiments

Mark Liberman
University of Pennsylvania

http://ling.upenn.edu/~myl

Reproducible (?) Replicable (?) Meaningful Computational Experiments

Mark Liberman
University of Pennsylvania

http://ling.upenn.edu/~myl

The "Common Task Method"

A research paradigm
in experimental computational science:

- Shared training and testing data
- Well-defined evaluation metric
- Techniques to avoid over-fitting

Setting:
Algorithmic analysis of the natural world.

Dozens of current examples –

Sometimes shared-task workshops:

Conference on Natural Language Learning Shared Task for 2015 (CoNLL2015)
Open Keyword Search Evaluation 2015 (OpenKWS2015)
Open Machine Translation Evaluation (OpenMT2014, OpenMT2016)
Reconnaissance de Personnes dans les Emissions Audiovisuelles (REPERE2014)
Speaker Recognition Evaluation (SRE2014, SRE2016)
Text Retrieval Conference (TREC2015)
DiscoMT 2015 Shared Task on Pronoun Translation
TREC Video Retrieval Evaluation (TRECVID2015)
IMAGENET Large Scale Visual Recognition Challenge 2015
. . . and many others . . .

Sometimes just shared datasets and evaluation metrics.

For example, TAC 2014:

"The Text Analysis Conference (TAC) is a series of evaluation workshops organized to encourage research in Natural Language Processing and related applications, by providing a large test collection, common evaluation procedures, and a forum for organizations to share their results. TAC comprises sets of tasks known as "tracks," each of which focuses on a particular subproblem of NLP. TAC tracks focus on end-user tasks, but also include component evaluations situated within the context of end-user tasks."

TAC 2014 hosts evaluations in two areas of research:

Knowledge Base Population (KBP) The goal of Knowledge Base Population is to promote research in automated systems that discover information about named entities as found in a large corpus and incorporate this information into a knowledge base.

Biomedical Summarization (BiomedSumm) The goal of BiomedSumm is to develop technologies that aid in the summarization of biomedical literature.

CoNLL Shared Task for 2015:

"Since the first CoNLL Shared Task on NP chunking in 1999, CoNLL shared tasks over the years have tackled increasingly complex natural language learning tasks. Early shared tasks focused on identifying text chunks or named entities that typically correspond to single words or short phrases within a sentence. Shared tasks on semantic role labeling are concerned with identifying arguments for individual predicates and characterizing the relationship between each argument and the predicate. Shared tasks on joint dependency parsing and semantic role labeling target the syntactic and semantic structure of the entire sentence, rather than the argument structure of individual predicates. More recently, shared tasks on coreference went beyond sentence boundaries and started to deal with discourse phenomena, and shared tasks on grammatical error correction dealt with detecting and correcting grammatical errors in texts."

TRECVID 2015:

"The main goal of the TREC Video Retrieval Evaluation (TRECVID) is to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation. TRECVID is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations."

In TRECVID 2015 NIST will continue 4 of the 2014 tasks with some revisions […], drop one […], separate out the localization task from semantic indexing, and add a new Video Hyperlinking task previously run in MediaEval:

Semantic indexing [IACC]
Interactive surveillance event detection [i-LIDS]
Instance search [BBC EastEnders]
Multimedia event detection [HAVIC]
Localization [IACC]
Video Hyperlinking [BBC for Hyperlinking]

Street View House Numbers (SVHN) dataset:

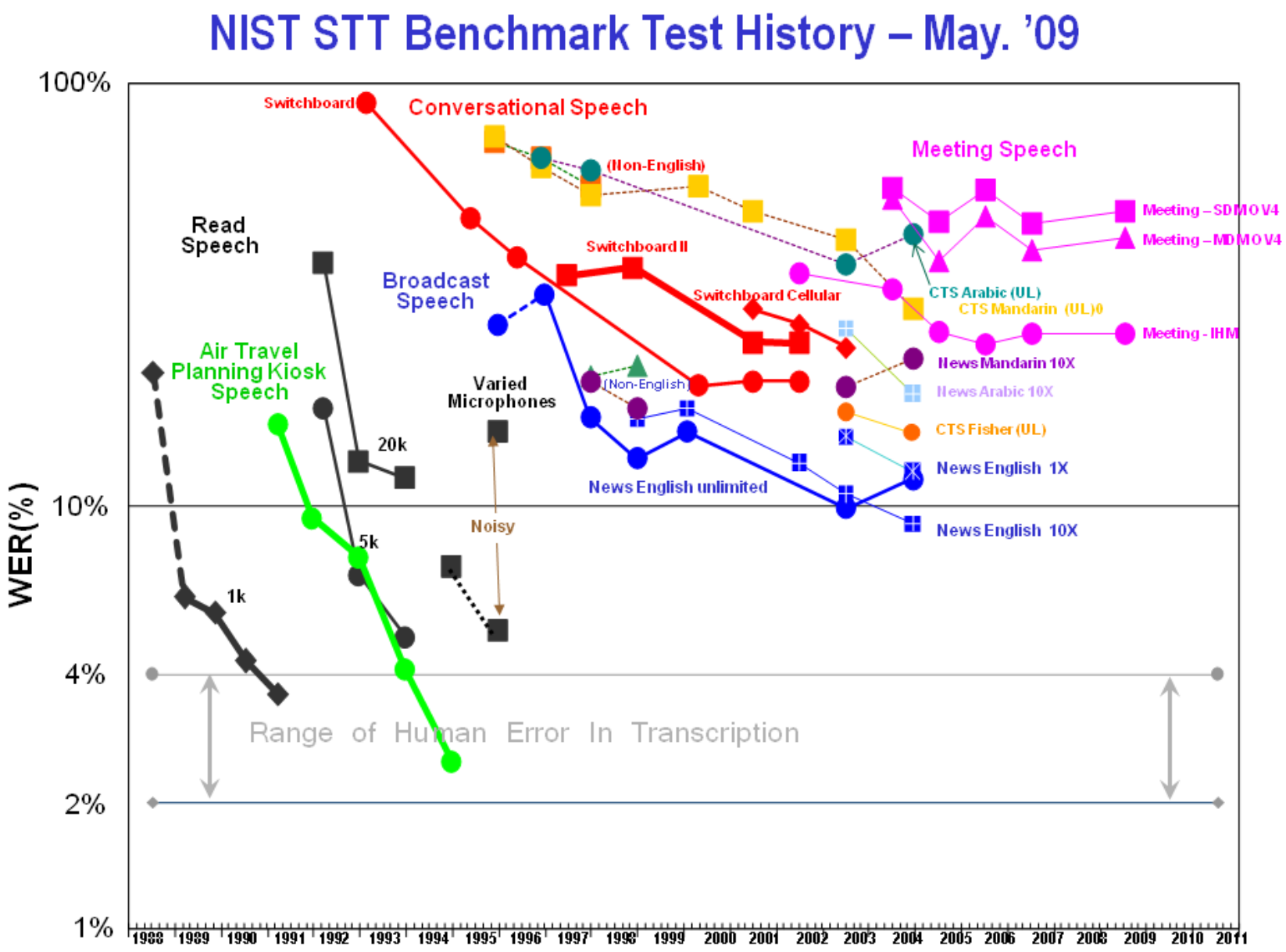
"SVHN is a real-world image dataset for developing machine learning and object recognition algorithms with minimal requirement on data preprocessing and formatting. It can be seen as similar in flavor to MNIST (e.g., the images are of small cropped digits), but incorporates an order of magnitude more labeled data (over 600,000 digit images) and comes from a significantly harder, unsolved, real world problem (recognizing digits and numbers in natural scene images). SVHN is obtained from house numbers in Google Street View images."

73257 digits for training,
26032 digits for testing,
and 531131 additional samples to use as extra training data.

Progress in performance on SVHN:

Error (%)	Method	Reference
36.7	WDCH	Netzer et al. (2011)
15	HOG	Netzer et al. (2011)
9.4	KNN	Netzer et al. (2011)
2.47	conv-DNN	Goodfellow et al. (2013)
2	Human	Netzer et al. (2013)
1.92	conv-DNN	Lee et al. (2015)

Progress is not always this rapid –
but steady progress almost always happens.


NIST STT Benchmark Test History – May. '09
100%
10%
4%
2%
1%
WER(%)
Switchboard
Conversational Speech
(Non-English)
Meeting Speech
Meeting – SDMOV4
Meeting – MDMOV4
Read Speech
Switchboard II
Broadcast Speech
Switchboard Cellular
CTS Arabic (UL)
CTS Mandarin (UL)0
Meeting - IHM
Air Travel Planning Kiosk Speech
Varied Microphones
(Non-English)
News Mandarin 10X
News Arabic 10X
CTS Fisher (UL)
20k
News English 1X
News English unlimited
Noisy
5k
News English 10X
1k
Range of Human Error In Transcription
1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011

Continued progress in Speech-to-Text

Switchboard Corpus of conversational telephone speech

Stalled at 20-30% word error rate 15 years ago:

WER (%)	Acoustic Model	Reference
48.7	GMM	Jeanrenaud et al. (1995)
18.6	GMM	Vesely et al. (2013)
17.1	DNN	Seide et al (2011)
14.3	DNN	Maas et al. (2014)
12.6	DNN	Vesely et al. (2013)
12.6	deep-RNN	Hannun et al. (2014)
10.4	conv-DNN	Soltau et al. (2014)

‘Twas not always thus –

The common task method is bred in the bone of researchers today.

But the situation 30 years ago was completely different.

And a cultural shift of this magnitude deserves an origin story.

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.


ASSOCIATES

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"

The ALPAC Report

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 hardheadedly 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's views
about automatic speech recognition
were similar to his opinions about MT.

And his 1969 letter to JASA,
expressing his personal opinion,
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.”

Tell us what you really think, John

"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 skeptical about the prospects.

At the same time,
many people believed that HLT was needed
and in principle was feasible.

1985: Should DARPA restart HLT?

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 1985 America,

no other sort of ASR program could have been gotten large-scale government funding.

NIST (Dave Pallett) 1985

"Performance Assessment of Automatic Speech Recognizers",
J. of Research of the National Bureau of Standards:

Definitive tests to fully characterize automatic speech recognizer or system performance cannot be specified at present. However, it is possible to design and conduct performance assessment tests that make use of widely available speech data bases, use test procedures similar to those used by others, and that are well documented. These tests provide valuable benchmark data and informative, though limited, predictive power. **By contrast, tests that make use of speech data bases that are not made available to others and for which the test procedures and results are poorly documented provide little objective information on system performance.**

“Common Task” structure

- A detailed “evaluation plan”
  - developed in consultation with researchers
  - and published as the first step in the project.
- Automatic evaluation software
  - 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 Piercians 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
   (because the project was glamour-and-deceit-proof)
   and to continue
   (because funders could measure progress over time)
2. Less obvious: it allowed project-internal hill climbing
   - because the evaluation metrics were automatic
   - and the evaluation code was public

   This obvious way of working was a new idea to many!
   … and researchers who had objected to be tested twice a year began testing themselves every hour…
3. Even less obvious: it created a culture
   (because researchers shared methods and results on shared data with a common metric)

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,
and has led to the flowering of "machine learning".

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 30 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, sentiment analysis, image analysis, video analysis, … , etc.

The general experience:

1. Error rates decline by a fixed percentage each year,
   to an asymptote depending on task and data quality
2. Progress usually comes from many small improvements;
   a change of 1% can be a reason to break out the champagne.
3. Shared data plays a crucial role – and is re-used in unexpected ways.
4. Glamour and deceit have mostly been avoided.

…and a self-sustaining process was started!

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.

A more recent sample

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

274 presentations –
Essentially 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.

Science is different…

But not that different.

Sharing data and problems

- lowers costs and barriers to entry
- creates intellectual communities
- speeds up replication and extension
- and guards against "glamour and deceit"
  (…as well as simple confusion)

A few initiatives (e.g. ADNI)
but NIH and NSF seem to be lagging

Thank you!