The Journal of Experimental Linguistics

Mark Liberman

University of Pennsylvania
myl@cis.upenn.edu

JEL: Overview

- What
- Why
- Who
- How
- When
- Whence
- Whither

- …discussion…

JEL: What?

- A journal of “reproducible research”
  - “reproducible computational experiments”
  - “executable articles”
  - Code to replicate all numbers, tables, figures from published data
- Open Access, web access only
  - Business model: “no one pays”
- LSA “co-journal” (eLanguage initiative)
- Articles, tutorials, squibs from **all language-related disciplines**

JEL: Why?

Goal: foster scientific communication in areas related to speech & language

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

-David Donoho, Stanford Statistics Dept.

Today, all science is computational science

JEL: Why?

- Benefits for authors:
  - Easier to continue old work
  - More influence on the field
- For readers:
  - Easier to check results
  - Extend work, transfer methods
- For language-related disciplines:
  - Faster diffusion of innovation
  - Avoid diffusion of nonsense
  - Speed the virtuous cycle of science

The factor of speed

- Scientific conversation is getting faster:
  - Journals: 1-3 years publication delay
  - Conference proceedings ~3 months
  - ePrint archives (arXiv): instantaneous
- More interactive:
  - Blogs, web forums
  - arXiv trackback feature
- JEL goals:
  - 4-6 weeks from submission to publication
  - Moderated forum for discussion

JEL: Who?

The editorial board, in alphabetical order:

Alan Black, Steven Bird, Harald Baayen, Paul Boersma, Tim Bunnell, Khalid Choukri, Christopher Cieri, John Coleman, Eric Fosler-Lussier, John Goldsmith, Jen Hay, Stephen Isard, Greg Kochanski, Lori Levin, Mark Liberman, Brian MacWhinney, Ani Nenkova, James Pennebaker, Stuart Shieber, Chilin Shih, David Talkin, Betty Tuller, and Jiahong Yuan.

JEL: How?

- Article texts – the usual process
- Data and scripts:
  - Checked
    - for re-creation of numbers, tables, figures
    - when run on independent system
  - Refereed
    - for clarity of code
    - for appropriateness and validity of methods
- No guarantee of long-term executability

JEL: When?

- Approved by LSA Exec 12/20/2009
- “Back end” mostly built, summer 2009
  - Facility for checking code/data
  - OJS site for online journal
- First submissions in process
- Goals
  - First articles online by 1/7/2010
  - Regular publication 1Q2010
- Process depends on volunteer labor
  (that’s a hint)

JEL: Whence?

- “Reproducible Research” movement
  - Geophysics
    - Jon Claerbout (1987++)
  - Signal processing
    - David Donoho, Jelena Kovacevic, Patrick Vandevelle
  - see:
    - reproducibleresearch.net (link)
    - Session on RR at Berlin 6 (link)

JEL: whence

- Doing it Old School: 19th c. philology
  - Claims about a body of shared data
  - Kudos for data curation and publication
  - Claims were explicit, checkable, extendable
- Classical corpus linguistics
  - Data was published
  - Programs were often shared
  - Claims were explicit, checkable, extendable

… a narrative interlude:

The DARPA “common task method”

A common theme

Nearly all modern research in the area of "Human Language Technologies"

(speech recognition, machine translation, information extraction from text, etc.)

is organized around stable, shared data and explicit, shared evaluation metrics.

Why is this?

The story starts in the 60's …

Pierce halts HLT funding

… with two bad reviews by John R. Pierce.

John Pierce was then the director of acoustics research at Bell Laboratories, and the man responsible for the development of communications satellites.

The topic of his first bad review was machine translation.

He chaired a committee whose 1966 report persuaded the U.S. government

to stop funding machine translation research.

The ALPAC report, 1966

Language and Machines: Computers in Translation and Linguistics [Report by the Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences, 1966]:

"Unedited machine output from scientific text is decipherable for the most part, but it is sometimes misleading and sometimes wrong …, and it makes slow and painful reading."

In other words, MT was not much good without post-editing -- and post-editing was just about as expensive as plain translation.

ALPAC MT Samples (1966)

Automatic translations, from 3 different systems, of a Russian article:

Biological experiments, conducted on different space aircraft/vehicles, astrophysical space research and flights of Soviet and American astronauts with/from sufficient convincingness showed that short-term orbital flights lower than radiation belts of earth in the absence of heightened solar activity in radiation ratio are safe.

Biological experiments, conducted on various/different cosmic aircraft, astrophysical researches of the cosmic space and flights of Soviet and American astronauts with the sufficient/rather persuasiveness showed/ indicated/pointed, that momentary/transitory/short orbital flights of lower/below than radiation belts/regions/flanges of earth/land/soil in the absence of the raised/increased/hightened sun/sunny/solar activity with respect to radiation are/appear/arrive/report safe/not dangerous/secure.

Biological experiments, which were conducted on different cosmic LETATEL6NYX APPARATI, the astrophysical investigations of cosmic space and the flights of Soviet and also American KOSMONAVTOV with the sufficient convincingness showed, that the short-term orbital flights of below radiation belts of ground upon the absence of the increased solar activity in radiation relation are safe.

ALPAC Conclusions

"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.

Pierce put his faith in science rather than engineering:

"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.

There is every reason to believe that facing up to this challenge will ultimately lead to important contributions in many fields."

Plus ça change…

Unfortunately, it's still true that "unedited machine output … is decipherable for the most part, but it is sometimes misleading and sometimes wrong …, and it makes slow and painful reading."

Here are 3 Chinese-English systems from a 2008 evaluation:

The new web November 23 the CPC Central Committee recently stepped up exchanges of Provincial Commission for Discipline Inspection of the CPC Central Committee, secretary of the Central Commission for Discipline Inspection, up to now, there have been six provinces of the adjustment.

Chinanews.com, November 23 - Recently intensified exchanges of secretary of the Provincial Discipline Inspection Commission of the Central Committee of the Communist Party of China (CPC) and secretary of the discipline inspection commission, so far, six provinces have.

Chinanews.com, November 23 (Xinhua) -- The CPC Central Committee recently of the provincial discipline inspection commission, as secretary of the discipline inspection commission so far, there have been six of the adjustment.

But wait…

…hasn't machine translation gotten better since 1966?

It certainly has.

But the samples given to Pierce's committee in 1966
were among the better outputs.
And the samples that I selected from the 2008 systems
were among their less good results.

And Chinese is harder than Russian.

The trouble is,
evaluation by example is not a reliable method.

The second bad review

Three years later,
John Pierce ended U.S. funding for speech recognition with a stinging letter to the Acoustical Society

J. R. Pierce, "Whither Speech Recognition?", *Letter to the Editor of JASA*, 1969:

"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."

Pierce pushes science again

The key problem, Pierce thought,
was failure to build on past accomplishments
in the way that successful science and engineering do:

"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."

HLT reborn

1986 -- CONTROVERSY: should DARPA start HLT research again?

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 limited and objective evaluation metric
  - applied by a neutral agent (NIST); and
- ensures that "simple, clear, sure knowledge is gained"
  - because participants must reveal their methods
  - to the sponsor and to one another
  - at the same time that the evaluation metric is applied.

In 1986 America,
no other sort of ASR program could have been funded.

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

   (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.

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 20 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, optical character recognition, … , 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,
   but artificiality remains a concern.
4. Interaction with speech and language science has been small.

… end of narrative interlude …

JEL: Whither?

- For JEL to work…

We need help!

- So submit! (articles, that is…)
  - Regular articles
  - “How we did it” supplements
  - Tutorials
  - Squibs
  - Reviews

JEL: Whither?

- Also needed: help with reviewing:
  - Fast publication requires fast refereeing
  - Code/data referees are a special need
- And help with “back end”:
  - graphic design
  - OJS hacking
  - ODT hacking
  - Proofreading
  - etc., etc.

JEL: Whither?

This is an experiment

in the future of scientific communication.

It can only work

if you use it in your work

and support it with your work.

So please join us

in exploring the possibilities.

Thank you!