Principles for Robust Evaluation Infrastructure

Justin Zobel

with Alistair Moffat, Mark Sanderson, & William Webber
Melbourne, Australia

Position

Critical elements:

- Shared testing environments, such as TREC's test collections.
- Resources such as shared, public-domain IR systems.

But other elements are also critical:

- Environments for publishing new data, runs, and systems;
- Shared, statistically based tools for measuring and recording experimental outcomes;
- Social frameworks that make openness the norm; and
- Provision of mechanisms by which restricted or private data can be evaluated, accessed, or inspected.

Score standardization

The problem: different topics have different score distributions (measured by mean and standard deviation).

The solution: adjust each topic's scores so that they have the same score distributions (measured by mean and standard deviation).

$$z = \frac{r - m}{s}$$

m	Mean score of (reference) systems against topic.
s	Score standard deviation of systems against topic.
r	Raw (unstandardized) score of a system against this topic.
z	Standardized score of that system against this topic.

Every topic has mean standardized score (on reference systems) of 0, and s.d. of 1.

Historical standardization

We used standardization to see if we could examine how TREC systems were changing over time.

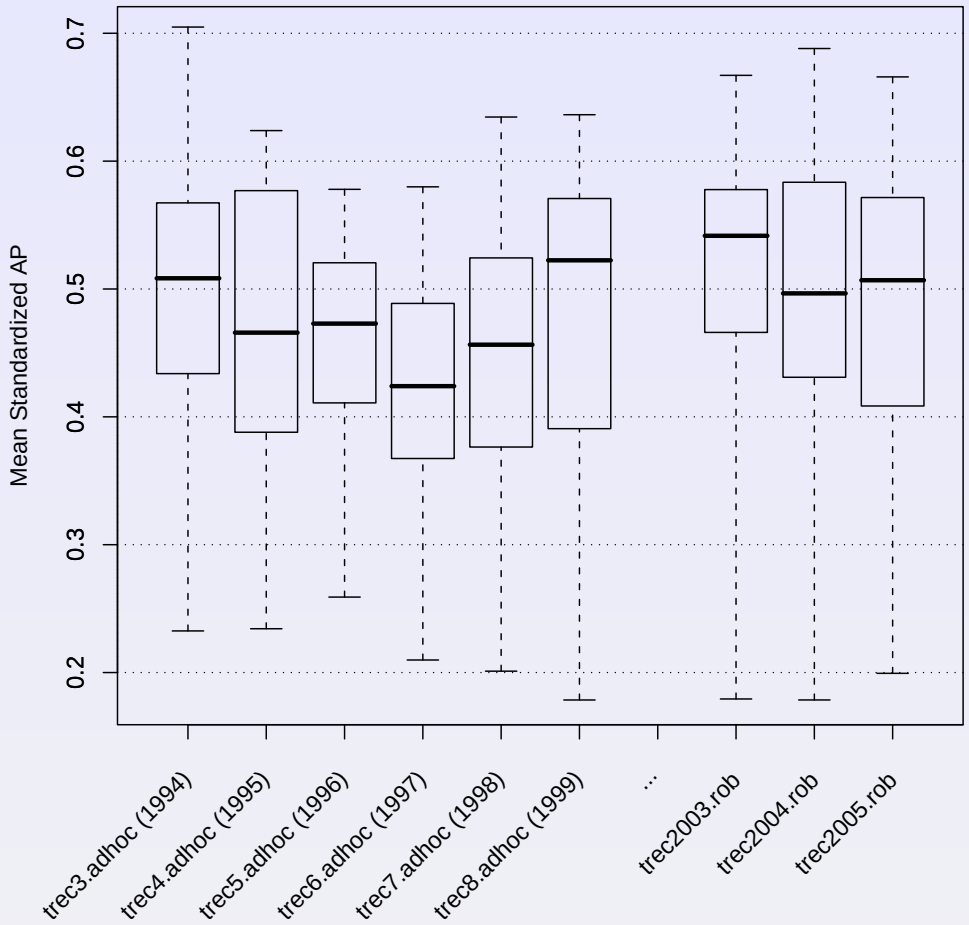
The reference set was a collection of public-domain search engines.

- Within these, we varied use of features such as query expansion to create additional, but non-independent, references.

With the reference set as benchmarks, we could then plot the effectiveness of the original TREC systems, using the original runs available from NIST.

Expectation: current public-domain systems would be much better than archaic (e.g., 1993) systems, and embody the best of recent innovations.

Historical standardization



- No overall improvement, perhaps since TREC 3 (1994).

Survey

We surveyed results published at SIGIR (1998–2008) and CIKM (2004–2008).

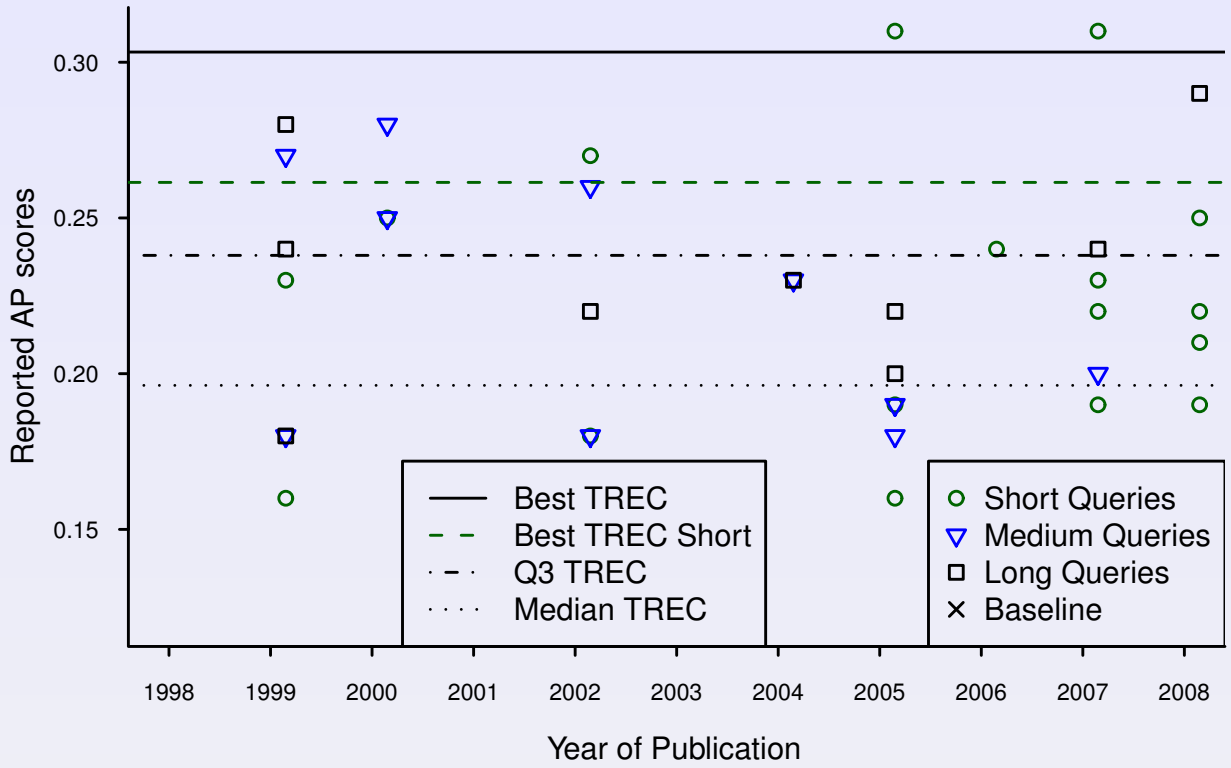
- All uses of a TREC collection (AdHoc, Terabyte, Web, Robust); a total of 106 papers and posters.
- Captured AP scores of improved and baseline systems.

Expectation: of an upward trend in reported effectiveness over time.

- Arrange reported scores by collection.
- Within collection, arrange by year of publication.

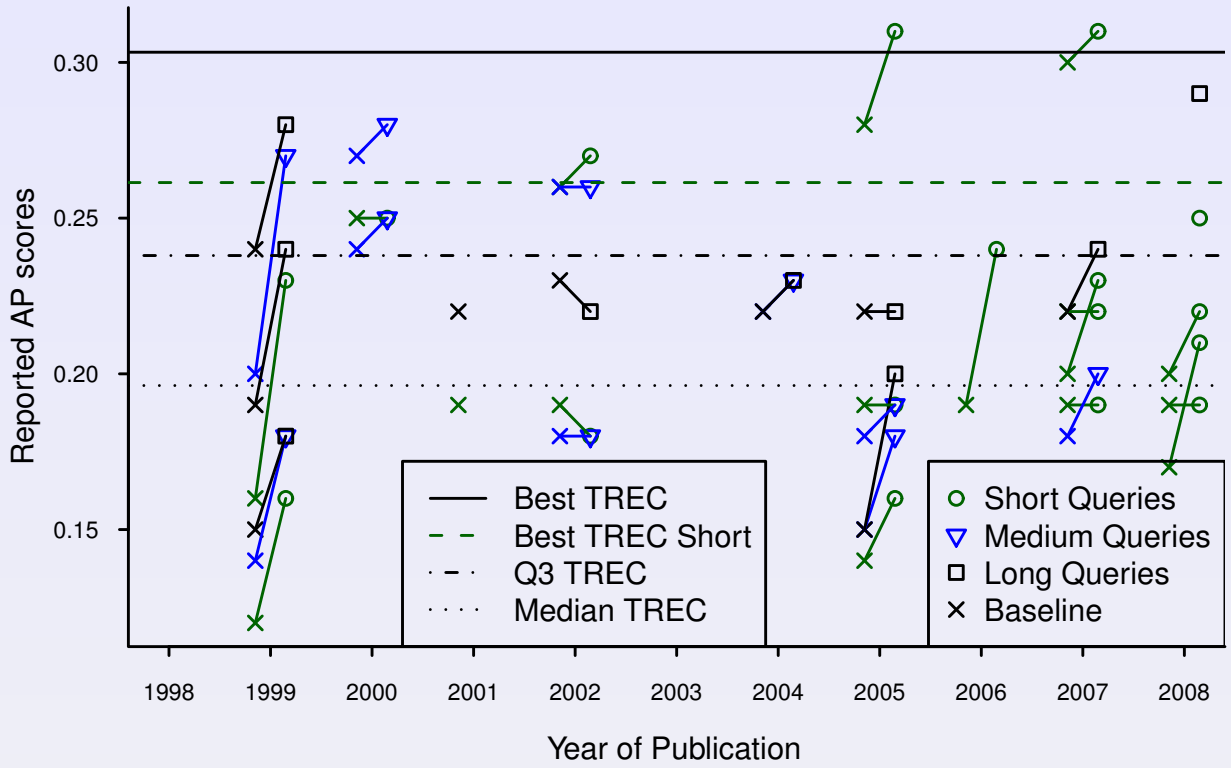
(Note: authors report average scores but don't publish runs or per-query scores, so standardization isn't an option.)

‘Improved’ scores over time



- No upwards trend in scores over time.
- Almost no results beat the best original TREC system.
- The median ‘improved’ score (0.225) is in the 58th percentile of original TREC systems.

Use of baselines



- Most baselines are below the median TREC system (the median baseline is at the 41st TREC percentile).
- Baselines do not improve over time
- . . . one year's improved result is not next year's baseline.

Position

Critical elements:

- Shared testing environments, such as TREC's test collections.
- Resources such as shared, public-domain IR systems.

But other elements are also critical:

- Environments for publishing new data, runs, and systems;
- Shared, statistically based tools for measuring and recording experimental outcomes;
- Social frameworks that make openness the norm; and
- Provision of mechanisms by which restricted or private data can be evaluated, accessed, or inspected.