Datasets and Benchmarks for AI

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

In October, 2016 the National Science and Technology Council's Networking and Information Technology Research and Development (NITRD) Subcommittee released the National Artificial Intelligence Research and Development Strategic Plan [1]. The plan articulates priorities for federally-funded research in AI organized along seven strategies. Two of the strategies focus on the need to develop infrastructural support for testing AI systems and techniques. The executive summary of the plan describes these two strategies as thus:

Strategy 5: Develop shared public datasets and environments for AI training and testing. The depth, quality, and accuracy of training datasets and resources significantly affect AI performance. Researchers need to develop high quality datasets and environments and enable responsible access to high-quality datasets as well as to testing and training resources.

Strategy 6: Measure and evaluate AI technologies through standards and benchmarks. Essential to advancements in AI are standards, benchmarks, testbeds, and community engagement that guide and evaluate progress in AI. Additional research is needed to develop a broad spectrum of evaluative techniques.

The National Institute of Standards and Technology (NIST) is the home of community evaluation programs that create datasets and benchmarks for a variety of tasks. One example is the Text REtrieval Conference (TREC)[1] program, which builds the infrastructure required for large-scale testing of information access systems such as search engines and question answering systems.

TREC began in 1992 and as run annually since then. Each year is organized around a set of challenge problems called "tracks". The set of tracks change from year to year to keep TREC fresh, though individual tracks generally run for at least a few years. For each track, TREC makes a dataset available to participants and publishes guidelines that define the parameters of the task. Participants perform the task and submit the results to NIST where human judges assess the quality of the results. NIST computes evaluation scores for the results using the human judgments. At the end of a cycle, participants gather at NIST to discuss the overall findings, to exchange research results, and to refine the research methodology.

This paradigm of individual experiments evaluated on a common task has proved to be highly successful. System effectiveness has improved, new research areas have been supported, and research ideas have flowed freely across different participant teams [3]. A large part of the success has resulted from defining carefully-calibrated evaluation tasks: abstract tasks that are general enough to be widely applicable and amenable to experimental control, while also realistic enough to be informative. Sparck Jones [2] calls such an abstraction a core competency.

[1] `http://trec.nist.gov/`

The core competencies tested in TREC tracks and other similar challenges have focused on effectiveness or accuracy divorced from such questions as fairness and trustworthiness. Accuracy is conceptually simple to operationalize, fairness and trustworthiness much less so. Is there a testable abstract task for trustworthiness? for fairness? What impact does domain or application area have? How is the impact of different training scenarios on a given system assessed and documented? These are fundamental questions to be answered to create a vibrant program targeting infrastructural support for trusted AI.

References

[1] National Science and Technology Council, Networking and Information Technology Research and Development (NITRD) Subcommittee. The National Artificial Intelligence Research and Development Strategic Plan. `https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf`, 2016.

[2] Karen Sparck Jones. Automatic language and information processing: Rethinking evaluation. *Natural Language Engineering*, 7(1):29–46, 2001.

[3] Ellen M. Voorhees, Paul Over, and Ian Soboroff. Building better search engines by measuring search quality. *IEEE IT Professional*, 16(2):22–20, March/April 2014.