The Role of Test and Evaluation in Social and Behavioral Sciences

Michael Maxwell, Ph.D.[1]
Center for Advanced Study of Language, University of Maryland

Introduction

This paper emphasizes the importance of Test and Evaluation (T&E, also known as Independent Verification and Validation, or IVV) in USG-supported research on social and behavioral sciences.[2] We argue that language-related research provides a model of T&E for other social and behavioral sciences.

Why language?

Why are we using language-related research as a model for social and behavioral sciences?

First, the study of language *is* a social and behavioral science, and is classed as such by the National Science Foundation (cf. also Larson 2010).

Second, language and linguistics has played an indisputably important role in American defense and intelligence, with impacts both to the IC/ DoD, and to academia. Many of the resources available for mid-sized languages such as Pashto, Bangla, and various "dialects" of Arabic, were developed with IC or DoD funding.

Third, it is relatively easy to point to significant successes in the growth of language technologies, especially in the past twenty years: machine translation, speech-to-text, and keyword search of the internet are obvious examples.

In sum, we believe that languages and linguistics (including computational linguistics) constitute a success story for how social and behavioral sciences can serve the IC and DoD. That said, progress has not always been linear, and there are lessons to learn from both failure and success. This white paper concentrates on one of those lessons: the need for language experts, linguists, and computational linguists to be involved in Test and Evaluation (T&E) teams of research projects.

The Role of Test and Evaluation in Language-related Research

We base our discussion largely on three IARPA projects for which CASL has provided T&E expertise:

- **Babel** (https://www.iarpa.gov/index.php/research-programs/babel), which developed search technology for spoken language. This project was started by Dr. Mary Harper, and completed under Dr. Carl Rubino in 2016.
- **Mercury** (https://www.iarpa.gov/index.php/research-programs/mercury), an on-going project in the prediction of significant events (political crises, disease outbreaks, terrorist activity, and military actions) based on SIGINT inputs. The Program Manager is Dr. Kristen Jordan.
- **Material** (https://www.iarpa.gov/index.php/research-programs/material), an on-going project in cross-language information retrieval (CLIR), with inputs and outputs in English. The Program Manager is Dr. Carl Rubino.

This white paper also draws on experience in DARPA-funded projects, including:

---

[1] Corresponding author: mmaxwell@casl.umd.edu.

[2] Chapter 8 of the Defense Acquisition Guidebook (https://www.dau.mil/tools/dag) addresses T&E; the responsibility for T&E falls under the Deputy Assistant Secretary of Defense for Developmental Test & Evaluation/ Director, Test Resource Management Center.

- **TIDES**, a project headed by Dr. Charles Wayne in the early 2000s, with the goal of building information extraction and summarization systems for foreign languages.
- **LORELEI** (http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents), a project developing methods for rapid ramp-up of technology to provide situational awareness in low resource languages during crisis situations. Dr. Boyan Onyshkevych is the Program Manager.

The following sub-sections discuss specific roles played by T&E in these projects.

The Role of T&E in Language Selection

For most IC applications, a goal is to develop Human Language Technology (HLT) methods that will work across a variety of languages and writing systems. But a project can afford to use as test cases only a limited number of languages. In the Babel project, The IARPA Program Manager tasked CASL to advise on the choice of 26 languages[3], ensuring that the sample was representative of a diverse language typology, to include both tone and non-tone languages, languages with inflectional morphology ranging from none to those with complex affixation, and languages with different kinds of syntax and writing systems.

The Role of T&E in Data QC

Data provided to performers, or used in evaluation, needs to be Quality Controlled. QC adds expense, both in the time needed for QC, and in the time needed to repair faulty data. Nevertheless, it is crucial: machine learning can easily go astray when its input data is biased, misleading or noisy (Angwin et al 2016). In IARPA Babel, linguists on the T&E team helped develop new computational methods to ensure the accuracy of speech transcription. Similarly, IARPA MATERIAL will use content and topic questions and answers created by annotators. First efforts at ensuring consistency among annotators were judged not good enough, and the T&E team developed new solutions.

The Role of T&E in Language Data Normalization

Data normalization is a method for making data internally consistent. While related to Data QC, normalization seeks to bring raw unannotated data into a consistent format. (The same methods are also applied later to real-world use cases.)

Some methods of data normalization rely on standards, such as Unicode Normalization (Unicode Consortium 2016, see also http://unicode.org/reports/tr15/), but language-specific complex normalization methods may also be needed. For example, while Pashto orthographic standards are reasonably specific, the usage of certain letters of the Pashto alphabet shows little consistency in real-world texts. Pashto normalization therefore often maps this set of letters to a single letter. While this removes noise, it throws away potentially relevant information. Members of the T&E team who are knowledgeable in the language, language computing, and the goals of the project must therefore weigh the gain against the loss.

The Role of T&E in Evaluation

As the term 'Test and Evaluation' would suggest, T&E teams devote a great deal of time to issues of evaluation, since performers may be de-selected on this basis, and even the continuation of the program can be at stake.

Evaluation data may be of the same type as was provided to performers as training data, but with annotations held back for grading; QC and data normalization ensure accuracy and fairness in this case. Alternatively, performers may be evaluated on different kinds of data. In the IARPA Mercury project, for example, performers train their prediction systems on classified and unclassified indicators, but their

[3] Of which 25 were used.

systems are evaluated against reports of actual events. It is therefore incumbent on the Mercury T&E team to extract accurate event records from those reports, a process still undergoing improvement.

The Role of T&E in Transition

Assuming a project is successful, its results must be transitioned to the IC or DoD. In many cases, transitioning is the responsibility of the T&E team.

Transition of research into operations has a long history, and is to some extent a well understood problem with standardized methodologies. However, there is one aspect of transitioning projects that rely on AI (machine learning) that is rather different from most earlier tech transitions: namely, AI systems are often black boxes, making the reasons for their results hard to understand. Users may therefore be reluctant to take the AI system's advice without justification, which impacts human-in-the-loop systems (Castelvecchi 2016). Worse, bias can be inadvertently introduced in the training data (Anguin 2016), and such a bias may only become apparent in certain situations—potentially only in crises.

Finding ways to make a research system more transparent to potential users, or using simulations to test for biases, would be a natural task for a T&E team. The DARPA project in Explainable Artificial Intelligence, http://www.darpa.mil/program/explainable-artificial-intelligence, is a research effort aimed at this problem. Cross-fertilization among such DARPA and IARPA projects should be encouraged, perhaps including cross-teaming.

The Role of T&E in Post-Analysis

In the Babel project, performance[4] varied widely among the 25 languages. While this was expected, there was no obvious correlation with typological qualities (such as morphology complexity). If the factors making languages "easy" or "hard" for Babel systems were better understood, it would be easier to transition the research into operation, e.g. to predict the costs for a new language. Because the Babel project did not have funding for more than very preliminary research on this question, an important opportunity was unfortunately missed.

The Role of T&E in Data Retention

Good data is expensive, as noted in our earlier comments about data QC. Fortunately, the uses of data need not end with the end of the project. It is therefore incumbent on IC- and DoD-sponsored projects to make provision for data retention and dissemination after project completion. Not only might the data enable future researchers to come up with better solutions than the performer teams did during the original project, it can be useful to researchers in other domains. Data cataloging (e.g. through the Government Catalog of Language Resources, GCLR), archiving and dissemination are all necessary parts.[5] Since institutions usually survive longer than people hold individual roles, and since DARPA and IARPA Program Managers in particular quickly move on to other positions, it may be advisable to fund T&E teams (who will already be familiar with the data) to carry out this work.

The Composition of T&E Teams

The T&E teams that we have been involved in included representatives from NIST, one or more government labs, and a UARC (ourselves). We assume that the roles of NIST and the government labs are clear, and will therefore focus on the role of the University Affiliated Research Center (UARC).

---

[4] As measured by the Actual Term Weighted Value (ATWV) metric.

[5] There are of course considerations that could prevent wider dissemination, such as IRB and classification; but the default should be open dissemination.

Because of its university affiliation, a UARC is able to draw not only on its own staff, but also on staff and faculty of the associated university. This academic connection also makes it easier for the UARC to interact with researchers in other institutions, both domestic and overseas--the latter is especially relevant in the context of language research, since foreign languages are by definition spoken mostly in foreign countries.

To the extent that a UARC has cleared personnel among its members, adjuncts and affiliates, the UARC can also serve as a mediator/ interpreter between the cultures of the IC and DOD on the one hand, and uncleared academics on the other.

Finally, a UARC has trusted agent status with the USG, enabling a relationship with an academic institution that is not otherwise possible.

Recommendations

In summary, we provide the following recommendations:

- The role of T&E teams in research projects aimed at producing results and technologies useful to the IC, and to the DOD in general, is clear and (as far as we know) undisputed. First and foremost, then, we urge that this role be continued.
- T&E teams have often been composed of some combination of government labs, NIST, and UARCs, functioning under the leadership of a DARPA or IARPA Program Manager. These three kinds of organizations play unique roles; we believe that it should be policy for T&E teams include members from each of these components.
- The role of T&E teams in post-project analysis should be expanded to address questions left unanswered during the fast-paced period of performance.
- The role of T&E teams in Tech Transition and in Data Retention should be expanded.
- There is on-going research into making AI/ Machine Learning systems more transparent in operational settings. This research should be incorporated into projects that use machine learning, a task which could be furthered by T&E teams.
- Finally, there are many things that have been done right in government-supported language research, and with the T&E teams that help guide that research. We advocate that this serve as a model for other kinds of social and behavioral research conducted under USG funding.

References

Anguin, Julia; Jeff Larson, Sury Mattu and Lauren Kirchner. 2016. “Machine Bias.” https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Castelvecchi, Davide. 2016. “Can we open the black box of AI?” Nature 538: 20-23. Available at http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

Larson, Richard K. 2010. *Grammar as Science*. Cambridge, MA: MIT Press.

The Unicode Consortium. 2016. *The Unicode Standard, Version 9.0.0*. Mountain View, CA: The Unicode Consortium. http://www.unicode.org/versions/Unicode9.0.0/.