NIST Language Technology Evaluation Cookbook

Alvin F. Martin, John S. Garofolo,
Jonathan C. Fiscus, Audrey N. Le, David S. Pallett, Mark A. Przybocki, Gregory A. Sanders

National Institute of Standards and Technology, Gaithersburg, MD, USA
{alvin.martin, john.garofolo, jonathan.fiscus, audrey.le, david.pallett, mark.przybocki, gregory.sanders} @nist.gov

Abstract

We review some of the methodology applied to the various NIST language technology evaluations. We discuss the elements included in each evaluation plan, and suggest what we believe are key practices for successful evaluations, and what pitfalls should be avoided. A couple of lessons learned are noted.

1. Introduction

The National Institute of Standards and Technology has been designing and administering community-wide formal evaluations of language technologies since 1987. Although the evaluated technologies differ considerably, a number of common elements must be addressed in each successful evaluation. Thus, over time, NIST has developed a methodology for implementing this type of evaluation. The common elements include clear specification of the research task(s); definition of informative evaluation metric(s); creation of publicly available scoring software useful for both developmental and summative evaluation; properly scoped test corpora for the training, development, and evaluation phases; clear-cut evaluation rules and protocols; concise system description requirements; simple yet expressive system output submission formats; and realistic schedules. In most NIST evaluations, these elements are documented in a detailed evaluation plan, which is made publicly available for all potential evaluation participants.

This paper will address the development of each of these elements and suggest some insight into what makes for a successful evaluation, and what pitfalls are to be avoided. This is our first effort at communicating this methodology in a formal document to the language research community.

2. Task Definition

One or several basic tasks are specified for each evaluation in its evaluation plan. NIST evaluations have focused on the core technologies to be developed rather than on specific applications that may be of more interest to some system developers than to others, with emphasis on technical capabilities beyond what is attempted in the contemporary commercial marketplace. To the extent that application specific decisions are required, the interests of the government sponsoring agencies have received priority.

Thus, in speaker recognition the NIST evaluations have focused on text-independent recognition using conversational telephone speech. This is of interest to the program sponsors, and avoids choosing text-dependent scenarios requiring decisions on the word sequences to be spoken by each participant, decisions which tend to be rather application specific. A wide range of evaluation participants have participated as a result, though there has perhaps been diminished interest by commercial vendors oriented to particular applications.

Evaluation	Domain	Tasks
Rich Transcription [1], [2]	BNews, CTS, MR	Word recognition Metadata detection Speaker segmentation
Conversational Telephone Recognition [1]	CTS	Word recognition
Broadcast News Recognition [1]	BNews	Word recognition
Topic Detection and Tracking (TDT) [3]	BNews, NWire	Story segmentation Topic tracking Topic detection First story detection Link detection
Machine Translation [4]	NWire	English translation
Language Recognition [5]	CTS	Language detection
Speaker Recognition [6]	CTS, Forensic	Speaker detection Multi-speaker detection Speaker tracking Speaker segmentation
Automatic Content Extraction (ACE) [7]	BNews, NWire, NPaper	Entity detection Event detection
Spoken Document Retrieval (SDR) [8]	BNews	Information retrieval Speech recognition

Table 1: *NIST Evaluation Tasks and Domains. Domain abbreviations: BNews = Broadcast News, CTS = Conversational Telephone Speech, MR = Meeting Room data, NPaper = Newspaper, NWire = Newswire*

In speech recognition, the NIST evaluations have long emphasized speaker independent recognition, even when most commercial systems supported only speaker dependent recognition. In recent years the evaluations have focused on the challenging domains of broadcast news and conversational telephone speech. Systems developed in these evaluations have become the basis of later commercially successful systems.

Table 1 summarizes the basic tasks of various NIST evaluations.

3. Metric(s)

For each evaluation task a primary evaluation metric is specified. This allows participants to focus their approaches to the evaluation tasks, knowing how evaluation scoring will be carried out and presented by NIST. Because participants will orient their efforts toward performing well on the metric specified, it is important that it should accurately reflect system capability for the task.

Choosing such an appropriate metric can be a challenge, and a subject of considerable discussion beforehand involving NIST, government sponsors, and the evaluation participants. If the metric can be specified long before the actual evaluation takes place, then participating sites are able to use it in their development work in preparation for the evaluation. This is very desirable in vetting the metric.

An intuitively meaningful metric that gets at the essence of the technical challenge is most desirable, though this is perhaps more achievable in some areas than others. The word error rate metric for automatic speech recognition (speech-to-text) is notable for meeting these criteria, being simple in concept (easily understood by managers) but accepted as meaningful by most researchers. It provides a simple way of classifying errors into the intuitively meaningful categories of substitutions (incorrect words), deletions (words actually spoken but omitted), and insertions (words reported though not spoken). The word error rate is then the total number of errors divided by the number of words actually spoken.

For detection tasks (speaker and language detection and TDT), system decisions may result in two types of errors, often denoted as misses and false alarms. NIST has generally used a specific linear combination of the two error rates as its primary evaluation metric in such evaluations. In addition to individual hard decisions for each trial, sites are required to provide likelihood scores that indicate the relative degree of confidence in each trial decision. These scores support the determination of the range of possible system operating points, as discussed further in the next section.

While there should be a primary metric specified for each task, additional metrics may also be defined which indicate other conditions or aspects of performance of interest to the sponsors and which will be presented in NIST's analysis of evaluation results. A proliferation of metrics is best avoided, however, as it dulls the research focus and makes the results harder to interpret.

4. Scoring and Analysis Software

NIST provides on its website [9] scoring and analysis software for each evaluation task. The public availability of the software assures that all participants, or potential participants, understand exactly what is entailed in the evaluation tasks.

The software implements the specified scoring metric(s), but often goes considerably beyond this. Various options are provided to give more or less detailed scoring information or to include or not include additional metrics or scoring options.

For detection tasks NIST provides software to support the display of graphs that show the range of possible operating points, based on the likelihood scores provided by systems for all trials. NIST calls these Detection Error Tradeoff (DET) Curves [10], a variation on traditional Receiver Operating Characteristic (ROC) Curves [11]. In DET Curves both axes are plotted on a normal deviate scale. This produces linear curves when the underlying likelihood score distributions are normal. Figure 1 shows a sample DET plot.

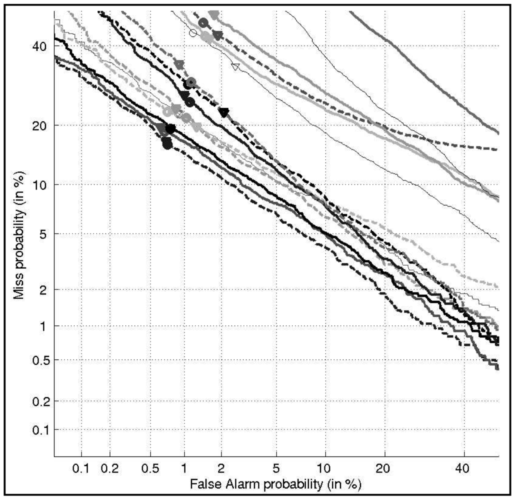

Figure 1*: DET plot of systems participating in the 2003 NIST Speaker Recognition Evaluation*

One important type of analysis of performance results of ongoing interest involves determining whether differences in performance between evaluation systems are significant in a statistical sense. Such tests determine whether or not, at a given confidence level (generally 95% or higher), one can reject the null hypothesis that observed differences between systems may be attributed to random variations. NIST has developed a suite of four types of paired comparison tests for the significance of observed performance differences between two evaluation systems. The suite was designed particularly for use in speech transcription evaluations, but has been applied to some other evaluations as well. These tests are described in detail on the NIST web site [12].

5. Data: Training, Development, Evaluation

It is well understood that evaluations are largely data driven. The key requirement, and cost, for a successful evaluation, is the creation of appropriate data sets to support the evaluation and the development effort leading up to it. NIST works closely with sponsors and the Linguistic Data Consortium (LDC) to arrange for the creation and annotation of appropriate speech and text data in a timely fashion for the evaluations it coordinates.

Because data is an expensive and scarce resource, and in some cases is subject to intellectual property restrictions, the right to possess and process data for research purposes is sometimes an issue. NIST arranges with the LDC to make evaluation data available to all participant and prospective participant sites for research and development purposes during the period leading up to and following each evaluation. This sometimes requires that non-members of the LDC sign an appropriate license agreement governing the use of the data involved. It should also be noted that the LDC has a general policy both of respecting all intellectual property rights for the data it distributes and of assuring that serious researchers should have access to needed data whatever their financial circumstances.

The evaluation plan specifies the data sources that may be used in the course of the evaluation cycle. This includes the data sources available for system training, the development test set on which systems may be repeatedly tested and tuned by the developer, and the evaluation test set on which systems are to be run once and then scored by NIST. Each data type needs to be made available to evaluation participants in a timely manner. Adherence to the collection and annotation schedules may be vital to evaluation success.

Training data should be available early to the research sites and should be abundant. The required transcriptions or annotations for the training data may be of lower quality than for the development and evaluation test data. Recent experiments involving conversational data for speech recognition have shown that a rapid but less accurate transcription process can produce training data that leads to little degradation in the resulting systems. And other work has shown that even automatically transcribed data can be of some value for training.

Development and evaluation test data need to have high quality transcription or annotation. Development test data is also needed early on in the evaluation process, and in a quantity and scope comparable to that of the evaluation data. This allows both participants and sponsors to know in advance the kind of results to be expected, and to anticipate possible problems with the annotation or evaluation procedures.

The early availability of both training and development test data allows effective research efforts by the sites to improve their algorithms in the period leading up to the actual evaluation. This is the period of time during which the greatest research gains are likely to be made.

For an ongoing series of evaluations, the evaluation data of one cycle will normally become the development test data of the next cycle. For evaluations in new domains or one-time evaluations, a “dry run” using development test data enhances confidence in the evaluation process and permits bugs in the evaluation infrastructure to be identified and removed.

At the conclusion of an evaluation, the community has results for two data sets, from both the development and the evaluation test data. This provides some insight into performance variability across data sets. Quantifying and (preferably) controlling such variability is a major concern in conducting an ongoing series of evaluations.

6. Evaluation Rules and Protocols

Issues may arise about how participants develop their evaluation systems and interact with the data that is provided. Appropriate specification of the evaluation rules and protocols in the evaluation plan helps to assure that all sites understand the procedures to be followed, and that none inadvertently (or intentionally) receives an unfair advantage.

One recurring issue is whether sites may utilize training data other than that which is provided and specified. For speech recognition of broadcast news type data it has been thought important not to allow training on data that is contemporaneous with that used for evaluation, as it provides an unrealistic advantage in the creation of language models (the news will not be new to the system). This needs to be carefully specified. More generally, there is an issue here of the relative merits of maintaining rules that make the playing field as level as possible versus encouraging the development of the best possible systems. Different NIST evaluations in the past have decided this issue in different ways, but it is important that the decisions be carefully considered by the sponsors, perhaps with input from the participants, and that the rules finally adopted be specified in the evaluation plan.

7. System Descriptions

A clear concise description of each system run in an evaluation is very useful. It provides the sponsor and the researchers with a concise overview of how each system was constructed and what makes it different from other participating systems. This information can help in determining why one system might have out-performed another. It is especially useful when comparing results from “contrast” tests for the same system using different parameters. A system description should not be a technical report but, rather, a brief abstract describing the pertinent features of the system. Parameter settings used for contrastive tests should be highlighted.

Generally, a template is provided to the community so that the descriptions are reasonably uniform across sites. This template usually includes: 1) site identification, 2) contact information, 3) a unique system ID for each evaluation system 4) specification of the training data and/or rule sets employed, 5), system execution time (as a multiple of real-time), and 6) any pertinent links or references to more detailed information.

8. Submission Format and Directions

The submission directions give detailed instructions, in terms of file names, file formats, and line-by-line requirements, for the results that sites must provide to NIST. This detailed specification helps to assure that the sites understand the evaluation tasks and that the scoring software will perform as expected. Initial versions of evaluation plans may leave parts of these directions to be filled in later.

9. Schedule

A complete schedule, which generally forms the concluding section of each evaluation plan, plays a vital role in keeping an evaluation on track. It may begin with the date for release of the evaluation plan and conclude with the dates for the evaluation workshop. The dates for NIST to release the different types of data to the sites, and for the sites to submit their results to NIST are included. Providing an evaluation plan with such a calendar as early as possible helps to lock in the task definitions, to put all parties on equal footing with regard to the evaluation, and to avoid schedule slips and unexpected surprises.

10. Workshop

NIST, together with the program sponsors, organizes a workshop following each evaluation. At the workshop NIST presents the evaluation performance results for all participating sites and analyses of various factors affecting performance. Each participating site is expected to have a representative at the workshop, who gives a technical presentation on the site's system(s) used in the evaluation.

NIST has regarded site participation in such a workshop as vital to the mission of each evaluation, even though some organizations may choose not to take part in the evaluation because they wish to avoid potential embarrassment among research peers. This makes it important that sites have access to detailed evaluation plans and data of the type to be used before deciding to participate in an evaluation.

The NIST evaluation workshops have generally been open only to representatives of participating sites and government organizations sponsoring or having an interest in the evaluation, along with organizations, such as the Linguistic Data Consortium, providing data or other resources that support the evaluation. Some feel that this has made evaluations less visible in the wider research community but, on the other hand, it is reasonably argued that smaller workshops promote greater technical interchange.

Differing policies have been followed in different areas in terms of making the workshop presentations publicly available after the workshop. Speech recognition proceedings have been made available (some are on the NIST website), but for speaker recognition the specific performance results of named individual sites have not been disclosed. This is an issue about which there is some ongoing discussion. Developers of successful systems are likely subsequently to publish some information on their work for the wider research community.

11. Lessons Learned

We conclude with a couple of observations from the experience of numerous evaluations.

As noted, the variability of test data sets is a major concern for ongoing series of evaluations. Running a control system, perhaps provided by a participant, on succeeding data sets in advance can help to limit such variability and to anticipate its extent. In the recent Rich Transcription speech recognition evaluations a further strategy has been adopted; NIST created a static "Progress Test". This is a fixed evaluation data set for sites to run their systems on each year. After the run the data is returned to NIST, and is not otherwise available to participants or anyone else. Thus the year-to-year results on it form a firm benchmark for measuring year-to-year performance progress. This Progress Test set is used in addition to the regular annual test sets that become development test data for each successive evaluation. It remains to be seen how acceptable such a fixed test set will come to be viewed in the community.

NIST and its sponsors have primary responsibility for creating evaluation plans. But it is of crucial importance that the relevant research community "buys-in" to the plans that are presented. It is important that evaluations address problems that are of interest to outside research organizations, that they complement ongoing research and development efforts, and that their size and scope, along with the tools provided to support them, make them doable by research organizations of limited size and budgets. Teaming is also to be encouraged. Thus it is important to get to know the key players in the relevant research community and to have appropriate preliminary discussions about their interests and capabilities before evaluations are planned and announced. NIST evaluations are open to all interested participants, and the degree of community response is an important indicator of an evaluation's impact and success.

12. References

[1] Pallett, D., "A Look at NIST's Benchmark ASR Tests: Past, Present, and Future", *Proc 2003 IEEE Workshop on Automatic Speech Recognition and Understanding*

[2] Garofolo, J., et al., "The NIST Meeting Room Pilot Corpus", *Proc. LREC* (2003)

[3] Fiscus, J., "Results of the 2003 Topic Detection and Tracking Evaluation", *Proc. LREC* (2003)

[4] Doddington, G., "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, *Proc. HLT* (2002)

[5] Martin, A. and Przybocki, M., "NIST 2003 Language Recognition Evaluation", *Proc. Eurospeech* (2003)

[6] Martin, A. and Przybocki, M., "The NIST Speaker Recognition Evaluations: 1996-2001", *Proc. 2001: A Speaker Odyssey* (2001), pp. 39-43

[7] Doddington, G., et al., "The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation", *Proc. LREC* (2003)

[8] Garofolo, J., Auzanne, C., and Voorhees, E., "The TREC Spoken Document Retrieval Track: A Success Story", *Proc. RIAO 2000*, Paris, France

[9] http://www.nist.gov/speech/

[10] Martin, A., et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. EuroSpeech* Vol. 4 (1997), pp.1895-1898

[11] Swets, John A. ed., "Signal Detection and Recognition by Human Observers", John Wiley & Sons, Inc., pp. 611-648, 1964

[12] http://www.nist.gov/speech/tests/sigtests/sigtests.htm