Session 3: Human Language Evaluation

Lynette Hirschman

MITRE Corporation
Bedford, MA 01730

This session focused on experimental or planned approaches to human language technology evaluation and included an overview and five papers: two papers on experimental evaluation approaches[1, 2], and three about the ongoing work in new annotation and evaluation approaches for human language technology[3, 4, 5]. This was followed by fifteen minutes of general discussion.

When considering evaluation, it is important to consider the basic issues involved in evaluation:

- **Why evaluate:** what are the goals of evaluation?
- **What to evaluate:** what function(s) of the system should be evaluated, e.g., what input/output pairs are compared?
- **How to evaluate:** what procedures can be used to evalute specific system functions (or to grade goodness of input/output pairs)?
- **Where to go from here:** what additional evaluations are needed and what can be developed to support future research?

1. WHY EVALUATE?

Evaluation serves a number of purposes:

- **Cross-system evaluation:** This is a mainstay of the periodic ARPA evaluations on competing systems. Multiple sites agree to run their respective systems on a single application, so that results across systems are comparable. This includes evaluations such as message understanding (MUC)[6], information retrieval (TREC)[7], spoken language systems (ATIS)[8], and automated speech recognition (CSR)[8].
- **Within-system progress:** This is perhaps the most important role because it supports incremental system development, debugging and even hill climbing and automated learning approaches, if fast evaluation methods are available.
- **Understanding design trade-offs:** It is well-known that there are trade-offs in system design, e.g., between speed and error rate for speech recognition systems; similarly, there may be trade-offs in error recovery and types of feedback in dialogue-based systems. Appropriate evaluation methods make it possible to design controlled experiments to investigate these trade-offs.
- **Directing research focus:** Evaluation (especially when associated with research funding) brings increased attention to the technology being evaluated. It also fosters increased infrastructure to support evaluation, and in turn, infrastructure supports evaluation.[1] The success of the ARPA human language technology program can be attributed in part to the judicious use of common evaluation to focus attention on particular research issues, resulting in rapid improvement in the technology, increased sharing of technical information, and broader participation in the research activities.

2. WHAT TO EVALUATE?

Once we decide to evaluate, the first question is what to evaluate? Where do we put probes to inspect the input and output, in order to perform an evaluation? This issue is discussed in the Sparck Jones paper[1]. In some cases, we can evaluate the language technology in isolation from any front-end or back-end application, as shown in Figure 1, where probes are inserted on either side of the language interface itself. This gives us the kind of evaluation used for word error rate in speech (speech in, transcription out) or for machine translation, as proposed in the Brew/Thompson paper (source text in, target text out)[2]. This kind of evaluation computes output as a simple function of input to the language system.

Unfortunately, it is not always possible to measure a meaningful output – for example, researchers have struggled long and hard with measurements for *understanding* – how can a system demonstrate that it has understood? If we had a general semantic representation, then we could insert a probe on the output side of the semantic component, independent of any specific application. The last three papers ([3, 4, 5]) take various approaches to the issue of predicate-argument

[1] The Penn Treebank parse annotations provide an interesting case where annotation supported evaluation. By creating a theory-neutral description of a correct parse, the Treebank annotation enabled researchers to take the next step in agreeing to use the parse annotations (bracketings) as a "gold standard" against which to compare system-derived bracketings[9]. This evaluation, in turn, has enabled interesting automated learning approaches to parsing.

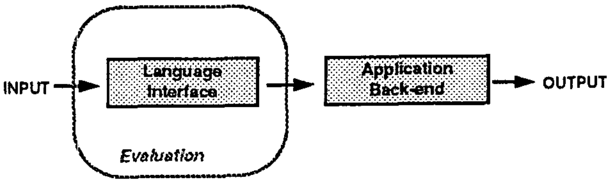

Figure 1: Evaluating Language Input/Output

structure in an attempt to define a more semantically-based and application-independent measure.

Right now, we can only measure understanding by evaluating an interface coupled to an *application* – Figure 2 shows the application back-end included inside the evaluation. This allows us to evaluate understanding in terms of *getting the right answer* for a specific task, as is done in the Air Travel Information (ATIS) system, which evaluates language input/database answer output pairs. However, this means that to evaluate spoken language understanding, it is necessary to build an entire air travel information system.

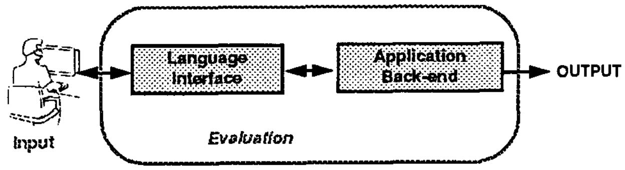

Figure 2: Evaluating Language Interface Plus Backend

Finally, for certain kinds of applications, particularly interactive applications, it is appropriate to enlarge the scope of evaluation still further to include the users. For interactive systems, this is particularly important because the user response determines what the system does next, so that it is not possible to use pre-recorded data.[2] Increasingly complex human-computer interfaces, as well as complex collaborative tools, demand that a system be evaluated in its overall context of use (see Figure 3).

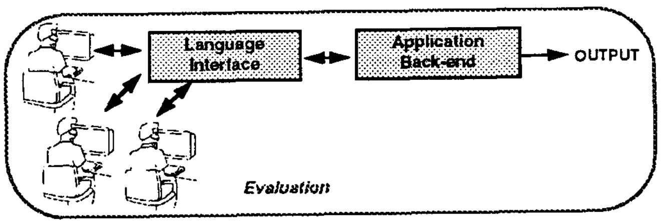

Figure 3: Evaluating Language Interfaces in Context of Use

3. HOW TO EVALUATE

We must not only decide what inputs and outputs to use for evaluation; we must decide **how** to evaluate these input/output pairs as well. Evaluation seems relatively easy when there is an intuitive pairing between input and output, for example, between speech signal and transcription at the word or sentence level. The task is much more complex when there is either no representation for the output (how to represent *understanding*?) or in situations where the result is not unique: what is the correct translation of a particular text? What is the best response to a particular query? For such cases, it is often expedient to rely on human judgements, provided that these judgements (or relative judgements) are reproducible, given a sufficient number of judges. Evaluation of machine translation systems[10] has used human judges to evaluate systems with differing degrees of interactivity and across different language pairs. The Brew and Thompson paper[2] also describes reliability of human judges in evaluating machine translation systems. Human judges have also been used in end-to-end evaluation of spoken language interfaces[11].

4. WHERE TO GO FROM HERE?

Because evaluation plays such an important role in driving research, we must weigh carefully what and how we evaluate. Evaluation should be theory neutral, to avoid bias against novel approaches; it should also push the frontiers of what we know how to do; and finally, it should support a broad range of research interests because evaluation is expensive. It requires significant community investment in infrastructure, not to mention time devoted to running evaluations and participating in them. For example, we estimate that the ATIS evaluation required several person-years to prepare annotated data, a staff of two to three people at NIST over several months to run the evaluation, time spent agreeing on standards, and months of staff effort at participating sites. Altogether, the annual cost of an evaluation certainly exceeds five person-years, or conservatively at least $500,000 per evaluation. Given this level of investment, it is critical to co-ordinate effort and obtain maximum leverage.

The last three papers[3, 4, 5] all reflect a concern to develop better evaluation methods for semantics, with a shared focus on predicate-argument evaluation. The Treebank annotation paper[3] discusses the new predicate-argument annotation work under Treebank. The paper by Grishman discusses a range of new evaluation efforts for MUC, which are aimed at providing finer grained component evaluations. The last paper, by Moore, describes a similar, but distinct, effort towards developing more semantic evaluation methods for the spoken language community.

5. DISCUSSION

The discussion began with the question: can we afford three somewhat similar but distinct predicate-argument evaluations? The resulting interchange helped to clarify the relationship between these three proposals. Both Marcus and Grishman argued that the Treebank annotation should directly support the MUC-style predicate-argument evaluation outlined in [4], although the Treebank annotations may be a sub-

[2] Pre-recorded data allows the same data to be used by all participating sites, effectively removing human variability as a factor in the evaluation.

set of what is used for MUC predicate-argument evaluation. The relation of the spoken language "predicate-argument" evaluation to the other two was less clear. Moore explicitly stated during the discussion (and Marcus agreed) that the Treebank annotation is quite different (more syntactic and more "surface") than the predicate-argument notation planned for spoken language. Moore believed that a deeper level (less syntactic and more semantic) was needed to meet the needs of (some parts of) the spoken language community. Thus, although the spoken and written language communities have an opportunity to converge on some common annotation and evaluation metrics, this may well not happen. These annotation and evaluation approaches are, however, "work-in-progress" and economic and time considerations may cause some convergence, even while theories and research agendas remain distinct.

References

1. Sparck Jones, K., "Towards Better NLP System Evaluation," this volume.
2. Brew, C., and Thompson, H. S., "Automatic Evaluation of Computer Generated Text: A Progress Report On the TextEval ·Project", this volume.
3. Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Ferguson, M., Katz, K. and Schasberger, B., "The Penn Treebank: Annotating Predicate Argument Structure," this volume.
4. Grishman, R., "Whither Written Language Evaluation," this volume.
5. Moore, R. C., "Semantic Evaluation for Spoken Language Systems," this volume.
6. Sundheim, B. and Chinchor, N., "Survey of the Message Understanding Conference," Proc. of the Human Language Technology Workshop, ed. M. Bates, Princeton, March 1993.
7. Harman, D., "Overview of the Second Text Retrieval Conference," this volume.
8. Pallett, D., Fiscus, J., Fisher, W., Garofolo, J., Lund, B., Pryzbocki, M., "1993 Benchmark Tests for the ARPA Spoken Language Program" this volume.
9. Black, E., et al., "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars," Proc. of the Speech and Natural Language Conference, ed. P. Price, 1992.
10. White, J. S., O'Connell, T., "Evaluation in the ARPA Machine Translation Program: 1993 Methodology," this volume.
11. Hirschman, L. et al., "Multisite Data Collection and Evaluation in Spoken Language Understanding," Proc. of the Human Language Technology Workshop, ed. M. Bates, Princeton, March 1993.