Issues in Reuse of Evaluation Data and Metrics

Donna Harman

National Institute of Standards and Technology

`donna.harman@nist.gov`

There has been a major explosion in the number of large-scale evaluations in NLP, along with an increasing interest in these evaluations by an expanding number of research groups. An example of this is the growth of the question-answering evaluation in TREC, starting with 20 participating groups in 1999 and now attracting more than 35 groups. Many of these groups are newcomers to these types of evaluations and sometimes also newcomers to research in a given area.

Each of these evaluations leaves behind large sets of evaluation data, which were expensive to create. They also leave behind sets of metrics and testing methodology that represent a huge investment in time by many of the participating research groups. Usually these metrics and methodologies were jointly created based on many discussions among the researchers and the evaluation organizers; additionally they are usually complex and require time for sufficient understanding by all who are involved.

The ability to reuse the data, testing methodology and/or metrics, either for the original testing environment or for a new type of environment is very attractive. In many cases reuse within a given community has been planned for from the beginning and the test sets and metrics become a valuable community resource. Examples of these types of resources are the data and metrics from the MUC evaluations and the TREC ad hoc evaluations.

This talk will examine the issues of reusability both within a given community and across communities. Whereas the benefits of reuse are obvi-

ous, there are many subtle problems that might not be clear, especially to people new to research in a given area.

All evaluations are based on many assumptions. For example they may be modelling a specific type of application, molding the data and the metrics to insure that they accurately mirror the intended use of the technology. So using material or metrics from these evaluations for testing in a different application may provide misleading results. Again this problem is often very subtle. Here an example would be using speech transcription data created based on broadcast news to predict how well a given system would perform in a live transcription application (different vocabularies, speaker characteristics, and acoustic devices).

Another danger involves the assumptions built into the scale of the evaluation and the metrics that allow for statistical significance of the results to be tested. Each metric has certain statistical properties that were (hopefully) carefully examined in the original evaluation, and these properties may make a given metric inappropriate in a different evaluation.

Despite all these dangers, it is critical that we get the maximum use of these expensive resources. Workshops such as this one should be encouraged as a way to both share resource information and carefully examine the assumptions built into these resources.