Pragmatic Influences on Sentence Planning and Surface Realization: Implications for Evaluation

Amanda Stent
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
amanda.stent@gmail.com

Abstract

Three questions to ask of a proposal for a shared evaluation task are: *whether* to evaluate, *what* to evaluate and *how* to evaluate. For NLG, shared evaluation resources could be a very positive development. In this statement I address two issues related to the what and how of evaluation: establishing a "big picture" evaluation framework, and evaluating generation in context.

1 Introduction

Recently, shared evaluation tasks have been used in IE, parsing, semantic role labeling, QA and MT. These shared tasks have resulted in new corpora, tools and performance metrics. Because NLG is a small field, shared evaluation resources could be a very positive development. However, we should avoid a common trap of shared evaluation tasks: a too-narrow evaluation framework and simplistic performance metrics leading to devaluing of interesting applications and research problems. In this statement, I address these two issues in turn.

2 An Evaluation Framework for NLG

We should avoid the urge to adopt shared evaluation tasks that unnecessarily limit NLG research. I propose a broad *shared evaluation framework* organized around the reference NLG architecture proposed in (Reiter and Dale, 1997). The framework has three dimensions: *discourse type* (e.g. summaries, explanations, comparisons), *application* (e.g. tutoring, question answering), and *generation task*. Generation tasks are further organized into task types (selection/organization) and levels (Table 1).

Level	Selection	Organization
discourse	content selection	discourse planning
paragraph	discourse cue assignment	sentence aggregation
sentence	lexical selection RE generation	surface realization
media	media selection	media coordination

Table 1: Generation tasks

This framework could be used immediately, while the evaluation discussion continues. If we set up a wiki organized according to this (or another) framework, researchers could immediately start sharing evaluation resources such as corpora and tools. Shared evaluation tasks could be chosen from discourse type/ application/generation task triples for which data and/or multiple implementations exist (Reiter and Belz, 2006). Lessons learned from evaluations for one discourse type/application/generation task could be applied to other discourse types and applications. Instead of focusing research on one generation task, a shared framework could lead to more substantial and interesting evaluations in a variety of areas.

3 Evaluation in Context

High-quality generation makes heavy use of context information such as user models, discourse history, and the physical context of the discourse. For example, generation tasks affected by user preferences include content selection and ordering, media organization, and sentence aggregation (Reiter et al., 2003; Stent et al., 2004; Stent and Guo, 2005). I am particularly concerned about existing automatic evaluation metrics for surface realization (e.g. BLEU, NIST) because they do not take context into account. In particular, they: use a small number of reference outputs selected without regard to the generation context; conflate the measurement of fluency and adequacy (meaning preservation); and conceal rather than reveal the types of errors found. Consequently, it is difficult to do error analyses or compute the relative impact of system changes on output fluency and adequacy (Stent et al., 2005; Scott and Moore, 2006). This makes it hard to evaluate how context information affects system performance.

In the evaluation framework presented here, each generation task includes a subtask devoted to 'selection' and another devoted to 'organization'. Selection subtasks can be evaluated by information extraction-like metrics (a combination of counts of correct, missing and spurious elements giving precision and recall measures). These metrics give counts useful in error analysis. Ordering subtasks are harder to evaluate automatically. Traditionally, most ordering subtasks are performed using tree data structures (e.g. sentence plan trees), so tree edit distance metrics can be used (Bangalore et al., 2000). For automatic evaluations, human judges can select reference outputs taking context into account.

In our research on ordering tasks, we use human evaluations. The evaluator is presented with the generation context, then given randomly ordered possible outputs from different systems (including the reference sentence(s)). The evaluator ranks the possible outputs from best to worst, and separately notes whether each possible output is inadequate or ambiguous, disfluent or awkward. We use standard statistical methods to compare the systems contributing outputs to the evaluation, and can easily perform error analyses. We could contribute our evaluation tools to an evaluation wiki. With a shared evaluation, the human evaluation effort could be shared across sites and the cost to any particular research group minimized.

4 Summary

In the NLG community, recent efforts to provide shared evaluation resources (e.g. the SumTime corpus) should be encouraged. A shared evaluation framework should encourage the full range of NLG research.

Because generation output quality is dependent on context, generation output should be evaluated in context and evaluation metrics and tools should be developed that incorporate context, or at least facilitate error analyses to permit exploration of the impact of context.

References

S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of INLG.*

E. Reiter and A. Belz. 2006. GENEVAL: A proposal for shared-task evaluation in NLG. In *Proceedings of INLG Special Session on Sharing Data and Comparative Evaluation.*

E. Reiter and R. Dale. 1997. Building applied natural-language generation systems. *Journal of Natural-Language Engineering*, 3:57–87.

E. Reiter, R. Robertson, and L. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.

D. Scott and J. Moore. 2006. An NLG evaluation competition? eight reasons to be careful. In *Proceedings of INLG Special Session on Sharing Data and Comparative Evaluation.*

A. Stent and H. Guo. 2005. A new data-driven approach for multimedia presentation planning. In *Proceedings of EuroIMSA.*

A. Stent, R. Prasad, and M. Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of ACL 2004.*

A. Stent, M. Marge, and M. Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of CICLing.*