Shared Tasks and Comparative Evaluation for NLG: to go ahead, or not to go ahead?

Barbara Di Eugenio
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60304
{bdieugen}@cs.uic.edu

When I read this call for paper, my initial reaction was quite enthusiastic at the perspective of a new, brighter day for NLG. However, a few doubts immediately arose in my mind. At this point, I lean more towards developing appropriate metrics for evaluation rather than shared tasks. I will discuss here why I find the idea attractive, but also why I cannot quite buy it.

The two areas I've worked in the most during my career as a researcher in NLP have been discourse / dialogue processing (DDP) and NLG. Not surprisingly, more than once I have felt a pang of envy for researchers in those other areas of NLP with clear evaluation metrics or at least an agreed upon dataset on which applications can be evaluated, e.g. the Penn Treebank for parsers. The envy is even greater, since I feel principled work in DDP and NLG requires humongous effort (Di Eugenio et al., 2003):

1. You need to start with data collection and annotation, since 99% of the appropriate corpora do not exist. For example, in the last 6-7 years I have been working on generating feedback in intelligent tutoring systems (ITSs). We have worked in three different domains (diagnosis of mechanical systems, letter pattern completion tasks, and basic data structures and algorithms in Computer Science). We had to collect and annotate data in each of these domains, since none existed we could use.

2. Then, you need to proceed through computational modeling and implementation.

3. Finally, you need to run an evaluation that, to be convincing, most often needs to include human subjects.

Shared tasks and comparative evaluations are very attractive because they would short circuit the first and the third steps in the process. To be realistic, the tasks to be shared would be based on at least some corpus analysis; and the comparative evaluations on the shared dataset would not require evaluation with human subjects.

The big question is, what would participating in such an enterprise do for each specific project, both theoretically and practically. For example, how does participating in a task on say generating route descriptions help me develop the feedback generator for my Computer Science ITS? This point is articulated very well by Donia Scott and Johanna Moore in their position paper at the INLG workshop in 2006 (Scott and Moore, 2006). In fact, they articulate seven additional reasons to be cautious. I agree with most of them, in particular with the danger of stifling research and the need for funding. I'll elaborate on these two here.

I am concerned with how the community uses shared tasks and evaluations. The danger is that anybody who does not participate or performs a different task is shunned, because then their work cannot be compared to the rest. For example, if you do summarization but you don't evaluate your system on DUC data, reviewers are quick to kill your paper. This can also happen with evaluation measures of course, as attested by the discussion of measures of intercoder agreement, specifically Kappa, in which I have been an active participant (Krippendorff, 1980; Carletta, 1996; Di Eugenio and Glass, 2004). Providing measures of intercoder agreement is essential to being able to assess the quality of coded data; however, the hard part is to understand what the values of Kappa mean. Especially when reviewing papers, most researchers still blindly adopt a scale tentatively proposed by Krippendorff that discounts any

$K < .67$, even if Krippendorff himself notes that his are just guidelines, and that Kappa values must be related to the researcher's specific purposes and his/her tolerance of disagreement.

I am also convinced that any effort to come up with shared resources needs to be financially supported, and cannot only be based on volunteer work. I am referring to e.g. actually paying somebody to run the competitions, as NIST does with TREC. An opposite point of view is reported in (Belz and Dale, 2006):

> Money would be needed for data resource creation, but not necessarily for anything else; evidence that this was possible could be found in successful and vibrant shared-task initiatives run on a shoe-string, such as CoNLL and SENSEVAL.

However, in my experience, volunteer work can only go that far, as I witnessed when I participated in the Discourse Resource Initiative in the mid nineties. The goal was to devise a tagging scheme for discourse / dialogue that could be used as a standard. I attended three workshops, all the participants did their homework prior to the workshops, but then the effort fizzled out because nobody could sustain it in their "spare" time. There was no funding to e.g. pay annotators to try out the coding schemes that were developed at those workshops. Mind you, the effort was not wasted, because it led to the DAMSL coding scheme for dialogue acts (Allen and Core, 1997), which in turn was the basis for a variety of coding schemes, e.g. (Jurafsky et al., 1997; Di Eugenio et al., 2000; Hardy et al., 2002).

To conclude, I'd be more inclined towards coming up with agreed upon evaluation measures that we can all use, as (Paris et al., 2006) has already proposed. As a start, we could adapt and build on the Paradise framework for dialogue systems evaluation (Walker et al., 1997).

References

J. Allen and M. Core. 1997. Draft of DAMSL: Dialog act markup in several layers. Coding scheme developed by the participants at two Discourse Tagging Workshops, University of Pennsylvania March 1996, and Schloß Dagstuhl, February 1997.

A. Belz and R. Dale. 2006. Introduction to the INLG'06 special session on sharing data and comparative evaluation. In Proceedings of INLG06, Special Session on Sharing Data and Comparative Evaluations.

J. Carletta. 1996. Assessing agreement on classification tasks: the Kappa statistic. Computational Linguistics, 22(2):249–254. Squib.

B. Di Eugenio and M. Glass. 2004. The Kappa statistic: a second look. Computational Linguistics, 30(1):95–101. Squib.

B. Di Eugenio, P. W. Jordan, R. H. Thomason, and J. D. Moore. 2000. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. International Journal of Human Computer Studies, 53(6):1017–1076.

B. Di Eugenio, S. Haller, and M. Glass. 2003. Development and evaluation of nl interfaces in a small shop. In 2003 AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue, Stanford, CA, March.

H. Hardy, K. Baker, L. Devillers, L. Lamel, S. Rosset, T. Strzalkowski, C. Ursu, and N. Webb. 2002. Multi-layer dialogue annotation for automated multilingual customer service. In ISLE Workshop: Dialogue Tagging for Multi-Modal Human Computer Interaction, Edinburgh, Scotland.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. Technical Report 97-02, University of Colorado, Boulder. Institute of Cognitive Science.

K. Krippendorff. 1980. Content Analysis: an Introduction to its Methodology. Sage Publications, Beverly Hills, CA.

C. L. Paris, N. Colineau, and R. Wilkinson. 2006. Evaluation of NLG systems: common corpus and tasks or common dimensions and metrics? In Proceedings of INLG06, Special Session on Sharing Data and Comparative Evaluations.

D. Scott and J. D. Moore. 2006. An NLG evaluation competition? eight reasons to be cautions. In Proceedings of INLG06, Special Session on Sharing Data and Comparative Evaluations.

M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In ACL-EACL97, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pages 271–280.