

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

Barbara Di Eugenio

Computer Science
University of Illinois at Chicago

The excitement

NLG is hard! whether targeted at monologue or at dialogue

1. Start with corpus collection and annotation – any new task / domain requires its corpus
2. Proceed through computational modeling and implementation
3. Run evaluation (often with human subjects)

The excitement (ctd.)

Data collection / analysis (step 1) is extremely time consuming

Evaluation (step 3) can be too, especially if one doesn't get it right the first time

STEC would potentially short circuit steps 1 and 3:

- tasks to be shared would be based on at least some corpus analysis, performed by community, not by individual site
- comparative evaluations on the shared dataset would not require evaluation with human subjects

Doubts

- Find shared task of sufficient interest to many researchers is unlikely (workshop may prove me wrong)
- “Sociology” of science: what happens when the community focuses on those tasks and competitions?
- How far can you go without funding?

Doubts (ctd.)

The *have's* and *have-not's*: those who participate in the STEC are in, the others are out.

- You need to use the same corpora as in the STEC so you can compare ... but then when you do it, you are still criticized
- The community gets “fossilized” in its evaluation measures:
 - ROUGE for summarization ... until PYRAMID came out
 - the magic .67 for Kappa for interannotator agreement [Krippendorff 80, Carletta 96, Di Eugenio & Glass 04]

Doubts (ctd.)

How far can you go without sustained funding?

Example: DRI (Discourse Resource Initiative), mid nineties, to devise standard annotation schemes for discourse and dialogue phenomena

- Funding for three well attended workshops (Philadelphia, USA; Dagstuhl, Germany; ?, Japan).
- Then effort fizzled out because nobody could sustain it: need money e.g. to pay annotators to systematically try out coding schemes
- Not wasted effort though. E.g. DAMSL code for speech acts [Allen & Core 97] spawned other efforts (SWBD-DAMSL [Jurafsky et al 97], COCONUT [Di Eugenio et al 00]). Referential expressions annotation effort was folded into MATE initiative

More fruitful: develop framework for evaluation

If we had a shared framework for evaluation (not just a single measure!), we could better situate the performance of our systems – not to compare them, but to be able to assess how they perform in relation to the difficulty of the task and many other factors.

Concretely: build on PARADISE scheme for dialogue system evaluation [Walker, Litman, Kamm, Abella 98], e.g. by bringing in factors proposed by [Paris et al, this workshop]

PARADISE

Objective: Maximize User Satisfaction

- Maximize Task Success (uses Kappa)
- Minimize Costs
 - Efficiency Measures: number of utterances, dialogue time, etc
 - Qualitative Measures: agent response delay, repair ratio, etc

Operationalized via performance function; multiple linear regression used to compute contribution (weight) of each factor to predicting objective, i.e. user satisfaction