To Share a Task or Not: Some Ramblings from a Mad (i.e., crazy) NLGer

Kathleen F. McCoy
Computer and Information Sciences
University of Delaware
Newark, DE 19716
mccoy@cis.udel.edu

To me the question is not whether or not there should be a shared task - the question is: what is the best way to move ”the field” forward. Part of the issue that I see here is that it is not at all clear how ”the field” should be defined (let alone how we should move it forward). For instance, one thing that struck me in the 2006 INLG Workshop was the variety in the problems addressed by the papers. Part of the issue that I see is that there is so much to do, so many things to solve, so many places where there are important problems that need to be addressed, that it isn't clear what should ”be chosen” as THE task.

The age old argument as to what makes INLG different from ”those other shared task fields” is that there is no clear consensus on what the input to INLG is. It is also the case that there is no clear consensus as to what is important in the output. Thus it is difficult to imagine a shared task.

From someone who is arguing for a shared task, there are some questions that I need to understand that might influence what my ultimate decision is.

- What do you envision a shared task being? The real question here has to do with both how and why you expect people to interact in this task.
  - A competition for money?
  - A funded activity in itself?
  - A competition just for the fun of it?
  - A competition or a cooperation? A competition would mean researchers go off and work on something, and then come together every so often for a competition where the fruits of their labor are pitted against each other. A cooperation would entail groups of researchers collaborating on a larger system. The cooperation may or may not also contain a competition but that's not the main goal.
- What is the desired outcome?
  - An advance in technology that may be applicable in lots of different places?
  - An advance in NLG technology that will allow more commercialization? bigger web presence? more excitement?
  - More funding for INLG research?
  - More publications of INLG research?
- What is the envisioned output that is going to lead to that outcome?
- On what basis is this output evaluated.

1 Some reasons for being against a shared task

One of my biggest fears with a shared task is that the evaluation may shut people out (or shut out ”the right” way of actually tackling the problem). My case in point here is the area of text summarization which is a task that (to any NLG person) cries out for strong NLG research (at least as a major component). The problem is that the evaluations they have adopted preclude doing any NLG work. That is, the scoring mechanisms do better with sentence extraction methods rather than some deeper extraction coupled with generation. But why is this? I

believe most would acknowledge that the actual results would be better with generation. But, in order to actually score the competition, a fairly automatic scoring mechanism was developed. After all, with generated text, how would it be evaluated? One must acknowledge that it is really hard to reduce features like text coherence (essential to NLG) down to a single number to be compared against others. No matter how you decide to measure text coherence, it won't be right. Text coherence is not well enough understood.

Just because the text summarization shared task chooses to be generation unfriendly is not such a big deal. Just because someone interested in generation is not going to score well in that particular competition, doesn't stop them from still doing generation; it just stops them from participating in that competition. But, this is not so. Perhaps because the competition is successful, it has created quite an exclusive community and that community has seeped into other areas - most notably, publications. What this means is that it becomes very difficult to get work published that has anything to do with text summarization if you don't play the game of that competition. The metric for the competition has become the metric by which research is judged in that area, to the exclusion of other research. This despite the acknowledgment from most of the shared task participants that the evaluation metric is sorely lacking.

So, the problem here is that a competition that on the face of it is good for INLG turns out to squelch it. The only ones that get to do work remotely related to the shared task have to devote substantial efforts to what scores well in the competition (and hope they can stand in long enough and fight for a change in the evaluation metrics).

Lesson: A poor choice of an evaluation method can adversely affect the outcome by discouraging (indeed discrediting) research that is ultimately necessary for forward progress in the field.

That is to say, a successful shared task may have the side effect of squelching research that is important just because it either looks at the problem differently or because it takes an approach that does not stand up well against the chosen evaluation metric.

A second, related, point has to do with the kind of processing that may be favored by shared task competitions. For example, the early MUC conferences generated a lot of work and had many accomplishments. But, in the end, the MUC conferences caused a lot of people to do "domain hacking" rather than finding deeper solutions to the problem. Is INLG at the stage where it is ready to go off with disregard to these deeper solutions? One important thing to guard against in any shared task/evaluation is that it not favor shallow processing methods (particularly to the exclusion of "deeper" methods requiring theoretical advances). But, if one also thinks about it, isn't just such an evaluation metric (i.e., a shallow/automatic one) almost necessary for shared task evaluation? My personal feeling is that we do not understand enough to be able to develop evaluations that are going to be broad enough to cover the really important aspects of the field. The consequence could be that those important aspects will be left unstudied as systems try to optimize on the selected metric.

Let's keep in mind what we want. What makes generation different from understanding? What is it that we like about this field? Generation puts emphasis on some aspects of processing that can be ignored in understanding. Two examples are syntax (which one might arguably ignore in understanding but it is pretty difficult to ignore if one is generating) and coherence (which one can get quite far by ignoring in understanding). Ignoring coherence in generation becomes very apparent very quickly (making the text very difficult for a reader to process). Yet these very same problems of such interest are very difficult to quantify into a metric.

It is not clear to me at this point that we understand what the problems are in generation well enough to posit a shared task for the field that is going to further things. I think there must be better ways to further the field.

2 Questions to Ponder

- What is the underlying purpose of the suggestion of a shared task?
- Is a shared task actually the way to accomplish that purpose?
- Is there another mechanism that might actually work better?