Share and Share Alike: Resources for Language Generation

Marilyn Walker
Department of Computer Science
University of Sheffield
Sheffield, S1 4DP, United Kingdom
M.A.Walker@sheffield.ac.uk

1 Introduction

It has been proposed that the NLG community could benefit from from the introduction of 'shared task evaluations', where systems with comparable functionality, that take the same input and produce similar outputs, are submitted to an evaluation 'bakeoff'. These STECs would provide shared data sets consisting of inputs and human-written text outputs for each input.

Scott and Moore (2006) argue that this approach may not make sense because: (1) the input and output for NLG, and for individual modules in NLG, is unclear, given the wide range of settings (e.g. dialogue vs. text) application domains, and theories used in NLG; (2) the evaluation metrics to be used are unclear, and recent work in machine translation evaluation has called into question the use of automatic metrics calculated from texts such as ROUGE and BLEU; (3) the ability to plug-and-play NLG components by clearly defining the interfaces between different NLG modules would contribute more to progress in the field than would STECs; and (4) STECs are supported by huge amounts of funding for applications that are regarded as 'killer aps', and it is unclear what those applications are for language generation.

As argued elsewhere, what I would characterize as the most essential difference between language generation and other language processing problems is that there is no single right answer for language generation [5]. Rather, there are a very large set of alternative possible outputs, which can be ranked along specific criteria, but these criteria will vary depending on the intended application and context of use. Thus any resource based on the assumption of a single correct output will be flawed. This is identical to the issue of resources for dialogue systems [2]. Thus for a resource to be useful, it must meet the LANGUAGE PRODUCTIVITY ASSUMPTION:

> An optimal generation resource will represent multiple outputs for each input, with a human-generated quality metric associated with each output.

This assumption does not imply that it is impossible to do any automatic evaluation of generation outputs. As we argued for dialogue systems [6], and was argued subsequently for generation [1], it is possible to approximate human judgements with an automatic evaluation metric learned from a corpus of outputs, automatically calculated metrics on those outputs, and human judgements.

However, it is also true that any almost type of shared resource would be helpful for scientific progress in language generation. Especially PhD and masters students could benefit from a large variety of different types of shared resources, but I believe that the most useful resources would not be of the type described for STECs, but rather resources for particular NLG modules, with their interfaces clearly specified (Mellish etal 2006). Moreover, it is unclear whether such resources could best be provided by a large government STEC. Rather, I would argue that resources developed by researchers in the field to support their own work would, if made available, contribute more to progress in the field.

Why hasn't this already happened? There are

shared tools for realization, such as Halogen, RealPro and Open-CCG, which are becoming widely used, but datasets of inputs and outputs that could be used to compare algorithms in evaluation experiments are needed. There are at several reasons why this has not already happened, i.e. why many scientists do not make resources that they have developed and used in their own work available:

1. There are many different problems and domains addressed by research in language generation, so that it has been unclear what could be shared usefully.

2. Resources are costly to develop and scientists often are not sure that they are 'finished' with a resource, and need to ensure their work is published before giving the resource away.

3. Scientists who are not used to sharing resources don't realize that having other scientists use your resource and therefore build on your work can be extremely valuable in the long term (e.g. use of your resource by other scientists is guaranteed to lead to more citations of your work);

4. Researchers are afraid if they release software or data resources to the community that they will end up spending a lot of time answering questions about how to use the resource;

5. It takes a lot of time to get a resource organized and documented and put on a web page for other people to use. If the scientist changes affiliation or the web page structure at the site changes, this infrastructure has to be recreated or maintained.

If these problems could be overcome, much of recent research in language generation could produce shared resources. NSF funding for small grant amounts to address problem (5) could help a lot. LDC involvement in resource databanking and provision would address the distribution and maintenance problems. In the following section I describe a resource that could be easily shared and which would be very useful in my view.

2 A Shared Resource for Information Presentation

Natural language interfaces to databases has been a primary application for language generation for many years[3]. Early work in NLG introduced two classic problems: (1) paraphrasing the user's input [4], and (2) generating information presentations of sets of database entities, such as summaries, comparisons, descriptions, or recommendations (McKeown, 1985; McCoy 1989; DembergMoore 2006; Polifroni etal 2003) *inter alia*. Given the databases currently in use in both civilian and military application, and the potential to use NLG in this context without the need for NL input, a language generation resource of potentially wide interest would consist of:

- INPUT: a speech act from the set *summarize, recommend, compare, describe*, and a set of one or more database entities in terms of slots and values representing the content.

- OPTIONAL INPUT: user model, dialogue context, or other parameters affecting output, to constrain and make apparent the context for generation.

- OUTPUT-1: a set of alternative outputs (possibly with TTS markup);

- OUTPUT-2: human generated ratings or rankings for outputs.

An example of the outputs and ratings, from my own work is given in Figure 1. An advantage of this resource is that the representation of sets of database entities is not controversial. It also allows for content selection, sentence planning, and realization differences, and for the derivation and use of inferences over database entities, such as time series. It supports the use of generation parameters such as different types of user models, and the use of constraints on the speaker such as style or personality type.

References

[1] Srinivas Bangalore, Owen Rambow, and Steve Whittaker. Evaluation metrics for generation. In *Proceedings of the First International Natural Language Generation Conference (INLG2000)*, Mitzpe Ramon, Israel, 2000.

[2] Lynette Hirschman. Evaluating spoken language interaction: Experiences from the DARPA spoken language program 1990–1995. In S. Luperfoy, editor, *Spoken Language Discourse*. MIT Press, Cambridge, Mass., 2000.

[3] Karen Sparck Jones. Natural language and databases, again. In *COLING84: Proc. 10th International Conference on Computational Linguistics*, pages 182–183, 1984.

Alt	Realization	Judge A	Judge B
8	Chanpen Thai is a Thai restaurant, with good food quality. It has good service. Its price is 24 dollars. It has the best overall quality among the selected restaurants.	4	2
1	Chanpen Thai has the best overall quality among the selected restaurants. This Thai restaurant has good food quality. Its price is 24 dollars, and it has good service.	4	3
2	Chanpen Thai has the best overall quality among the selected restaurants. Its price is 24 dollars, and it is a Thai restaurant. It has good food quality and good service.	4	4
10	Chanpen Thai has the best overall quality among the selected restaurants. It has good food quality. Its price is 24 dollars. It is a Thai restaurant, with good service.	3	3
3	Chanpen Thai has the best overall quality among the selected restaurants. Its price is 24 dollars. It is a Thai restaurant, with good service. It has good food quality.	3	3
5	Chanpen Thai has the best overall quality among the selected restaurants. It has good service. It has good food quality. Its price is 24 dollars, and it is a Thai restaurant.	3	2
9	Chanpen Thai is a Thai restaurant, with good food quality, its price is 24 dollars, and it has good service. It has the best overall quality among the selected restaurants.	2	4
4	Chanpen Thai has the best overall quality among the selected restaurants. It has good food quality, with good service, it is a Thai restaurant, and its price is 24 dollars.	2	4
7	Chanpen Thai has the best overall quality among the selected restaurants because it has good service, it has good food quality, it is a Thai restaurant, and its price is 24 dollars.	2	5
6	Chanpen Thai has the best overall quality among the selected restaurants since it is a Thai restaurant, with good service, its price is 24 dollars, and it has good food quality.	1	4

Figure 1: Some Alternative Sentence Plan Realizations for a recommendation given the DB entity for ChanPen Thai, with feedback from User A and User B

[4] Kathleen R. McKeown. Paraphrasing questions using given and new information. *Computational Linguistics*, Jan-Mar 1983.

[5] Marilyn A. Walker. Can we talk? methods for evaluation and training of spoken dialogue systems. *Language resources and evaluation*, 39(1):65–75, 2005.

[6] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language*, 12(3), 1998.