

Lessons Learned from Large Scale Evaluation of Systems that Produce Text: Nightmares and Pleasant Surprises

Kathleen R. McKeown
Department of Computer Science
Columbia University
New York, NY 10027
kathy@cs.columbia.edu

Extended Abstract

As the language generation community explores the possibility of an evaluation program for language generation, it behooves us to examine our experience in evaluation of other systems that produce text as output. Large scale evaluation of summarization systems and of question answering systems has been carried out for several years now. Summarization and question answering systems produce text output given text as input, while language generation produces text from a semantic representation. Given that the output has the same properties, we can learn from the mistakes and the understandings gained in earlier evaluations. In this invited talk, I will discuss what we have learned in the large scale summarization evaluations carried out in the Document Understanding Conferences (DUC) from 2001 to present, and in the large scale question answering evaluations carried out in TREC (e.g., the definition pilot) as well as the new large scale evaluations being carried out in the DARPA GALE (Global Autonomous Language Environment) program.

DUC was developed and run by NIST and provides a forum for regular evaluation of summarization systems. NIST oversees the gathering of data, including both input documents and gold standard summaries, some of which is done by NIST and some of which is done by LDC. Each year, some 30 to 50 document sets were gathered as test data and somewhere between two to nine summaries were written for each of the input sets. NIST has carried out both manual and automatic evaluation by comparing system output against the gold standard summaries written by humans. The results are made public at the annual conference. In the most recent years, the number of participants has

grown to 25 or 30 sites from all over the world.

TREC is also run by NIST and provides an annual opportunity for evaluating the output of question-answering (QA) systems. Of the various QA evaluations, the one that is probably most illuminating for language generation is the definition pilot. In this evaluation, systems generated long answers (e.g., paragraph length or lists of facts) in response to a request for a definition. In contrast to DUC, no model answers were developed. Instead, system output was pooled and human judges determined which facts within the output were necessary (termed “vital nuggets”) and which were helpful, but not absolutely necessary (termed “OK nuggets”). Systems could then be scored on their recall of nuggets and precision of their response.

DARPA GALE is a new program funded by DARPA that is running its own evaluation, carried out by BAE Systems, an independent contractor. Evaluation more closely resembles that done in TREC, but the systems’ scores will be compared against the scores of human distillers who carry out the same task. Thus, final numbers will report percent of human performance. In the DARPA GALE evaluation, which is a future event at the time of this writing, in addition to measuring properties such as precision and recall, BAE will also measure systems’ ability to find all occurrences of the same fact in the input (redundancy).

One consideration for an evaluation program is the feel of the program. Does the evaluation program motivate researchers or does it cause headaches? I liken Columbia’s experience in DUC and currently in GALE to that of Max in *Where the Wild Things Are* by Maurice Sendak. We began with punishment (i.e., if you don’t do well, your funding will be in jeopardy), encounter monsters along the way (seemingly arbitrary methods for

measuring output quality), finally tame the monsters and sail back peacefully across time. DUC has reached the peaceful stage, but GALE has not. The TREC definition pilot had less of a threat of punishment.

Evaluation in all of these programs began at the request of the funders, with the goal of comparing how well different funded systems perform. Improvement over the years is also measured in order to determine if funding is well spent. This kind of goal creates anxiety in participants and makes it most important to get the details of the evaluation right; errors in how evaluation is carried out can have great consequences. Coming to agreement on the metrics used, the methodology for measuring output and the tasks on which performance is measured can be difficult; the environment does not feel friendly. Even if evaluation within the language generation community was not initiated with the same goals, I think it is reasonable to expect a certain amount of disagreement as the program gets off the ground.

However, over time, researchers come to agreement on some portion of the task and these features become accepted. At this point in time, it is possible to see the benefits of the program. Certainly, within DUC, we are at this stage. DUC has generated large amounts of data, including both input document sets and multiple models of good output for each input set, which has spurred studies both on evaluation and summarization. Halteren and Teufel, for example, provide a method for annotation of content units and study consensus across summarizers (van Halteren and Teufel, 2003; Teufel and van Halteren, 2004b). Nenkova studies significant differences across DUC04 systems (Nenkova, 2005) as well as the properties of human and system summaries (Nenkova, 2006). We can credit DUC with the emergence of automatic methods for evaluation such as ROUGE (Lin and Hovy, 2003; Lin, 2004) which allow quick measurement of systems during development and enable evaluation of larger amounts of data. We have seen the development of manual methods for evaluation developed both within DUC (Harman and Over, 2004) and without. The Pyramid method (Nenkova and Passonneau, 2004) provides an annotation method and metric that addresses the issues of reliability and stability of scoring. Thus, research on evaluation of summarization has become a field in its own right result-

ing in greater understanding of the effect of different metrics and methodologies.

¿From DUC and TREC, we have learned important characteristics of a large-scale evaluation, of which the top three might be:

- Output can be measured by comparison against a human model, but we know that this comparison will only be valid if multiple models are used. There are multiple good summaries of the same input and if system output is compared against just one, the results will be biased.
- If the task is appealing to a wide audience, the evaluation will spur research and motivate researchers to join in. We have seen this with growth of participation in DUC. One benefit of summarization and QA is that the task is domain-independent and thus, no one site has an advantage over others through experience with a particular domain.
- Given the different ways in which evaluation can be carried out and the fact that different researchers may be biased towards methods which favor their own approach, it is important the evaluation be overseen by a neutral party which is not deeply involved in research on the task itself. On the other hand, some knowledge is necessary if the evaluation is to be well-designed.

While my talk will focus on large scale evaluation programs that feature quantitative evaluation through comparison with a gold standard, there has been work on task-based evaluation of summarization (McKeown et al, 2005). Task-based evaluation is more intensive and to date, has not been done on a large scale across sites, but shows potential for indicating the usefulness of summarization systems.

In this brief abstract, I've suggested some of the topics that will be covered in my talk, which will tour the land of the wild things for evaluation, illuminating monsters and highlighting events that will allow more peaceful sailing. Evaluation can be a nightmare, but over time and particularly if carried out away from the influence of funding pressures, it can nurture a community of researchers with common goals.

Acknowledgments

This material is based upon work supported in part by the ARDA AQUAINT program (Contract No. MDA908-02-C-0008 and Contract No. NBCHC040040) and the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 and Contract No. N66001-00-1-8919. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or ARDA.

References

- Harman, D. and Over, P. 2004. The effects of human variation in document summarization evaluation. In *Text Summarization Branches Out Workshop, ACL 2004*.
- Lin, C.-Y. 2003. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop in Text Summarization, ACL'04*.
- Lin, C.-Y. and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*.
- McKeown, K. and Passonneau, R.J. and Elson, D.K., and Nenkova, A., and Hirschberg, J. 2005. A Task-Based Evaluation of Multi-Document Summarization. In *Proceedings of SIGIR 2005*.
- Nenkova, A. 2005. Automatic text summarization of newswire: Lessons learned from the Document Understanding Conference. In *Proceedings of AAAI 2004*.
- Nenkova, A. 2006. Understanding the process of multi-document summarization: content selection, rewrite and evaluation. Ph.D Dissertation, Columbia University.
- Nenkova, A. and Passonneau, R. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*.
- Teufel, S. and van Halteren, H. 2004. Evaluating information content by factoid analysis: human annotation and stability. In *EMNLP-04*.
- van Halteren, H. and Teufel, S. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*.