Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation

Position Papers

Workshop Organizers: Robert Dale and Michael White

April 20-21, 2007
Arlington, Virginia

Sponsored by the US National Science Foundation and endorsed by SIGGEN, the special interest group in natural language generation of the Association for Computational Linguistics.

Invited Speaker

Kathleen McKeown, Columbia University, USA

Workshop Organizers

Robert Dale, Macquarie University, Australia
Michael White, The Ohio State University, USA

Review Committee

Anja Belz, University of Brighton, UK
Robert Dale, Macquarie University, Australia
Helmut Horacek, University of Saarbrücken, Germany
Donia Scott, The Open University, UK
Michael White, The Ohio State University, USA

Workshop Website

http://www.ling.ohio-state.edu/~mwhite/nlgeval07/

Preface

We are pleased to present the position papers for the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation. The aim of this workshop is to bring together leading international researchers in the field of NLG, with the aim of establishing a clear, community-wide, position on the role of shared tasks and comparative evaluation in NLG research. There are many views held in the NLG community as to the proper role of evaluation, but so far there has been little scope to hammer out these positions in a forum that provides the time and involvement required to subject the different views to rigorous debate. Our aim is for this workshop to provide that forum. We expect the workshop to result in the working out of a number of clearly argued positions on the issue that will serve as a base resource for the field moving forward. Further, we expect that, in line with the wishes of a number in the community, basic specifications will be worked out for a variety of shared task evaluation campaigns that can then be considered by the wider community.

The workshop schedule begins with an invited presentation by Kathleen R. McKeown entitled "Lessons Learned from Evaluation of Summarization Systems," and is followed on the first day by presentations of each of the 15 accepted position papers. On the second day, the schedule includes time for elaborating joint positions in working groups.

We would like to thank our invited speaker and authors for their participation, and the review committee for their assistance in putting together the program. We would also like to express our thanks to the US National Science Foundation for sponsoring the workshop, and to Tanya Korelsky in particular for her help and advice in organizing the event. We hope it is an enjoyable and productive experience!

Michael White and Robert Dale
Workshop Organizers

Table of Contents

Putting development and evaluation of core technology first

Anja Belz
NLTG, CMIS, University of Brighton, UK
A.S.Belz@itri.brighton.ac.uk

NLG needs comparative evaluation

NLG has strong evaluation traditions, in particular in user evaluations of NLG-based application systems (e.g. M-PIRO, COMIC, SUMTIME), but also in embedded evaluation of NLG components vs. non-NLG baselines (e.g. DIAG, ILEX, TAS) or different versions of the same component (e.g. SPOT). Recently, automatic evaluation against reference texts has appeared too, especially in surface realisation.

What has been missing are comparative evaluation results for comparable but independently developed NLG systems. Right now, there are only two sets of such results (for the SUMTIME weather forecasts, and for regenerating the Wall Street Journal Corpus). As a result, we have no idea at present what NLG techniques generally work better than others.

If NLG is a field of research that can progress collectively, rather than a loose collection of groups each progressing more or less independently, then it needs to develop the ability to comparatively evaluate NLG technology. This seems to me an absolutely fundamental principle for any branch of science and technology: without the ability to compare, results cannot be consolidated and there is no collective progress (Spärck Jones, 1981).

Shared tasks, but not necessarily shared data

That comparable techniques, components and systems need to perform comparable tasks — that comparative evaluation needs to be in that sense based on shared tasks — goes almost without saying. However, such tasks can be more or less loosely defined: implicitly by a set of paired inputs and outputs, or explicitly by a set of specifications and input/output requirements. Comparability increases if systems take the same type of inputs, and evaluation can be performed on the basis of a set of test inputs. Test-set evaluation can be useful in research-oriented evaluation, where results need to be obtained quickly and cost-efficiently. However, for evaluation at the application level, especially if it is user-based, test-input evaluation is often not necessary.

Core technology first, applications second

The single biggest challenge for comparative NLG evaluation is identifying sharable tasks: this is problematic in a field where systems are rarely developed for the same domain, let alone with the same input and output requirements.

One possibility is to propose an application for NLG researchers to develop systems for. These could then be evaluated according to ISO 9126 and 14598 on software evaluation, and this would shed light on the real-world usefulness of the systems.

However, NLG is a varied field with many applications and it will be hard to choose one that is recognised by a large enough number of researchers as their task. Moreover, evaluation at the application level would necessarily include application-specific content-determination techniques, and results would therefore not automatically generalise beyond the application. It would also not shed light on the usefulness or otherwise of any component technology.

We need an approach that unifies NLG, not one that creates a new subfield specialising in the chosen application. We need to focus on what unites NLG not what diversifies it. The way to do this is in my view to focus on the development and evaluation of core technology that is potentially useful to all NLG and to utilise the commonalities that have already evolved, in particular the more generally agreed subtasks such as GRE, lexicalisation, content ordering, or even a larger component like surface realisation.

Focus on output evaluation

The evaluation criteria general to all software systems covered by ISO standards 9126 and 14598 of course also apply to evaluating NLG systems, but we

still need to decide how to evaluate the — necessarily domain-specific — goodness of their outputs (one of the ISO criteria), and that is what research needs to focus on. Depending on how a shared task has been defined and whether a system or component is being evaluated, output evaluation could be in the form of added-value evaluation of components embedded within applications, direct evaluation of outputs or indirect evaluation by comparison against a set of reference texts. In terms of evaluation criteria, in the neighbouring disciplines of MT and summarisation, fluency and accuracy have emerged as standard criteria, and the latter now also assesses 'responsiveness' of a summary to the given topic, a criterion approximating 'real-world usefulness'.

Towards common subtasks, corpora and evaluation techniques

There are some subfields that have developed enough common ground to make it feasible to create a shared task specification straight away and have enough researchers able to participate (e.g. GRE). However, there is a lot that needs to be done to make this possible across larger parts of NLG.

Subtasks and input/output requirements need to be standardised to make core technologies truly comparable (as well as potentially reusable). In other NLP fields standardisation is often driven by evaluation efforts (e.g. in parsing), but it is probably more productive to work towards this in dedicated research projects. E.g. in the newly funded Prodigy Project, one of our core aims is to develop an approach to content representation that generalises to five different data-to-text domains.

Building data resources of NLG inputs and/or outputs may be the most straightforward way to encourage researchers to create comparable NLG systems. There are very few such resources at the moment, among them are the SumTime corpus, and the GREC corpus of short encyclopaedic texts for generating referring expressions in context that we are currently developing (Belz and Varges, 2007).

Creating NLG-specific evaluation techniques and assessing their reliability is essential so that we know how to reliably evaluate NLG technology. Such techniques should a assess the three criteria mentioned above: (i) language quality; (ii) appropriateness of content; and (iii) task-effectiveness, or how well do the generated texts achieve their communicative purpose.

We need a range of evaluation methods suitable for quick low-cost evaluation during testing of new ideas as well as reliable, potentially time and cost-intensive methods for evaluating complete systems. The aim of the GENEVAL initiative (Reiter and Belz, 2006) is to develop a range of evaluation techniques for NLG and to assess their reliability, ultimately aiming to provide NLG researchers with knowledge to decide which technique to use given their available time, resources and evaluative aim.

Concluding remarks

Comparative evaluation doesn't have to be in the shape of competitions with associated events (as opposed to just creating resources and encouraging other researchers to use them), but I happen to like the buzz and energy they create, the way they draw new people in, and the hot-housing of solutions they foster (Belz and Kilgarriff, 2006). It should at least be tried out to see whether it can work for NLG.

There's a lot of virtue in talking: discussing the options and trying to find consensus. But there's also virtue in doing — creating data and tasks and putting them out there for researchers to use if they want. Even organising competetive events to see if they work. The risks of getting it wrong seem small to me — shared-task evaluations can be run on a shoe-string (as SENSEVAL and CoNLL continue to demonstrate), and anyway, these things have a habit of self-regulating: if an event, task or corpus fails to inspire people, it tends to quietly go away.

References

A. Belz and A. Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *Proc. INLG'06*, pages 133–135.

A. Belz and S. Varges. 2007. The GREC corpus: Main subject reference in context. Technical Report NLTG-07-01, Natural Language Technology Group, CMIS, University of Brighton.

E. Reiter and A. Belz. 2006. GENEVAL: A proposal for shared-task evaluation in NLG. In *Proceedings of INLG'06*, pages 136–138.

K. Spärck Jones, 1981. *Information Retrieval Experiment*, chapter 12, page 245. Butterworth & Company.

Generating Instructions in Virtual Environments (GIVE): A Challenge and an Evaluation Testbed for NLG

Donna Byron[⋆] **Alexander Koller**[†] **Jon Oberlander**[‡] **Laura Stoia**[⋆] **Kristina Striegnitz**[◦]
[⋆] The Ohio State University [†] Columbia University [‡] University of Edinburgh [◦] Northwestern University
{dbyron|stoia}@cse.ohio-state.edu koller@cs.columbia.edu jon@inf.ed.ac.uk kris@northwestern.edu

Would it be helpful or detrimental for the field of NLG to have a generally accepted competition? Competitions have definitely advanced the state of the art in some fields of NLP, but the benefits sometimes come at the price of over-competitiveness, and there is a danger of overfitting systems to the concrete evaluation metrics. Moreover, it has been argued that there are intrinsic difficulties in NLG that make it harder to evaluate than other NLP tasks (Scott and Moore, 2006).

We agree that NLG is too diverse for a single "competition", and there are no mutually accepted evaluation metrics. Instead, we suggest that all the positive aspects, and only a few of the negative ones, can be achieved by putting forth a *challenge* to the community. Research teams would implement systems that address various aspects of the challenge. These systems would then be evaluated regularly, and the results compared at a workshop. There would be no "winner" in the sense of a competition; rather, the focus should be on learning what works and what doesn't, building upon the best ideas, and perhaps reusing the best modules for next year's round. As a side effect, the exercise should result in a growing body of shareable tools and modules.

The Challenge The challenge we would like to put forth is instruction giving in a virtual environment (GIVE). In this scenario, a human user must solve a task in a simulated 3D space (Fig. 1). The generation module's job is to guide the human player, using natural language instructions. Only the human user can effect any changes in the world, by moving around, manipulating objects, etc.

We envision a system architecture in which a central game server keeps track of the state of the world. The user connects to this server using a graphical client, and the generation system also connects to the server. Thus the implementation details of the virtual world are hidden from the generation system,

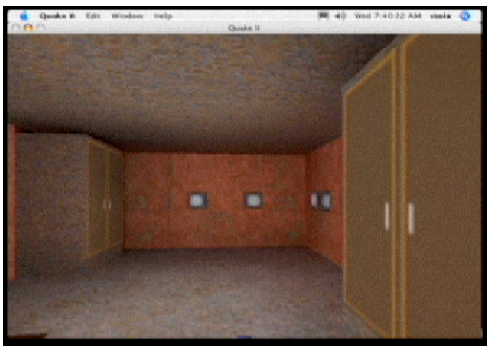

Figure 1: A sample virtual environment

which gets access to a symbolic representation of the world and a description of the task goal, and receives regular updates on the user's position, objects in his field of vision and their properties, etc. A sequence of actions that will achieve the goal is provided by an off-the-shelf planner.

There are numerous ways in which such a system could be evaluated. Quantitative measures can be collected automatically (completion time, success rate, percentage of generated referring expressions that the user resolved correctly), and subjective ones can be gathered from user satisfaction surveys. Since some 3D game engines, such as the open-source Quake II engine, support network play, it is technically possible to collect data cheaply from participants over the Internet.

Why this is a good challenge The proposed challenge spans a wide range of sub-problems of NLG, such as referring expression generation, aggregation, grounding, realization, and user modeling. On the other hand, the challenge can be scaled up and down along a number of different dimensions, both on the level of the challenge as a whole and on the level of individual systems. The output modality could be either text or speech; the system may or may not accept and process language input from the user; the user's position can be made discrete or even

simplified to a text-adventure-like "room" concept (Koller et al., 2004); and the system might choose to present all instructions in one block and expect the user to follow them without any further intervention. Furthermore, most tasks require only a simple ontology and a limited vocabulary, and the challenge is completely theory-neutral in that it makes no assumptions about the representations that a system uses internally. All this means is that many NLG researchers could find something interesting in the challenge, and even small research teams could participate, focusing on one module and implementing all others with simple template-based systems.

We are aware that generalized instruction-giving is beyond the capabilities of the current state of the art. That's what makes it a challenge. Comparable events, such as the Textual Entailment challenge (Dagan et al., 2005), have been very successful in revitalizing a research field and attracting outside interest. Furthermore, like the highly successful Robocup challenge and its more resource-light variants, GIVE has the benefit of addressing hard research issues in the context of a "fun" game-based scenario. Such scenarios can bring visibility to a field and encourage the entry of young researchers.

Finally, the GIVE challenge has the potential to lead to the development of practically relevant technologies. It is closely related to the problem of pedestrian navigation assistance (termed the "Black Hawk Down problem" in military circles; Losiewicz, p.c.), object manipulation tasks (the "Apollo 13" or "Baufix" problem), and training systems (Rickel and Johnson, 1998). On a more theoretical level, the GIVE problem has already been found to shed new light on standard NLG tasks. For example, Stoia et al. (2006) observed that human instruction givers avoid the generation of complex referring expressions; instead, they guide the user into a position where a simple RE is available.

Logistics Assuming that we decided to organize such a challenge, we would provide the computational infrastructure. We would distribute a software package to interested participants, including the 3D engine (perhaps based on the modified version of Quake created by Byron's research group), a framework for the generation system servers, a planner, and example maps.

During the challenge itself, the participating research teams would run their generation servers on machines at their own institutions. These would communicate with the central game server we provide. Experimental subjects would be made available by the challenge organizers. While we hope to be able to let subjects interact with the systems online, such a setup makes it difficult to ensure that the sample of subjects is representative. Thus we would probably run a dual evaluation for the first challenge, at which we have both online and controlled subjects, to verify the comparability of the results.

Finally, we would communicate the evaluation results to the participants and invite them to present system descriptions at a workshop. This would also serve as a forum for participants to evaluate the challenge, modify it for the future, and identify interesting subchallenges. To encourage cooperation and ensure a benefit for the community as a whole, we are considering to require participants to make their code available to the public. However, we recognize that this suggestion may discourage some from participating and needs to be discussed within the NLG community along with the other details of how to implement the proposed GIVE challenge.

References

I. Dagan, O. Glickman, and B. Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

A. Koller, R. Debusmann, M. Gabsdil, and K. Striegnitz. 2004. Put my galakmid coin into the dispenser and kick it: Computational linguistics and theorem proving in a computer game. *Journal of Logic, Language, and Information*, 13(2):187–206.

J. Rickel and W.L. Johnson. 1998. Steve: A pedagogical agent for virtual reality. In *Proceedings of the Second International Conference on Autonomous Agents*.

D. Scott and J. Moore. 2006. An NLG evaluation competition? eight reasons to be cautious. Technical Report 2006/09, Department of Computing, The Open University. `http://mcs.open.ac.uk/ds5473/publications/TR2006_09.pdf`.

L. Stoia, D. Byron, D. Shockley, and E. Fosler-Lussier. 2006. Sentence planning for realtime navigational instruction. In *Companion Volume to Proceedings of HLT-NAACL 2006*, pages 157–160, June.

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.

Corpus-based evaluation of Referring Expressions Generation

Albert Gatt and **Ielka van der Sluis** and **Kees van Deemter**
Department of Computing Science
University of Aberdeen
{agatt,ivdsluis,kvdeemte}@csd.abdn.ac.uk

1 Introduction

Corpus-based evaluation of NLP systems has become a dominant methodology. Typically, some metric is invoked to evaluate the results produced by a system against a 'gold standard' represented in the corpus. Despite growing recognition of the importance of empirical evaluation in NLG, resources and methodologies for evaluation of Generation of Referring Expressions (GRE) are in their infancy (but c.f. Viethen and Dale (2006)), although this area has been studied intensively since the publication of the Incremental Algorithm (IA) by Dale and Reiter (1995). This contribution describes some of the difficulties which inhere in any corpus-based evaluation exercise involving GRE, as well as a methodology to create a corpus aimed at overcoming these difficulties.

GRE is a **semantically intensive** task. Given an intended referent, a GRE algorithm searches through a knowledge base (KB) to find a combination of properties that uniquely identifies the referent. In order to apply the 'human gold standard' rationale of a corpus-based evaluation to this task, the corpus in question must satisfy at least the following:

1. **Semantic transparency**:
   (a) The domain knowledge of authors in the corpus must be known in advance, so that the algorithm is exposed to exactly the same knowledge. Deviations from such knowledge by humans must be clearly indicated.
   (b) If it is 'standard' GRE that is being evaluated, where output is a semantic or logical form, the corpus should permit the compilation of a normalised logical form from the human data (i.e., abstract away from variations in syntactic and lexical choice).
2. **Pragmatic transparency**:
   (a) If it is 'standard' GRE that is being evaluated, then the communicative intentions of authors in the corpus must be restricted as far as possible to the *identification* intention.
   (b) The communicative situation in which descriptions are produced must be controlled. For instance, a fault-critical situation might elicit more informative descriptions than a non-fault-critical one, which would affect the performance of algorithms in the evaluation.

The rest of this contribution describes our methodology to construct and annotate the TUNA Reference Corpus (TRC). Since its introduction in van Deemter *et al.* (2006a), the TRC has been completed, and consists of ca. 1800 descriptions with annotations about domain knowledge, semantics, and some aspects of communicative context.

2 A corpus for GRE

The TRC was constructed by eliciting descriptions of objects in a controlled experiment, conducted over the internet over a period of three months. The structure of the corpus is shown below, with reference to the experimental conditions manipulated.

	+FC		-FC		
domain	**sing**	**plur**	**sing**	**plur**	**total**
household	210	390	105	195	900
photographs	180	360	90	180	810

Subjects interacted with a computer system and referred to objects in domains where the precise combination of properties that was minimally required to identify the objects was known in advance. Two domains were used, one consisting of artificially constructed pictures of household items, the other of real photographs of people. It was made clear to subjects that they had to identify objects for the system, which in turn 'interpreted' their description and removed objects from the screen. Some of the subjects were placed in a fault-critical situation (+FC) and were told that the system was being tested for use in critical situations where errors could not be corrected; for the other, non-fault-critical situation (-FC), subjects were given the opportunity to correct the system's mistakes by clicking on the correct targets. Descriptions were to both singular and plural referents, and also varied in whether or not subjects could use locative expressions.

The corpus is fully annotated in an XML representation designed to meet the four desiderata outlined above; see (van Deemter et al., 2006b) for details. Description are paired with an explicit domain representation (entities and their attributes) which also indicates the communicative situation (±FC). Domain properties are tagged with an `ATTRIBUTE` tag, which takes a `name` and a `value`. The logical form of a description is indicated by means of a `DESCRIPTION` tag. An example of the annotation for the description *the small desk and the red sofa* is shown below.

```
<DESCRIPTION NUM='PLURAL'>
<DESCRIPTION NUM='SINGULAR'>
<DET value='definite'>the</DET>
<ATTRIBUTE name='size' value='small'>small</ATTRIBUTE>
<ATTRIBUTE name='type' value='desk'>desk</ATTRIBUTE>
</DESCRIPTION>
and
<DESCRIPTION NUM='SINGULAR'>
<DET value='definite'>the</DET>
<ATTRIBUTE name='colour' value='red'>red</ATTRIBUTE>
<ATTRIBUTE name='type' value='sofa'>sofa</ATTRIBUTE>
</DESCRIPTION>
</DESCRIPTION>
```

Using the `DESCRIPTION` tag, a logical form can be compiled by the recursive application of a finite set of rules. Thus, `ATTRIBUTE`s within a `DESCRIPTION` are conjoined; sibling `DESCRIPTION`s are disjoined. Attribute names and values are normalised to match those in the domain, irrespective of the wording used by an author. For example, the above annotation is compiled into $(small \wedge desk) \vee (red \wedge sofa)$.

3 GRE Evaluation

We have used the corpus to conduct an evaluation of the IA against some earlier algorithms, whose perceived shortcomings the IA was designed to address (Gatt et al., In preparation). Logical forms compiled form human-authored descriptions were compared to those generated by an algorithm within the same domain.

Because domain properties are known, human-algorithm comparisons can be based on various metrics, for example, (dis-)similarity of sets of attributes using metrics such as some version of of edit distance or the Dice coefficient. Moreover, the design of an evaluation study can vary. For instance, it is possible to compare an algorithm to a single subject in the corpus, or to an average of all descriptions in the corpus. Overall, a corpus built in line with the requirements outlined in this paper will provide the possibility of more refined algorithm evaluations compared to those conducted in the past. We plan to make this corpus available to the research community in the near future.

References

R. Dale and E. Reiter. 1995. Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(8):233–263.

A. Gatt, I. van der Sluis, and K. van Deemter. In preparation. Assessing algorithms for the generation of referring expressions, using a semantically and pragmatically transparent corpus.

K. van Deemter, I. van der Sluis, and A. Gatt. 2006a. Building a semantically transparent corpus for the generation of referring expressions. In *Proc. INLG-06.*

K. van Deemter, I. van der Sluis, and A. Gatt. 2006b. Manual for the tuna corpus: Referring expressions in two domains. Technical report, University of Aberdeen.

J. Viethen and R. Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proc. INLG-06.*

Position Statement for Workshop on STEC in NLG

Nancy L. Green
University of North Carolina at Greensboro
nlgreen @ uncg.edu

According to the call for participation of this workshop, a shared-task evaluation campaign (STEC) is a competitive approach to research funding where "different approaches to a well-defined problem are compared based on their performance on the same task". The proposed benefits of this approach are "enhancing the wider NLP community's view of work in NLG, and in providing a focus for research in the field". These benefits may not justify the risk.

NLG researchers should be careful that promoting STECs does not have a reductionist effect on the field and does not lead to the marginalization of other important NLG research areas. As many have noted, NLP starts with a well-defined input – text; while NLG does not. Thus, it is possible for the NLP community to use STECs to attack certain well-defined problems, e.g. in text classification, without having to first solve harder computational problems such as understanding all the nuances of meaning in a text. (However, even some computational linguists have complained that this trend in NLP has resulted in neglect of other key research.)

The starting point of an NLG system is not as well-defined since it is often non-linguistic, e.g., a Bayesian network for tumor classification used in an existing decision support system, a database about museum artifacts, or quantitative data requiring further computational analysis to detect trends and other significant features. A STEC providing common inputs might enable researchers to focus on problems in subsequent stages of the NLG "pipeline". However, use of a common starting point in the STEC may limit the general applicability of the solutions. Also, it may result in decreased support for NLG research on "what to say", e.g., reasoning required of an animated agent designed to engage in persuasive conversation with a user about the user's diet; deciding what to say may require not only nutrition information and dialogue history, but also a model of emotion and argument schemes.

At the other end of the NLG pipeline, application-independent research on how variation in surface generation, rhetorical features, and physical presentation features (such as layout) influences communicative effectiveness is needed. While several NLG systems for generating text variants for use in this kind of experimentation have been developed, the experimentation itself does not fit into a STEC funding model. A STEC could show that one generated result was more successful for a particular task than that of the competitors, but would not address the more fundamental questions whose answers could inform design of many different systems. Also, it is not clear how the narrow focus of a STEC could support the multi-disciplinary research required for multimedia generation, i.e., generation of integrated text and paralinguistic features (speech and gesture) or graphics (pictures, maps, diagrams, data graphics). It would be a mistake to limit the scope of NLG research to the medium of print.

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?

Flexibility counts more than precision

Position paper for the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation

David McDonald

BBN Technologies
Cambridge, MA 02138 USA
dmcdonald@bbn.com

Abstract

Today's NLG efforts should be compared against actual human performance, which is fluent and varies randomly and with context. Consequently, evaluations should not be done against a fixed 'gold standard' text, and shared task efforts should not assume that they can stipulate the representation of the source content and still let players generate the diversity of texts that the real world calls for.

1 Minimal competency

The proper point of reference when making an evaluation of the output of a natural language generation (NLG) system is the output of a person. With the exception of the occasional speech error or other predicable disfluencies such as stuttering or restarts, people speak with complete command of their grammar (not to mention their culturally attuned prosodics), and with complete command of their discourse context as it shapes the coherence of what they say and the cohesion of how they say it.

Any NLG system today that does not use pronouns correctly (assuming they use them at all), that does not reduce complex NPs when they describe subsequent references to entities already introduced into the discourse, that does not reduce clauses with common subjects when they are conjoined, or that fails to use any of the other ordinary cohesive techniques available to them in the language they are using is simply not in the running. Human-level fluency is the entrance ticket to any comparative evaluation of NLG systems.

2 Real sources

Similarly, any system that started from a hand-made source representation (as we all did in the 1970s) would not be meeting the minimal standards by which we should measure an NLP system today. Any proposal for a shared evaluation campaign should provide source representations that reflect real data used to do real work for real (preferably commercial) systems.

A good example of a class of real sources is minimally interpreted numerical data sources such as raw instrument readings for weather reports (SumTime) or data points in the movement of stock averages during a day of trading (Kukich 1988). I will propose a more versatile source later.

3 Variation is expected

When I read Winnie-the-Pooh to my daughter at bed time what comes out of my mouth is not always what was in the book, though it always carries the same message. Overworked phases aside, people rarely phrase their content the same way time after time even when they are talking about something they know very well.

This natural level of variation that people exhibit is something that our NLG systems should do as well. It is the only way, for example, that a syn-

thetic character in a computer game that incorporated a proper NLG system would ever be seen as realistic, which is crucial in game-based training systems where suspension of disbelief is required if the training is to be effective.

4 Context is everything

Consider these passages that I clipped from today's news.[1] The first is the title pointing to the full article and was positioned next to a graphic. The second was the small blurb that summarized the content of the article. The third is the equivalent text close to the top of the full article. If we looked at Apple's press release or its quarterly earnings report that prompted this BBC article we would see still different phrasings of this same content.

"Apple profits surge on iPod sales"

"Apple reports a 78% jump in quarterly profits thanks to strong Christmas sales of its iPod digital music player."

"Apple has reported a 78% surge in profits for the three months to 30 December, boosted by strong Christmas sales of its iPod digital music player."

From the point of view of the source representation that a NLG system would use, these three texts are arguably based on the identical content. Some leave out details, others choose different phrasing. What drives the differences is the purpose that the text serves—the context in which is will be used—a flashy title to catch the eye; a short summary; the lead in to a full write up.

5 Where does flexibility come from?

As these example show, a good generator will be sensitive to its context and adapt what it produces accordingly. Still, other than things like freely varying choices of synonyms and semantically neutral variations in linguistic constructions that could be governed by genuinely random 'decisions', most NLG systems prefer to have rationales behind their choices, whether they are the design of the features sets that govern statistical systems or symbolic rules. Where are the rationales for such widely varying surface forms going to come from, and how might they be incorporated in a common data set for evaluation?

I don't believe that we know the answer to this question yet other than that is has something to do with the set and setting deep within the computational entity for whom the generator is working. This calls for research on the kinds of representations that initiate and drive generation and how they encode teleology and psychological motive. No two researchers are likely to agree on what this representation looks like, and for texts like these examples it cannot be reduced to numerical data.

Let me suggest that a clean way to handle this problem is to make the shared data set be *the texts themselves*, with their settings, and to let the players construct whatever representation they want by *parsing* them. Taking the interpretations back far enough to identify a common core content among a set of different texts that are stipulated by a consensus of judges to be conveying essentially the same content should provide some insight into the reason for the difference that just starting from the generation direction would not.

Parsing and regenerating is also a worthy problem in its own right. There is a vast wealth of information that is only available as texts, and DARPA and others are actively developing efforts in 'learning by reading'. I believe that a natural sweet spot for commercial generation work in the future (besides the game world) is in regenerating a common body of content in different genres and with different functions, just as human journalist does after reading a press release. If we can take up this problem collectively as part of a shared task, so much the better.

References

Karen Kukich, 1988. *Fluency in Natural Language Reports*. McDonald & Bolc (eds.) Natural Language Generation Systems. Springer-Verlag series in Symbolic Computing

SumTime: "Generating English Summaries of Time-Series Data. http://www.csd.abdn.ac.uk/~ereiter/sumtime.html

[1] BBN News, 17 January 2007.

NLG Evaluation: Let's open up the box

Chris Mellish
Computing Science
University of Aberdeen
Aberdeen AB24 3UE, UK
cmellish@csd.abdn.ac.uk

Donia Scott
Centre for Research in Computing
The Open University
Milton Keynes, MK7 6AA, UK
d.scott@open.ac.uk

Abstract

There is a spectrum of possible shared tasks that can be used to compare NLG systems and from which we can learn. A lot depends on how we set up the rules of these games. We argue that the most useful games are not necessarily the easiest ones to play.

The Lure of End-to-End Evaluation

Mellish and Dale (1998) discuss a number of different approaches to NLG system evaluation that had been used by 1998. Systems can be evaluated, for instance, in terms of accuracy, fluency or in their ability to support a human task. Independent of this is the question as to whether evaluation is *black box* or *glass box*, according to whether it results in an assessment only of the complete system or also of its contributing parts.

End-to-end evaluation is black box evaluation of complete NLG systems. It involves presenting systems with "naturally occurring" data and evaluating the language produced (according to accuracy, fluency, etc.). End-to-end evaluation is a tempting way to start doing NLG evaluation, because it imposes minimal constraints on the structure of the systems. Therefore as many people as possible can take part. This is important, because at the beginning critical mass is needed for things to "take off".

The Dangers of End-to-End Evaluation

Unfortunately there are dangers in using an end-to-end task as the basis of comparative NLG system evaluation:

- Danger of overfitting the task. The best systems may have little to say about language in general, but may encode elaborate stimulus-response type structures that work for this task only.
- Lack of generalisability. The best systems may have nothing to say about other NLG tasks. Or the way that systems are presented/ compared may prevent researchers in nearby areas from seeing the relevance of the techniques. So you may actually end up attracting *fewer* interested people.

Opening the box

End-to-end evaluation emphasises a "black box" approach that ignores what the NLG systems are doing inside. And yet we have some good ideas about the general tasks carried out in NLG (e.g., lexical choice, referring expression generation, aggregation) and it is at this level that we exchange knowledge at conferences and the field progresses independent of particular applications.

Opening the box for NLG evaluation would be analogous to the move in the MUC conferences from a unitary task to a set of much more structured subtasks. This was able to make MUC much more interesting to people involved in, for instance, named entity recognition and anaphora resolution. It also helped to bridge the large disconnect between 'success' in the MUC competition and 'progress' in the field of NLP.

Perhaps NLG evaluation could start simple and

progress in a similar way, moving in time from *application*-tasks to *NLG*-tasks. But without the significant funding that initiatives like MUC have had access to, it might well never make it beyond the first step.

How to Start?

How can we design evaluation tasks that stretch NLG systems in interesting ways? We need to have an agreement on which subtasks of NLG are of general interest and we need to have an agreement about what their inputs and outputs look like. This relies on a degree of theoretical convergence — something that the NLG field is not renowned for.

In this context, it is relevant to review whether RAGS (Mellish et al., 2006) might provide a good basis for defining tasks which would evaluate NLG systems, components and algorithms in a meaningful way.

RAGS

RAGS (Reference Architecture for Generation Systems) was an attempt to exploit previous ideas about common features between NLG systems in order to propose a reference architecture that would help researchers to share, modularise and evaluate NLG systems and their components without having to commit to particular theoretical approaches or implementational requirements. In practice, the project found that there was less agreement than expected among NLG researchers on the modules of an NLG system or the order of their running. On the other hand, there was reasonable agreement (at an abstract level) about the kinds of data that an NLG system needs to represent, in passing from some original non-linguistic input to a fully-formed linguistic description as its output.

RAGS took as a starting point eight commonly-agreed low-level NLG tasks (lexicalisation, aggregation, rhetorical structuring, referring expression generation, ordering, segmentation and centering/salience), and provided abstract type definitions for six different types of data representations (conceptual, rhetorical, document, semantic, syntactic and "quote"). It produced and made available sample implementations of the RAGS technology and complete implementations of RAGS systems, along with some sample datasets.

The final product of the RAGS project is undeniably incomplete, and the framework itself is difficult to use — both practically (e.g., many find the type descriptions hard to understand) and conceptually (one is forced to make hard decisions about the data at hand, answering questions such as "is this conceptual or semantic?").

Moving forward

There is a sense in which RAGS was slightly ahead of its time. Were we to start again, it would be more sensible to cast RAGS in terms of the Semantic Web (Berners-Lee et al., 2001). This would allow us to take advantage of the Web Ontology Language (OWL) (Antoniou and van Harmelen, 2003) and a great deal of technical infrastructure that has developed independently of, and in parallel to, RAGS.

We have begun to re-cast RAGS in terms of OWL, but this is still at an early stage. When complete, this work will help NLG researchers to use RAGS for the purpose for which it was intended: making it easier to create reusable data resources, communicate data between program modules, and allow modules (or at least their inputs and outputs) to be defined in a relatively formal way. This should make RAGS more useful for defining "glass box" evaluations of NLG systems.

This will not, of course, mean that evaluation would be an *easy* game to play; but, the game would be much more *meaningful*. And probably a lot more fun.

References

Grigoris Antoniou and Frank van Harmelen. 2003. Web Ontology Language: OWL. In S. Staab and R. Studer, editors, *Handbook on Ontologies in Information Systems*. Springer-Verlag.

T. Berners-Lee, J. Hendler, and O. Lassila. 2001. The semantic web. *Scientific American*, 284(5):35–43.

C. Mellish and R. Dale. 1998. Evaluation in the coatext of natural language generation. *Computer Speech and Language*, 12:349–372.

Chris Mellish, Donia Scott, Lynne Cahill, Daniel Paiva, Roger Evans, and Mike Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12(1):1–34, March.

NLG Systems Evaluation: a framework to measure impact on and cost for all stakeholders

Cécile Paris [a] **Nathalie Colineau** [a] **Ross Wilkinson** [b]

CSIRO – ICT Centre

[a] Building E6B Macquarie University Campus, North Ryde NSW 2113, Australia

[b] Computer Science & Information Technology Building, ANU Campus, Acton ACT 2601, Australia

{cecile.paris, nathalie.colineau, ross.wilkinson}@csiro.au

1 Enlarging the view of evaluation

The weaknesses of most current evaluation methods is that the conclusions are based not on whether a system performs as expected and on the consequences of its deployment, but on how well it scores against references. In other words, systems are mostly evaluated on some properties (in particular, the "accuracy" of their output), but hardly ever on their ability to fulfil the purpose for which they have been developed and their impact on their (various) users. We argue here that a better way to look at NLG system evaluation would consist in determining the effectiveness of the whole system – not simply its correctness under particular conditions.

Another major drawback of current evaluation practices is to look at only one side of the equation: the benefit. We believe that both the cost and the benefit of the system are important to decide on a system's success.[1] While there is clearly a recognition that there are costs involved, in particular, in obtaining the various resources required (e.g., domain models, task models) – as evidenced by the number of tools developed to help author complex knowledge bases (Power & Scott, 1998; Paris *et al.*, 2005; Androutsopoulos *et al.*, in press) – these costs are typically not measured and not taken into account when evaluating a system. Similarly, the trend towards common evaluation metrics and competitive evaluation tasks does not account for the cost incurred to fine-tune systems for years – a cost also pointed out in (Scott & Moore, 2006). The actual benefit of the improvements may be questionable compared to the cost incurred (e.g., time and effort involved). The benefit-cost tradeoffs (the "bang for buck") are important if we want technology to be adopted and potential users to make an informed choice as to what approach to choose when.

In addition, competitive evaluation tasks often decontextualise systems from their real use by setting artificial tasks. We argue that the context in which a system's effectiveness is evaluated is fundamental – a system exhibiting the 'best performance' might not be the best for a specific task as other task characteristics may be more important.

In this position paper, we consider an NLG system in the context of its stakeholders, their goals and tasks, and the information sources that the system requires. We propose an evaluation framework that allows for all the stakeholders, capturing who benefits from the system and at what cost.

2 A Comparative Framework for Measuring the Effectiveness of NLG Systems

As mentioned in (Paris *et al.*, 2006), and building on work from management and information system, e.g., (Mclean & Delone, 1992; Cornford *et al.*, 1994), we need to enlarge our view of evaluation and identify for each stakeholder role a set of benefits and costs that should be considered. As a first step, we have identified four main stakeholder roles, and, for each, what to evaluate, what questions to ask, as illustrated in Table 1:

– The information *consumer*. The person(s) who will use the generated text.

[1] It might even be useful to look at benefits and costs of a *proposed* system to determine whether it is worth developing and deploying.

	Information Consumer	Information Provider	Information Intermediaries	System Provider
Benefits	Task effectiveness Knowledge gained Satisfaction	Audience reach Audience accuracy Message accuracy	Ease of knowledge creation Ease of context modelling	System usage Reliability Response time Correctness
Costs	Time to complete the task Cognitive load Learning time	Metadata provision Structured information Currency of data	Time to create and integrate the resource Time to capture contextual characteristics	Implementation cost (hardware and software) System maintenance System integration

Table 1. Comparative framework for NLG systems' stakeholders

– The information *provider*. The person(s) (or organisations) with a message to convey. When the generated text is composed of existing text fragments, this person is responsible to provide the content. If the text is generated from first principles, the provider is responsible for the goal(s) and message(s) to be conveyed.

– The information *intermediaries*. They work prior to generation time to create the appropriate set of resources needed by the system (e.g., grammar, lexicon, domain and user models, or potentially text fragments).

– The system providers. They are responsible for the development and maintenance of the technology.

This framework provides us with a context to evaluate different approaches and systems. Given a system (approach) and purpose, the framework forces us to think explicitly about the stakeholders involved, their needs and expectations, how the system meets these and at what cost. This guides us with respect to what experiment(s) to conduct (e.g., test response time or satisfaction of consumers). Ideally, one would want to conduct experiments for each cell in the table. Realistically, we need to identify our priorities for a specific system and carry out the relevant experiments. The results then gives us a way to decide whether the system is worth adopting (developing), given the specified priority(ies) for a given situation (e.g., optimising the benefits to the provider, in particular accuracy of message *vs.* minimising the cost to the intermediary). Note that, the benefits and costs measures might be of a qualitative nature only (e.g., the type of changes required for maintenance and the expertise needed).

When we compare systems within this framework, we do not need the same input and output. What is important is the priority(ies) at stake. In addition, the point is not to average results across the table. Instead, the priorities tell us how to interpret the results. Finally, the framework is not defined around any specific task but can be used to evaluate systems developed for different tasks, given their respective priorities. Note that this approach is whole-of-system oriented.

To conclude, we believe we need to enlarge the view of evaluation, adopting a "consumer-oriented product review" type of evaluation (i.e., whole-of-system), and explicitly thinking of the "bang-for-buck" equation. We have adopted this approach in our own work.

References

Androutsopoulos, I, Oberlander, J., and Karkaletsis, V., in press. Source Authoring for Multilingual Generation of Personalised Object Descriptions. *Natural Language Engineering*, Cambridge University Press.

Cornford, T, Doukidis, G.I. & Forster, D., 1994. Experience with a structure, process and outcome framework for evaluating an information system, *Omega, International Journal of Management Science*, 22 (5), 491-504.

DeLone, W. H. & McLean, E. R., 1992. Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*, 3(1), 60-96.

Paris, C., Colineau, N, Lu S. and Vander Linden, K. 2005. Automatically Generating Effective Online Help. *International Journal on E-Learning*, Vol.4, No.1, 2005. 83-103.

Paris, C., Colineau, N. and Wilkinson, R. 2006. Evaluations of NLG Systems: common corpus and tasks or common dimensions and metrics?. In *Proc. of INLG-06*, held as a workshop on the COLING/ACL Conference, Sydney, Australia, July 15-16. 127-129

Power, R. and Scott, D. 1998. Multilingual authoring using feedback texts. In *Proc. of COLING-ACL 98*, Montreal, Canada.

Scott, D. and Moore, J., 2006. An NLG evaluation competition? Eight Reasons to be Cautious. Technical Report 2006/09, Department of Computing, The Open University.

NLG Shared Tasks: Lets try it and see what happens

Ehud Reiter
Department of Computer Science
University of Aberdeen
Aberdeen AB24 3UE, UK
ereiter@csd.abdn.ac.uk

1 Pros and Cons of Shared Tasks

I must admit that I have mixed feelings about shared task evaluations. Shared task evaluations of course have many advantages, including allowing different algorithms and approaches to be compared, producing data sets and evaluation frameworks which lower the "barriers to entry" to a field, and more generally getting researchers to interact more, and realise how their assumptions about inputs, outputs, knowledge sources, and processing constraints differ from those made by other researchers.

Shared task evaluations could also help us understand evaluation better. I would like to get a better idea of how different evaluation techniques (such as statistical evaluation, human preference judgements, and human task performance) correlate with each other. In order to carry out such studies, it would be very useful to have a number of systems with similar input/output functionality and knowledge sources; a shared-task evaluation could provide these systems (Reiter and Belz, 2006).

On the other hand, there are also dangers to shared tasks. In particular, focusing on a shared task can cause a community to narrow the scope of what it investigates. For example, colleagues of mine in the Information Retrieval community have suggested to me that the academic IR community's focus on the TREC shared evaluation in the mid and late 1990s limited its contribution to web search when this emerged as the "killer app" of IR. This is because the 1990s academic IR community had little interest in web-search algorithms (such as Google's page rank) which could not be used in TREC shared tasks.

In other words, TREC encouraged the IR community to focus on one specific type of IR problem, and probably helped it make progress in this area. But this was at the cost of ignoring other types of IR problems, which turned out to be more important.

My personal opinion is that we should try to organise some shared task evaluations in NLG, but do this (at least in the first instance) as one-off exercises. I think a yearly "NLGUC" event would be a mistake; but I think one-off shared evaluations could be worthwhile and should be tried.

2 Issue: Topic

From a practical perspective, I suspect that the main challenges in running an NLG shared evaluation are going to be (1) choosing a topic that attracts enough participants to make the exercise meaningful, and (2) deciding how to evaluate systems. Looking at the topic issue first, the NLG community is quite small (recent International NLG conferences have attracted on the order of 50 people), and the NLG problem space is enormous. Since a shared task evaluation must focus on specific NLG problem(s), it is not easy to find a topic which would attract a reasonable number of participants (at least 6, say).

One possible topic that could attract this number of people is generating referring expressions. This has attracted a lot of attention in recent years; for example in INLG 2006 there were papers on this topic from groups in Australia, Brazil, Germany, Japan, UK, and USA. There are also some corpora available which could be used for a reference generation shared task, such as Coconut (Jordan and Walker, 2005) and the Tuna corpus (van Deemter et

al., 2006).

Another possibility, which focuses on an application instead of on an NLG task, is generating weather forecasts. This has been one of the most popular NLG applications over the past 20 years; Bateman and Zock's list of NLG systems[1] (which is not complete) lists 13 systems in this area. And there are corpora available, such as the SumTime corpus (Sripada et al., 2005).

A third possibility is medical, in particular patient information. Medical applications of NLG are popular according to Bateman and Zock's list, and there are many people outwith the NLG community who are interested in generating personalised health information; indeed there are workshops on this topic. However, I suspect it would be harder to organise a shared task evaluation in this area because data resources would need to be created (I'm not aware of any existing corpora in this area).

3 Issue: Evaluation

Another challenge in organising a shared task evaluation is deciding how to evaluate the systems. I believe that most shared task evaluations in Language Technology use corpus-based evaluation, but this can be controversial, not least because corpus-based evaluation metrics seem to be biased towards systems built using corpus-based techniques (Belz and Reiter, 2006). In NLG in particular, it is clear that writers do not always produce optimal texts from the perspective of readers (Oberlander, 1998; Reiter and Sripada, 2002); this is another argument against using metrics which compare machine-generated texts to human written texts.

But reader-based evaluations have problems as well. The easiest kind to carry out is rating exercises, where human subjects are asked to rate the quality of generated texts. However, we know that in many cases such ratings are not good predictors of how useful texts actually are in helping real users carry out real tasks (Law et al., 2005). Task-based evaluations are more robust in this sense, but they are expensive and time-consuming, and we have no guarantees that texts that are useful in supporting one task will also be useful in supporting other tasks.

[1] http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.htm

Given this uncertainty, I think any shared task evaluation should use a number of different evaluation techniques. Indeed, as mentioned above, I think one of the goals of a shared task evaluation should be to get empirical data on how well different evaluation metrics correlate with each other, so that discussions about evaluation techniques can be informed by real data.

The other advantage of multiple evaluation techniques is that it makes it harder to say who "won" a shared task evaluation. This is good, because I think the NLG community will be more willing to participate in shared task evaluations if they are primarily seen as scientific ventures instead of as contests.

References

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *Proceedings of EACL-2006*, pages 313–320.

Pamela Jordan and Marilyn Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Anna Law, Yvonne Freer, Jim Hunter, Robert Logie, Neil McIntosh, and John Quinn. 2005. Generating textual summaries of graphical time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing*, 19:183–194.

Jon Oberlander. 1998. Do the right thing ... but expect the unexpected. *Computational Linguistics*, 24:501–507.

Ehud Reiter and Anja Belz. 2006. Geneval: A proposal for shared-task evaluation in nlg. In *Proceedings of INLG 2006*.

Ehud Reiter and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceedings of the Second International Conference on Natural Language Generation*, pages 97–104.

Somayajulu Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2005. SUMTIME-METEO: Parallel corpus of naturally occurring forecast texts and weather data (revised 2005 edition). Technical Report AUCS/TR0201, Computing Science Dept, Univ of Aberdeen, Aberdeen AB24 3UE, UK.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of INLG 2006*.

Evaluation in Natural Language Generation: The Question Generation Task

Vasile Rus
Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
vrus@memphis.edu

Zhiqiang Cai
Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
zcai@memphis.edu

Arthur C. Graesser
Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
a-graesser@memphis.edu

Abstract

Question Generation (QG) is proposed as a shared-task evaluation campaign for evaluating Natural Language Generation (NLG) research. QG is a subclass of NLG that plays an important role in learning environments, information seeking, and other applications. We describe a possible evaluation framework for standardized evaluation of QG that can be used for black-box evaluation, for finer-grained evaluation of QG subcomponents, and for both human and automatic evaluation of performance.

1 Introduction

Natural Language Generation (NLG) is one of the grand challenges of natural language processing and artificial intelligence (Dale et al., 1998). A robust NLG system requires the modeling of speaker's intentions, discourse planning, micro-planning, surface realization, and lexical choices. The complexity of the task presents significant challenges to NLG evaluation, particularly automated evaluation. Major progress towards standardized evaluation exercises of NLG systems will be achieved in shared-task evaluation campaigns (STEC) that are planned over a number of years. They start with simple (sub)tasks in the early years that invite wide participation by various research groups and then gradually increase the difficulty of the problems addressed. The selected shared task should minimize restrictions on alternative approaches. For instance, the test data should not be specified in representations that are favored by particular systems and researchers. The task should also allow evaluation of different aspects of NLG and should be relevant to a variety of applications.

We propose an evaluation framework for the task of Question Generation (QG). QG is defined as a task with simple input and output. The framework accommodates black-box evaluation of alternative approaches and finer-grained evaluation at micro-planning, surface realization, and lexical choice levels. The initial task is extendable to permit evaluation at all levels, including speaker's intentions and discourse planning. QG is an essential component of learning environments, help systems, information seeking systems, and a myriad of other applications (Lauer et al., 1992). A QG system would be useful for building an automated trainer for learners to ask better questions and for building better hint and question asking facilities in intelligent tutoring systems (Graesser et al., 2001). In addition to learning environments, QG facilities could help improve Question Answering systems by launching questions proactively and jumping in with suggested queries when dead-ends in inquiry inevitably occur.

QG as a testbed can benefit from previous experience on standardized evaluations of related shared tasks in Question Answering (TREC-Question Answering track; http://trec.nist.gov) and from evaluations of Intelligent Tutoring Systems such as AutoTutor (Graesser et al., 2001). Data sources from those previous shared tasks can be easily adapted to a QG task with relative efficient costs.

This paper defines the task of QG, briefly describes the QG evaluation framework, and presents evaluation metrics.

2 The Question Generation Task

Our approach to QG assumes that there are one or more sentences (i.e., possible answers to a user question) given as input, whereas the task of a QG approach is to generate questions related to this input. This textual specification of both input and output should encourage wide adoption of the task by many research groups because it does not impose any representational restrictions on the input or output. Various approaches can of course use their own internal representations for input. The input is limited to 1-2 sentences to simplify the task and minimize complexities of discourse level processing. The task can eventually be extended to incorporate discourse by specifying a paragraph as input and asking for a set of related questions as output.

Two data sources are available to extract input and output data. Both consist of a set of sentences and each sentence's associated human-generated questions. The first one is Auto-Tutor (Graesser et al., 2001), an Intelligent Tutoring System that holds dialogues with the learner in natural language. For each input sentence taken from such dialogues, there is an associated set of questions. The second source is the TREC Question Answering track, where thousands of Question-Answer pairs are available from Question Answering evaluations since 1999. In this case, for each sentence (answer) we have a single associated question.

The input (Expectation, Answer) and output data (Questions) are sufficiently well formulated to make the setup of such standardized evaluation quick and easy. The researcher community can target specific feature evaluations of generation systems. For example, by selecting sentences with associated Who? or What person? questions from the TREC QA source, one can focus on testing the capabilities of a system for generating person-related questions. Similarly, one can select sentence-question pairs tailored to the evaluation of lexical choice characteristics of a generation system.

3 Evaluation

The output of a QG system can be evaluated using either automated evaluation or manual evaluation. Automated evaluation can use methods similar to ROUGE in summarization and BLEU/NIST in machine translation which are based on N-gram co-occurrence. An extreme solution is to consider exact question matching in which the generated question and the expected question in the gold standard, containing the ideal/expected questions, have to be identical for a hit. Manual evaluation recruits experts to assess the output of various approaches along different criteria.

The evaluation of any NLG system includes multiple criteria, such as user satisfiability, linguistic well-foundedness, maintainability, cost efficiency, output quality, and variability. Other metrics can serve as proxies for some criteria. For example, precision may be a proxy for user satisfiability. In a recent study (Cai et al., 2006), our group used *precision* and *recall*. *Precision* is the proportion of good questions out of all generated questions. *Recall* or coverage is difficult to objectively compute because the number of questions generated from a sentence is theoretically indeterminate. A recall measure can be observed in specific experiments. In the TREC QA data set, there is only one question for each each answer. *Recall* would be the proportion of those TREC QA questions that are present in the output of a QG system.

References

Z. Cai, V. Rus, H.J. Kim, S. Susarla, P. Karnam, and A.C. Graesser. 2006. NLGML: A natural language generation markup language. In T.C. Reeves and S.F. Yamashita, editors, *Proceedings of E-Learning Conference*, pages 2747–2752, Honolulu, Hawaii. AACE.

Robert Dale, Donia Scot, and Barbara di Eugenio. 1998. Special Issue on Natural Language Generation. *Computational Linguistics*, 24(3):346–353, September.

Arthur C. Graesser, Kurt VanLehn, Carolyn P. Rose, Pamela W. Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22(4):39–52.

T.W. Lauer, E. Peacock, and A.C. Graesser. 1992. *Questions and Information Systems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

An NLG evaluation competition? Eight Reasons to be Cautious

Donia Scott
Centre for Research in Computing
The Open University, U.K.
FirstInitial.Lastname@open.ac.uk

Johanna Moore
Human Communication Research Centre
The University of Edinburgh, U.K.
FirstInitial.Lastname@ed.ac.uk

Abstract

It is our view that comparative evaluation of the type used in MUC, TREC, DUC, Senseval, Communicator, may not be sensible for NLG and could be a misguided effort that would damage rather than help the field.

Most would agree that NLG has to date failed to make as significant impact on the field of NLP and on the world–as measured by the number publications, existing commercial applications, and the amount of funding it has received. While it may be useful to look at other subfields of NLP (e.g., message understanding, machine translation, summarization, word sense disambiguation) and speculate why this should be the case, we urge caution in proceeding under the assumption that a good path to progress in NLG would be to jump on the evaluation competition bandwagon.

All that glitters is not gold: For evaluation competitions to have much meaning, there has to be a gold standard to aspire to. With a clearly defined input and a fully-specified output, one may be able to establish a reasonable criterion for success that can be applied to all competitors. In the case of NLG, this is extremely hard to achieve–some may say impossible–without distorting the task to a degree that renders it otiose.

What's good for the goose is not necessarily good for the gander: NLG systems have been, and continue to be built to serve a wide range of functions. It makes little sense to compare the output of systems that are designed to fulfull different functions, especially since the most important criterion for any system is its "fitness for purpose". NLG, unlike MT and parsing, is not a single, well-defined task but many, co-dependent tasks.

Don't count on metrics: Both the summarization and the MT communities, who have for several years been working towards shared metrics, are now questioning the usefulness of the metrics. For the past 3–4 years, to claim that one has made progress in MT, one simply needed to report an increase in BLEU score. Yet in the past year, there have been several papers published decrying the usefulness of BLEU (e.g., Callison-Burch *et al.* (2006)), and showing that it does not correlate well with human judgements when it comes to identifying high quality texts (despite prior reports to the contrary). Indeed, the recent word on the street is that BLEU should only be used as one of many metrics to tell if one is improving their own system, *not* as a metric to compare systems (Kevin Knight, invited talk, EACL 2006). Simply put: so-called 'quality metrics' often don't give you what you want, or what you think they give.

What's the input? The difference between NLU and NLG has been very aptly characterised as the difference between counting from one to infinity or from infinity to one (Yorick Wilks, invited talk, INLG 1990). A huge problem in NLG is that, quite simply, different applications have different input. But, even if we were to agree on a shared

task (and this is a huge problem in itself) such as producing reports of stock market activity, some would advocate starting with the raw data coming off the ticker tape, while others would say that the data analysis program needed to identify significant events in the data stream has nothing to do with the generation process. But surely the quality of data analysis will affect the quality of the text that is generated.

What to standardize/evaluate? So what can we hope to provide evaluation metrics for? Some would argue that realization is clearly an area for which we can provide standard metrics because surely we can all agree on what the input and output specification should be. But even here, there will be heated debate not only over what formalism to use, but what information must be specified in the input. For example, should the input to the realizer be required to include information structure? Should the output include markup for pitch accents and boundary tones (which is needed for high-quality speech synthesis)? If information structure is essential to your theory of how many generation choices are made, you will argue vehemently for it. But if it does not fit your theory or you don't have a content and sentence planner capable of producing the semantically rich input representation required, you will argue just as vehemently against it.

The plug-and-play 'delusion': One of the main selling points of the DARPA Communicator program was the idea of plug-and-play. It was intended to give researchers a full end-to-end dialogue system, in which they could test competing hypotheses about one component of a system (e.g., the parser, the dialogue manager, the response generator) without building all the other components. Great idea; horrific execution. Communicator specified a low-level agent communication architecture (Galaxy Communicator), *not* the interfaces between components of a dialogue system. The result was that the plug-and-play dream never came to fruition. And despite a large scale NIST evaluation of nine systems all performing the same task, many would claim that the dialogue community has learned virtually nothing about how to build better dialogue systems from this time-consuming and expensive exercise.

Who will pay the piper? The reason that ATIS, Communicator, BLEU, ROUGE, DUC, TREC, etc., made it past the coffee room is literally *millions* of U.S. dollars of research funding. If NLG hopes to get any momentum behind any evaluation initiative, there has to be a funder there to pay the bills. Who will do this, and why should they? Put another way: what's the 'killer app' for NLG in the Homeland Security domain?

Stifling science: To get this off the ground we have to agree the input to realization. And you can push this argument all the way up the NLG pipeline. And whatever we agree on will limit the theories we can test. So what is really needed is a theory neutral way of representing the subtask(s) of the generation process to be evaluated. If we cannot do this, we will stifle new and truly creative ideas that apply new advances in linguistics to the generation process.

We believe that a good starting point in being able to compare, evaluate and maybe even reuse NLG technologies could be for the community to engage with something like the RAGS initiative, which provides a language for describing the interfaces between NLG components (Mellish et al., 2006). We also think that the NLG community would benefit from becoming better versed in the experimental methods for conducting human evaluation studies. Until then, there is a real risk that too many people will engage in wasted efforts on invalid or irrelevant evaluation studies, and some good but unsexy evaluation studies will continue to be misunderstood.

References

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

C. Mellish, D. Scott, L. Cahill, D. Paiva, R. Evans, and M. Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12:1–34.

Pragmatic Influences on Sentence Planning and Surface Realization: Implications for Evaluation

Amanda Stent
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
amanda.stent@gmail.com

Abstract

Three questions to ask of a proposal for a shared evaluation task are: *whether* to evaluate, *what* to evaluate and *how* to evaluate. For NLG, shared evaluation resources could be a very positive development. In this statement I address two issues related to the what and how of evaluation: establishing a "big picture" evaluation framework, and evaluating generation in context.

1 Introduction

Recently, shared evaluation tasks have been used in IE, parsing, semantic role labeling, QA and MT. These shared tasks have resulted in new corpora, tools and performance metrics. Because NLG is a small field, shared evaluation resources could be a very positive development. However, we should avoid a common trap of shared evaluation tasks: a too-narrow evaluation framework and simplistic performance metrics leading to devaluing of interesting applications and research problems. In this statement, I address these two issues in turn.

2 An Evaluation Framework for NLG

We should avoid the urge to adopt shared evaluation tasks that unnecessarily limit NLG research. I propose a broad *shared evaluation framework* organized around the reference NLG architecture proposed in (Reiter and Dale, 1997). The framework has three dimensions:

Level	Selection	Organization
discourse	content selection	discourse planning
paragraph	discourse cue assignment	sentence aggregation
sentence	lexical selection RE generation	surface realization
media	media selection	media coordination

Table 1: Generation tasks

discourse type (e.g. summaries, explanations, comparisons), *application* (e.g. tutoring, question answering), and *generation task*. Generation tasks are further organized into task types (selection/organization) and levels (Table 1).

This framework could be used immediately, while the evaluation discussion continues. If we set up a wiki organized according to this (or another) framework, researchers could immediately start sharing evaluation resources such as corpora and tools. Shared evaluation tasks could be chosen from discourse type/ application/generation task triples for which data and/or multiple implementations exist (Reiter and Belz, 2006). Lessons learned from evaluations for one discourse type/application/generation task could be applied to other discourse types and applications. Instead of focusing research on one generation task, a shared framework could lead to more substantial and interesting evaluations in a variety of areas.

3 Evaluation in Context

High-quality generation makes heavy use of context information such as user models, discourse history, and the physical context of the dis-

course. For example, generation tasks affected by user preferences include content selection and ordering, media organization, and sentence aggregation (Reiter et al., 2003; Stent et al., 2004; Stent and Guo, 2005). I am particularly concerned about existing automatic evaluation metrics for surface realization (e.g. BLEU, NIST) because they do not take context into account. In particular, they: use a small number of reference outputs selected without regard to the generation context; conflate the measurement of fluency and adequacy (meaning preservation); and conceal rather than reveal the types of errors found. Consequently, it is difficult to do error analyses or compute the relative impact of system changes on output fluency and adequacy (Stent et al., 2005; Scott and Moore, 2006). This makes it hard to evaluate how context information affects system performance.

In the evaluation framework presented here, each generation task includes a subtask devoted to 'selection' and another devoted to 'organization'. Selection subtasks can be evaluated by information extraction-like metrics (a combination of counts of correct, missing and spurious elements giving precision and recall measures). These metrics give counts useful in error analysis. Ordering subtasks are harder to evaluate automatically. Traditionally, most ordering subtasks are performed using tree data structures (e.g. sentence plan trees), so tree edit distance metrics can be used (Bangalore et al., 2000). For automatic evaluations, human judges can select reference outputs taking context into account.

In our research on ordering tasks, we use human evaluations. The evaluator is presented with the generation context, then given randomly ordered possible outputs from different systems (including the reference sentence(s)). The evaluator ranks the possible outputs from best to worst, and separately notes whether each possible output is inadequate or ambiguous, disfluent or awkward. We use standard statistical methods to compare the systems contributing outputs to the evaluation, and can easily perform error analyses. We could contribute our evaluation tools to an evaluation wiki. With a shared evaluation, the human evaluation effort could be shared across sites and the cost to any particular research group minimized.

4 Summary

In the NLG community, recent efforts to provide shared evaluation resources (e.g. the SumTime corpus) should be encouraged. A shared evaluation framework should encourage the full range of NLG research.

Because generation output quality is dependent on context, generation output should be evaluated in context and evaluation metrics and tools should be developed that incorporate context, or at least facilitate error analyses to permit exploration of the impact of context.

References

S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of INLG*.

E. Reiter and A. Belz. 2006. GENEVAL: A proposal for shared-task evaluation in NLG. In *Proceedings of INLG Special Session on Sharing Data and Comparative Evaluation*.

E. Reiter and R. Dale. 1997. Building applied natural-language generation systems. *Journal of Natural-Language Engineering*, 3:57–87.

E. Reiter, R. Robertson, and L. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.

D. Scott and J. Moore. 2006. An NLG evaluation competition? eight reasons to be careful. In *Proceedings of INLG Special Session on Sharing Data and Comparative Evaluation*.

A. Stent and H. Guo. 2005. A new data-driven approach for multimedia presentation planning. In *Proceedings of EuroIMSA*.

A. Stent, R. Prasad, and M. Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of ACL 2004*.

A. Stent, M. Marge, and M. Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of CICLing*.

Automatic Evaluation of Referring Expression Generation Is Possible

Jette Viethen
Centre for Language Technology
Macquarie University
Sydney NSW 2109
jviethen@ics.mq.edu.au

Shared evaluation metrics and tasks are now well established in many fields of Natural Language Processing. However, the Natural Language Generation (NLG) community is still lacking common methods for assessing and comparing the quality of systems. A number of issues that complicate automatic evaluation of NLG systems have been discussed in the literature. [1]

The most fundamental observation in this respect is, in my view, that speaking about "evaluating NLG" as a whole makes little sense. NLG is not one task such as Syntax Parsing or Information Retrieval, but comprises many different subtasks. Just as the subtasks of NLU are evaluated separately using different metrics, corpora and competitions, the subtasks of NLG can only be evaluated individually. With its relatively clear defined task and input characteristics, referring expression generation (REG) is a subtask of NLG for which a shared evaluation scheme is conceivable. In this position paper, I therefore aim to take a solution-oriented look at the challenges of evaluating REG. Although it is unclear just how far any solutions for REG evaluation can be transferred directly to other NLG subtasks, progress in one task might help find solutions for others.

Gold standards: Natural language provides almost countless possibilities to say the same thing in a different way and even under the same external circumstances people use different descriptions for the same object. This variability of human language poses a huge difficulty in terms of what could be used as a gold standard corpus for the evaluation of any NLG task, including REG. It would be unfair to penalise a REG system for not delivering the exact referring expression contained in a corpus, when a large number of alternatives might be equally good or acceptable.

My position: A corpus for REG evaluation would have to contain a large number of descriptions for each referent, as opposed to just one solution per instance. It is unlikely that such a corpus can be drawn from naturally occurring text; the corpus would need to be constructed 'artificially'. This might be done by asking many online participants to provide descriptions for objects from scenes displayed on the screen.

Nevertheless, we will need to keep in mind that an evaluation corpus in NLG will never be really golden: a bad evaluation score might only be due to the 'bad luck' that the perfectly viable solutions a system delivers do not occur in the corpus.

What output do we expect? Three questions need to be answered with respect to the expectations we have of the output of a REG system. Firstly, we lack a definite *Goodness Measure* with which to assess the quality of a referring expression. Secondly, the *Linguistic Level* of the output of existing systems varies and it is not clear at which level we should evaluate. Most researchers are mainly interested in content determination, while some are concerned with the property ordering or even full lexical and syntactic surface realisation. A third question concerns *Solution Counts*: are we contented with one *good* referring expression for each referent, or do we expect a system to be able to produce all the possible descriptions for a referent used by humans.

My position: Psycholinguistic theories such as Grice's maxims of conversational implicature might provide an accurate model of speakers' behaviour when they refer. However, they do not

[1] A bibliography on recent literature relevant to the evaluation of referring expression generation and NLG can be found at http://www.ics.mq.edu.au/~jviethen/evaluation.

provide a straightforward way to reverse-engineer from these behavioural rules to practical guidelines for judging the actual referring expressions produced. A simple and feasible way to find a *Goodness Measure* for the output of REG systems would be to ask human participants not only to provide a description for the gold standard corpus, but also to rank different referring expressions for the same object.

It is clear that output at different *Linguistic Levels* cannot be evaluated using the same corpus and metrics. Before we enter a long and possibly fruitless discussion, we could get started by limiting ourselves to evaluation of REG systems only concerned with content determination. However, we should ensure the possibility to extend the corpus and metric to take word order and surface realisation into account with little extra effort.

If a *Solution Count* of one per referent is expected, the evaluation score can depend directly on the goodness rank of that solution in the corpus (if present at all). If more than one description is allowed, the number of descriptions provided and penalties for over-generation need to be incorporated in the evaluation metric to avoid 'blind' attempts at listing hundreds of descriptions.

Parameters: Most REG systems take a number of parameters such as preference orderings or cost functions over properties and objects, which can have a huge impact on the output. In view of the variability of human-produced referring expressions, it could be argued that algorithms should be allowed to use multiple parameter settings for an evaluation to produce different referring expressions. However, in some cases the parameters are so fine-grained that virtually any desired output can be engineered by carefully choosing the right settings.

My position: This means either that the parameter setting should be considered part of the algorithm proper allowing only one setting to be used throughout the whole evaluation, or that the evaluation metric must penalise systems for switching parameter settings during the evaluation.

A wide field with few players: Just as NLG is a huge field with many subfields, REG can be subdivided into different subtasks such as descriptions involving relations, incorporating object and property salience, or describing sets, and higher-level surface realisation tasks. This is compounded by the high domain-specificity of NLG systems in general. At the same time, the number of researchers in REG, as in most NLG subfields, is comparatively low.

My position: A competitive evaluation scheme for REG bears the potential to stifle research in this field by drawing the attention of the few people working in it to a race for slight percentage increases in a small subtask and domain, instead of advertising the advantages of working on the many untouched research questions.

To cater for evaluation of different subtasks of REG, the gold standard corpus needs to be subdividable and contain referring expressions of different kinds and different domains. To get started, it could be restricted to the most commonly considered types of referring expressions and subsequently extended.

Input Representation: Arguably, the problem of agreeing on the input for NLG is the biggest obstacle in the way towards automatic evaluation. Not only are input representations highly dependent on the application domain of a system, but in existing REG systems the design of the knowledge base from which the algorithm can draw the content for a description is usually tightly intertwined with the design of the algorithm itself. The amount and detail of information contained in the system input differs from case to case, as well as the form it takes: this can range from raw numerical data, over premeditated ontologies of domains, to natural newspaper text.

My position: In order to automatically evaluate REG systems, we have no other choice but to agree on the type of knowledge representation required for the domains covered in the evaluation corpus. As a minimum, the properties and relations of the objects in the different scenes that a system can draw from will need to be predetermined in a simple standard knowledge representation.

Conclusion: There are a number of challenges that have to be overcome in developing useful evaluation metrics for any NLG subtask. However, I am convinced that, for REG, automatic evaluation is possible and would be highly beneficial to the development of systems, if it is based on a large, divisible corpus of ranked descriptions and on basic agreements regarding input representation, parameters, and output expectations.

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.