SHARED TASKS AND COMPARATIVE EVALUATION IN NATURAL LANGUAGE GENERATION

Workshop Report

Edited by

ROBERT DALE
Macquarie University

MICHAEL WHITE
The Ohio State University

Contents

Preface

This document constitutes the final report on an NSF-funded workshop on *Shared Tasks and Comparative Evaluation in Natural Language Generation* held in Arlington, Virginia on April 20–21, 2007.

The collected proceedings[1] of the workshop, consisting of 15 position papers accepted after reviewing by an international program committee, are available as a separate document; this might be a considered a snapshot of opinions regarding comparative evaluation in natural language generation prior to the workshop itself. The present volume[2] contains material that reflects the discussions at the workshop that arose in response to the presentation of those position papers, as well as subsequent discussions between the members of the working groups that were formed at the event.

The report consists of an introductory chapter by the organisers, and four subsequent chapters authored by the members of the working groups. The introductory chapter provides relevant background information about the workshop, describes what happened at the event, and provides an overview of the chapters that follow.

ROBERT DALE AND MICHAEL WHITE

NOVEMBER, 2007

[1] *Position Papers of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, available from `http://www.ling.ohio-state.edu/nlgeval07/papers/NLGEval07-Position-Papers.pdf`.

[2] *Report from the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, available from `http://www.ling.ohio-state.edu/nlgeval07/NLGEval07-Report.pdf`.

Acknowledgements

The editors would like to thank:

- the National Science Foundation, and in particular Tanya Korelsky, for making the workshop possible;
- Kathy McKeown, who provided an insightful and entertaining invited talk at very short notice;
- our reviewers — Anja Belz, University of Brighton, UK; Helmut Horacek, University of Saarbrücken, Germany; and Donia Scott, The Open University, UK — who provided useful feedback on all the initial submissions to the workshop; and
- Claudia Morettini, for her extensive assistance in administering the workshop.

Finally, we would like to thank all the workshop participants, who together made this event more lively and fruitful than one could have hoped for.

Participants

The workshop was attended by the following people:

- *Anja Belz*, Natural Language Technology Group, School of Computing, Mathematical and Information Sciences, University of Brighton
- *Giuseppe Carenini*, Department of Computer Science, The University of British Columbia
- *Robert Dale*, Centre for Language Technology, Department of Computing, Macquarie University
- *Barbara Di Eugenio*, Department of Computer Science, University of Illinois at Chicago
- *Reva Freedman*, Department of Computer Science, Northern Illinois University
- *Albert Gatt*, Department of Computer Science, University of Aberdeen
- *Art Graesser*, Department of Psychology, The University of Memphis
- *Nancy Green*, Department of Computer Science, University of North Carolina at Greensboro
- *Alexander Koller*, Department of Computer Science, Columbia University
- *Tanya Korelsky*, National Science Foundation
- *James Lester*, Department of Computer Science, North Carolina State University
- *Kathy McCoy*, Computer and Information Sciences, University of Delaware

- *David McDonald*, BBN Technologies
- *Kathy McKeown*, Department of Computer Science, Columbia University
- *Johanna Moore*, Human Communication Research Centre, University of Edinburgh
- *Uzzi Ornan*, Department of Computer Science, Technion - Israel Institute of Technology
- *Cécile Paris*, CSIRO-ICT Centre
- *Ehud Reiter*, Department of Computer Science, University of Aberdeen
- *Vasile Rus*, Department of Computer Science, Institute for Intelligent Systems, The University of Memphis
- *Donia Scott*, Centre for Research in Computing, The Open University
- *Amanda Stent*, Department of Computer Science, Stony Brook University
- *Laura Stoia*, Department of Computer Science and Engineering, The Ohio State University
- *Jette Viethen*, Centre for Language Technology, Department of Computing, Macquarie University
- *Marilyn Walker*, Department of Computer Science, University of Sheffield
- *Michael White*, Department of Linguistics, The Ohio State University

Chapter 1

Introduction

Robert Dale[a] and Michael White[b]

[a]*Centre for Language Technology, Macquarie University, Sydney, Australia*
[b]*Department of Linguistics, The Ohio State University, Columbus, OH, USA*

Robert.Dale@mq.edu.au, mwhite@ling.osu.edu

1.1 Background and Motivation

In November 2006, in response to encouragement from Tanya Korelsky at the National Science Foundation (NSF), the editors of this report submitted a proposal to the NSF requesting funding for a workshop aimed at exploring the role of evaluation in Natural Language Generation. That proposal motivated the request for funding in the following terms:

> In recent years, the inclusion of an evaluation component has become almost obligatory in any publication in the field of natural language processing. For complete systems, user-based and task-oriented evaluations are used in both the natural language understanding (NLU) and natural language generation (NLG) communities. A third, more competitive, form of evaluation has become increasingly popular in NLU in the form of shared-task evaluation campaigns (STECs). In a STEC, different approaches to a well-defined problem are compared based on their performance on the same task. A large number of different research

> communities within NLP, such as Question Answering, Machine Translation, Document Summarisation, Word Sense Disambiguation, and Information Retrieval, have adopted a shared evaluation metric and in many cases a shared-task evaluation competition.
>
> The NLG community has so far withstood this trend towards a joint evaluation metric and a competitive evaluation task, but the idea has surfaced in a number of discussions, and most intensely at the 2005 European Natural Language Generation Workshop in Aberdeen, Scotland, and the 2006 International Natural Language Generation Conference in Sydney, Australia. There are a significant number of researchers in the community who believe that some form of shared task, and corresponding evaluation framework, would be of benefit in enhancing the wider NLP community's view of work in NLG, and in providing a focus for research in the field. However, there is no clear consensus on what such a shared task should be, or whether there should be several such tasks, or what the evaluation metrics should be.

Our proposal was driven by the perception that the level of community interest in this topic had reached such intensity that simply having yet another special session or panel discussion at a more general workshop, as had happened on several occasions in the past, would be unlikely to provide the amount of time or sustained interaction required to really thrash out the key questions and issues here. The time was ripe, we felt, for a workshop focussed specifically on this topic.

The NSF agreed with our position; the request for funding was granted, and a workshop was organised to be held at the Hilton Hotel in Arlington, Virginia in the USA on April 20-21, 2007.

To ensure the widest reach possible, we advertised the workshop via a call for papers which was distributed via the SIGGEN mailing list as well as a number of other more general mailing lists and news digests read by members of the natural language processing community. Each submission to the workshop was reviewed by two members of our international pogram committee, and the majority were accepted, consistent with our desire to have as many voices as possible heard; given the potential influence of the workshop on the community's future activity in this area, we wanted to be maximally inclusive.

In terms of numbers, the accepted submissions were broadly split between contributions which were cautious about community-wide evaluation programs, and those which had specific proposals to make in regard to evaluation exercises that might be carried out. This division was consistent with what we had seen in earlier discussions on the topic. Our aim for the workshop, then, was to see if we could further the debate between these two camps; it was unlikely that a consensus position could be reached, but we hoped that two days of intensive interaction would at least make sure that each side understood the pros and cons of the other's position, and that some of the key issues could be properly elucidated.

1.2 The Workshop

We organised the first day in such a way that the more cautious contributions were presented first, in the morning; then, later in the day, we scheduled the presentations that made specific proposals for evaluation.

At the beginning of the day, we took an informal poll to determine the audience's position with regard to shared task evaluations. We asked audience members to anonymously rate, on a scale of 1 to 10 (where 10 constituted strong agreement and 1 constituted strong disagreement), their attitude towards the proposition that holding a shared task evaluation in natural language generation would be a good idea. We repeated this at the middle of the day (after the cautious presentations), and again at the end of the day (after the specific proposals had been presented), to determine whether the presentations had swayed anyone's opinions. The morning vote revealed, again much as expected, an approximately even split between those who were strongly in favour of a shared task evaluation, and those who were not. The midday vote revealed a slight lessening of agreement with the proposition, but by the end of the day the proportions returned much to as they were in the morning, but with some people admitting that they had changed their views (and so, we must suppose, there were just as many switching from the cautious to the optimistic as in the other direction).

We had deliberately left the schedule for the second day underspecified: although we had some ideas as to how we might organise the second day's activities, we wanted to see what the outcomes of the first day's presentations and discussions were before committing to a particular structure.

On the basis of the first day's activities, we identified what we saw as the most prominent themes and ideas that had arisen:

- One group of attendees had raised a range of general methodological concerns in regard to evaluation, as represented by the contributions from Scott and Moore, Di Eugenio, McCoy, and Green.
- There had also been a number of specific proposals for frameworks for evaluation, in the presentations by Paris, Stent, Belz, Reiter, and Mellish and Scott.
- There had been specific proposals for resources that might be of use in evaluation: Stent had suggested a Wiki, and Walker had suggested shared data resources.

In addition, a number of specific proposals for shared tasks had been put forward:

- the use of DUC or other existing STECs as a host (McKeown);
- Referring Expression Generation (Gatt, Viethen);
- Virtual Environments (Koller et al);
- Question Qeneration (Rus and Graesser); and
- Textual Variation (McDonald).

Given that the positions taken with regard to evaluation were still spread between those who favoured a shared task (one or many), and those who remained cautious, we prefaced our proposal to the attendees for the second day's activities with the following hypotheses:

- Those who favour a shared task will push ahead anyway.
- Those who are cautious are the best people to identify the relevant desiderata that should be considered by those who want to push ahead.

We then proposed to split the attendees into a number of breakout groups to develop further the themes that had arisen. After some discussion, groups were formed to work on the following topics:

1. Desiderata: What questions should anyone considering a shared task evaluation keep in mind?
2. Shared Task #1: Referring Expression Generation

3. Shared Task #2: Text to Text

4. Shared Task #3: Virtual Environments

We wanted the working groups to progress the debate as much as possible, so we provided quite specific requirements for what the groups should do, in the hope that these would deliver some raw material that could subsequently be developed into a well-structured report.

The first group was asked to cover three specific questions:

- What's unique about NLG Evaluation?
- What approaches and frameworks might be relevant?
- What would a 'due diligence checklist' — an enumeration of the things to be considered if one is going to pursue a shared task in NLG — look like?

The groups working on shared task specifications were asked to develop, for their particular task, the following elements:

- a definition of the shared task and its type;
- the aims of the shared task: what it would achieve;
- the subcommunity it seeks to engage, and how it would do this;
- whether the task would involve evaluation and how;
- the resources required, and how they would be obtained; and
- a plan of execution.

The morning was then spent in these working groups, with a break for coffee halfway through where we reconvened to establish whether any clarification or fine tuning of the process was necessary. The workshop organisers moved between the groups to ensure that any useful ideas raised in one group could be made use of in the others.

After lunch, the groups came together and reported back on the results of their activity. The report-backs and associated discussion were lively and participative, and occupied nearly two hours.

Finally, around mid-afternoon, following a much needed ice cream break, we had a final wrap-up session where we mapped out a plan for what should happen after the workshop. We developed a schedule whereby the working groups would, on returning home after the workshop, produce written-up versions of the outcomes of their discussions, which would then be distributed to all attendees for comment before being revised for publication to the NLG community.

1.3 Post-Workshop Activity

Our initial schedule for the production of the final report from the workshop was, in hindsight, too optimistic; as is always the case, once people have returned to their respective institutions and normal daily live intrudes, the best of intentions are easily defeated by other pressures. We had intended in our original schedule that a draft of this report would be distributed amongst the wider NLG community before being finalised, but we have decided to skip this step in the process in the interests of making the report widely available as soon as possible. The results of this process are brought together in the document you are reading now.

1.4 An Overview of the Rest of this Document

The remainder of this document provides the results of the post-workshop working group activity. There are four chapters, corresponding to the four working groups:

- Chapter 2: This chapter lays out desiderata for consideration by anyone considering carrying out an evaluation exercise in natural language generation. It can be thought of as a set of guidelines and questions to ask to ensure that at least some of the pitfalls are avoided.

- Chapter 3 focusses on Referring Expression Generation as a shared task. This topic seems, at least at first sight, to be the most amenable to a shared task evaluation, since it is an area that generates a considerable amount of interest, and amongst the subcommunity working on the problem there appears to be broad agreement about the nature of the task.

- Chapter 4, on Text-to-Text Generation, leverages current interest in a number of other subareas of natural language processing where, rather than having some symbolic knowledge representation as either source or target of processing, the raw material to be worked with is text. The chapter sketches a number of directions in which the availability of resources and interest in other communities can be taken advantage of.

- Chapter 5 focusses on Virtual Environments as a platform for shared task evaluations. Rather than provide a specific proposal for a shared task, this group developed in some detail the idea of a broader platform that could play host to a large number of shared tasks.

Subsequent to the workshop, those involved in the group working on referring expression generation as a shared task have gone on to pilot the ideas they developed in Arlington through the *Attribute Selection for Generating Referring Expressions (ASGRE) Challenge*, which has been reported on at the UCNLG+MT workshop co-located with MT-SUMMIT in Copenhagen in September 2007. In specifying the shared task challenge, the organisers have clearly endeavoured to take into account both sides of the debate in Arlington.

While the referring expressions shared task challenge will be the first evaluation exercise out of the gate, we expect the efforts of the Virtual Environments and Text-to-Text working groups will also lead to additional shared task challenges in NLG in the not too distant future. We're particularly hopeful that we'll see results from the Virtual Environments group before long, since this clearly created a buzz of excitement at the workshop. With the deliberations of the Desiderata working group providing a valuable set of considerations for any task definition, we believe that all involved in the workshop have contributed to a significant step forward in our combined understanding of the role that evaluation can and should play in natural language generation. We look forward to seeing how these ideas and plans develop.

Chapter 2

Desiderata for Evaluation of Natural Language Generation

Cécile Paris,[a] Donia Scott,[b] Nancy Green,[c] Kathy McCoy,[d] and David McDonald[e]

[a]*CSIRO - ICT Centre, North Ryde NSW, Australia*
[b]*Centre for Research in Computing, The Open University, UK*
[c]*Dept. of Computer Science, University of North Carolina at Greensboro, USA*
[d]*Computer and Information Sciences, University of Delaware, USA*
[e]*BBN Technologies, Cambridge, MA, USA*

cecile.paris@csiro.au, D.Scott@open.ac.uk,
nlgreen@uncg.edu, mccoy@cis.udel.edu,
dmcdonald@bbn.com

2.1 Introduction

Evaluation is a crucial aspect of any scientific endeavour. It therefore goes without saying that work in Natural Language Generation (NLG) must be evaluated. Establishing good practice for evaluation in NLG will make it easier to judge and show advances in our field and the impact of our work. It will also make it easier for our work to be appreciated and published in reputable fora. In addition, it can set standards of evaluation practice that others in the field can learn and use.

In this chapter, we first describe the requirements for any evaluation. We then review the unique characteristics of NLG, in particular with respect to evaluation, and end with a statement of the agreed desiderata for evaluation in NLG. As these requirements apply to any evaluation of NLG work, they apply in particular to Shared Task Evaluations in NLG. In fact, we believe it is particularly important to take these requirements and desiderata into consideration when planning a shared task evaluation; the resources required to conduct a shared task evaluation are such that it is important to ensure they are not wasted.

2.2 Evaluation: Goals and Characteristics

The primary goal of evaluation is to shed light on the contribution of research (on theories, algorithms, techniques, tools) or development (of systems or components thereof) to the current state of the art in the field of study. Evaluations thus critically inform the progress of the field, allowing scientists to compare the contributions of their theories and algorithms for a given situation (context, need or purpose). Evaluation should advance the field in some meaningful way(s).

As a scientific enterprise, evaluations must therefore conform to the accepted norms of scientific methodology: they must be hypothesis-driven, employ an appropriate methodology (which would normally follow from the hypothesis), be conducted with rigour, be scrutable and replicable.

The goal of any research, of course, is to answer *specific questions* pertaining to what we will call 'the big picture'. A common mistake is to confuse the two, articulating the aim of an evaluation study loosely (e.g., 'to learn more about NLG/referring expression generation/document structuring/. . . ') rather than specifically (e.g., 'do referring expressions of this type, produced in this context, lead to improved performance on this task, carried out in this setting, for this purpose, over some other type of expression?'). Other common mistakes include failing to relate the immediate goals of the research to the wider long term goals of the field, or indeed, to relate the function of the particular NLG component being tested to the larger system of which it forms part. Similarly, many studies fail to specify the scope of the problem being studied. Examples would be a study of referring expression generation that didn't make clear that what was being addressed related only to first-mentions in a setting where the referent was not already known in the discourse context; or that the scope of the study was limited to a context that was 'given a semantic representation of such-and-such granularity, produce a logical form of such-and-such type from which a referring expression suitable for such-and-such setting can be generated'.

In evaluations involving human participants, it is important to ensure that the subjects used are appropriate; for example, linguistics undergraduates would not be suitable for evaluating the output of a system generating medical information for a physician in a clinical setting, while final year medical students could be suitable. Subjects/judges must of course also be independent and unbiased.

In both human-based and automatic evaluations, the measures employed for analysing the results must be meaningful, well-described and appropriate to the evaluation task at hand. In cases where gold standards are employed to benchmark system performance, it is critical that the gold standard provides a true measure of 'quality'. Similarly, the metric that is chosen to compute results and rank or compare systems or approaches must be clearly described and its meaning well understood, particularly in respect to how it is to be interpreted. For example, answers to questions such as the following must be clearly articulated: How meaningful is an increment in the metric? Are only relative rankings or ranges meaningful? How well does the metric correlate to human judgements? If it does not, how is it useful?

While these points may be obvious in retrospect, their importance cannot be stressed enough, and the consequences of not giving careful consideration to them during the design, execution and description of an evaluation study range from difficulties in getting reviewers to appreciate what has been evaluated to wasted effort spent on invalid studies. The latter problem can be particularly devastating in situations where repeating the study is difficult (e.g., where suitable subjects are difficult to come by).

Moving away from methodological concerns, it is worth noting that benefits often come at a cost, and (especially in commercial settings) it is helpful to make these costs explicit: for example, one might gain a 1% improvement in reader comprehension, but this was achieved at the cost of say, a reduced range of appropriate outputs in a context where variety is important, or a reduced system response in a context where speed is important, or four years developing a representation of common sense reasoning in a particular domain. It is often only through the explicit recognition of the cost/benefit tradeoffs that one can properly compare approaches in terms of which best suits the practical need at hand [PCW06, CPW07].

2.3 Unique Properties of NLG vis à vis Evaluation

The unique properties of NLG vis à vis evaluation have been clearly articulated in several contributions to the debate, both in Sydney and in Arlington – e.g., [DE07, Gre07, McC07, McD07, MS07, PCW06, PCW07, SM06a, SM07a, Wal07a]. We will not repeat them here; instead we refer the reader to those works. However, there are two attributes of NLG that impact enormously on the evaluation task and of which we feel compelled to remind the reader, namely: the input can vary widely depending on a system or application (e.g., from numerical data to logical forms), and the text to be produced depends crucially on its intended audience, purpose and application (see, e.g., [McD93, PCW06, SM06a]. In this chapter, we add new considerations.

Our first point relates to an obvious observation: given that language is intrinsically *context dependent*, what is relevant in one application task may be quite irrelevant in another. As a result, the processes involved in 'generating text' may change — sometimes quite radically — from one application to another. This raises issues especially with respect to the notion of 'core' NLG tasks: Are there such things? Notice that while some would argue that lexical choice is an activity that every NLG system must perform, the required sophistication of the lexical choice component can vary quite widely. With this variation in sophistication come differences in underlying knowledge sources, in interactions with other system modules, and differences of the goals of the lexical choice itself (e.g., is it concerned with tuning to the knowledge of the listener, with making text more coherent, with adhering to formality considerations, with generating text that is easier to read or more retainable?). Lexical choosers with different goals and/or different underlying knowledge sources in NLG systems with different architectures are necessarily going to look quite different from each other and will require very different kinds of evaluation. Furthermore, some systems might actually not do much lexical choice at all, if, for the purposes of their application, they are more concerned with issues at the overall text structure and re-use existing text or employ templates at the lexical level. *Studies aiming at establishing appropriate methods for 'core NLG tasks' might not be as meaningful as assumed at first glance.*

Similarly, the definition of an NLG task and its appropriate evaluation are greatly affected by the purpose for which the text is being generated. This will obviously vary not only from one application task to another, but with different contexts in which those application tasks are set. To take an obvious example: reading for the purpose of comprehension is rather different from reading for identification, requiring rather different types of text; reading for recall might be best supported by yet a different type. To consider 'reading' as a universal task thus risks relying on the superficial. An evaluation that ignores the fact that texts must be appropriate to the function that they are intended to perform will be meaningless. *Evaluation studies based on de-contextualised generation of texts will thus have little value.*

It is generally accepted that an important characteristic of an NLG system is its flexibility and ability in producing a range of results. In a practical setting, this feature is in fact often crucial to the system's success. Without it, there is no need for NLG technology — the text would have been more easily and more economically written by hand. One of the most powerful arguments for NLG technology is precisely that it provides the flexibility needed to produce a variety of texts, usually depending on context. *Evaluation studies that ignore the potential of the system to generate a range of appropriate outputs will be necessarily limited.*

NLG systems often draw on disciplines other than NLP alone. For example, systems sometimes employ theories from psychology or psycholinguistics, especially when reading comprehension is paramount. But other disciplines often also come into play: an NLG system is often part of another larger system, and human computer interaction issues play a huge role in how the system is perceived, or even how text is read and understood. Similarly, the way a text is presented will typically affect its understanding and potentially its recall. It might thus be difficult to decouple 'the text being generated' from its presentation, and this can affect the outcome of an experiment. *Evaluation studies that ignore or underestimate possible confounds that arise from the presentational setting of the generated text are likely to lead to invalid conclusions.*

2.4 Desiderata for an NLG Evaluation

Following from the discussion above, we hold that any NLG evaluation must consider the following issues:

Clarity of purpose

- What hypothesis is being tested?
- What larger goal does the study serve?
- What will it shed light on?
- How will it advance the field?

- On which other disciplines does it draw (e.g., psycholinguistics, HCI, etc.) and how?

Clarity of scope

- What is included?
- What is excluded?

Clarity of context

- Under what circumstances is the task applicable?
- Under what circumstances is it not applicable?
- What are the required inputs and outputs?
- What are the required resources?

Clarity of methodology

- What is the experimental set-up?
- What method of analysis and measure/metric is used, and why? How is it to be interpreted?
- If human subjects/judges are used: how appropriate are they to the task at hand? Are they also independent and unbiased?
- If a gold standard is used, how appropriate is it to the task at hand?

Clarity of outcome

- What do the results tell us vis à vis the initial hypothesis?
- What are the limitations of the study?
- How do the findings extend the state-of-the-art?
- How do the findings relate to known results from related fields?

Clarity of cost

- What are the cost/benefit tradeoffs?

2.5 Other Considerations for Evaluation

During the Arlington workshop, and in some of the position papers, it was pointed out several times that there are areas other than the much-touted information extraction, summarisation and message understanding that we should draw and learn from when establishing evaluation methods for NLG. In particular, human-computer interaction and information systems have a long tradition of evaluations, including more holistic views of evaluations than taken by our NLU siblings. In addition, many disciplines in the humanities (e.g., psychology) have long established methods and methodologies for experimental work. We should ensure we draw from these disciplines where possible.

We also want to emphasise again that (good) evaluation is crucial to progress in NLG, and that a wide range of possible evaluation scenarios can be applied to this end, of which shared-task evaluation is but one among many. However, whichever scenario is used, the validity of the outcome, and its contribution to the field will be determined by the extent to which the study in question adheres to the desiderata outlined in this document. This does not, of course, extend to issues such as robustness, the establishment of shared software platforms or the development of shared resources, each of which require rather different evaluation methods from those described here.

As a final note, we believe that, as a community, we should also expend effort and energy into establishing appropriate frameworks to encourage and facilitate groups to collaborate, enable researchers to place their work in a bigger picture and allow for comparisons, without the constraints imposed by the necessarily narrow focus and de-contextualisation of a shared-task evaluation. Such efforts could truly bring the community together in a meaningful way, and contribute to real progress in the field. There is already work on such frameworks — e.g., [PCW06, CPW07, MSC$^+$06] — on which we could build and capitalise.

Chapter 3

Referring Expression Generation

Anja Belz,[a] Albert Gatt,[b] Ehud Reiter[b] and Jette Viethen[c]

[a]*Natural Language Technology Group, University of Brighton, UK*
[b]*Department of Computer Science, University of Aberdeen, UK*
[c]*Centre for Language Technology, Macquarie University, Australia*

A.S.Belz@itri.brighton.ac.uk, agatt@csd.abdn.ac.uk,
ereiter@csd.abdn.ac.uk, jviethen@ics.mq.edu.au

3.1 Introduction

A working group on shared evaluation efforts in referring expression generation (GRE) seemed to be a natural development given the many mentions of GRE as a candidate task for shared evaluation in NLG during the Arlington Workshop and the number of people working on or interested in GRE at present. We, the GRE Working Group, started our discussion with an optimistic attitude toward the feasibility of shared evaluation campaigns in GRE, which is reflected in the concrete suggestions brought forward at the workshop and the implementation of these suggestions following the workshop. This is not to say that we are not aware of the potential risks and difficulties inherent in shared task evaluation in NLG which were discussed at the workshop ([SM07b, SM06b, Vie07]), nor are we attempting to prove that these risks don't exist or are negligible. We rather see the implementation of a shared task evaluation campaign in GRE as an opportunity to investigate ways to overcome the difficulties we are facing, so the research field can gain from them.

Section 3.2 gives a short overview of the structure of the evaluation task we propose. Following this we address the desiderata from Chapter 2 that apply to our proposal. In Section 3.4 we describe existing corpora that could be used for an automatic evaluation scheme and suggest a two-step timeline for putting GRE STECs into place. The first step, a pilot challenge on attribute selection for GRE to gauge feasibility and community interest, has already been concluded. The second step, a more large-scale STEC in GRE, is in planning.

3.2 The Task

3.2.1 Task Structure

The basic structure of a shared task evaluation for GRE as envisaged by the working group involves preparing a data source and setting a small set of clearly defined tasks that can be evaluated against that data source. Such a data source would typically take the form of a corpus of human-generated referring expressions for a certain domain. A number of existing candidate corpora for GRE tasks are discussed below in Section 3.4.1.

While STECs generally consist of a limited number of pre-defined tasks, our proposal seeks to promote greater openness, reflecting the potentially broad range of interests in the community, even within a single sub-task of NLG such as GRE. We therefore propose to include an “open category”, in which participants are free to use the STEC dataset(s) in any way they choose, without directly competing in the STEC proper, but with the option of disseminating their results and descriptions of their approaches in the STEC proceedings. This also goes some way towards addressing one of the chief concerns voiced at the Workshop by those who had a more cautious outlook on STECs, namely, that there is a risk of narrowing the scientific interests in a field by legitimising only a small subset of possible goals, thereby stultifying innovation.

The open category is also intended to encourage participation by people who might not be able to spend a lot of time and resources on developing new approaches to solve our predefined tasks, but have existing systems that could easily be adjusted to work on the data set. We believe that a workshop with a number of people who have worked with the same data set would result in very fruitful discussions, even if their systems are not tackling exactly the same task.

3.2.2 Evaluation

Choice of evaluation method and metrics is the most critical point to be solved in setting up a shared task evaluation campaign in GRE. Task-based evaluation in experiments involving human participants, as well as evaluation by letting humans judge the quality of system output directly, will most likely retain its importance in evaluating NLG systems, including those for GRE.

However, automatic evaluation methods comparing system output to human-produced referring expressions are also useful for a number of reasons. Firstly, only one large time- and resource-expensive experiment involving human subjects is required to collect the corpus that will serve as the standard of comparison, rather than an individual experiment for evaluating each system. Secondly, this allows developers to use training and development sets of the data together with the automatic evaluation metric to gauge the performance of their system during development. A third reason for using automatic evaluation methods lies in their re-usability at a later stage, either in a follow-on shared task or by individual researchers. Furthermore, they guarantee replicability of the evaluation results on the same data set, which can be useful in a number of ways such as verifying the correctness of a re-implemented system. Of course it is not trivial to develop or choose an adequate evaluation method for a certain task and each different method will accentuate different aspects of the task.

We therefore propose that one of the main aims of initial GRE STECs should be to test different evaluation techniques and metrics, automatic ones as well as human based ones, and to study the correlation between them on the outcome of a shared task evaluation. As a default metric we suggest the use of a standard metric, such as the DICE coefficient of similarity, but to also include a call for proposals for more elaborate evaluation metrics as part of early STECs.

Issues Arising from the Pilot Event

As noted above, a pilot challenge on attribute selection for GRE, called the *Attribute Selection for Generating Referring Expressions (ASGRE) Challenge*, has already been concluded and reported on at a workshop at MT-SUMMIT in Copenhagen in September 2007; this is described further in section 3.4.2.

Two main issues became apparent during the evaluation of the systems submitted to the ASGRE Challenge. These are issues that apply to the evaluation of individual systems and need to be addressed in the planning stage of future evaluation campaigns.

Firstly, evaluation against human-produced corpora, such as the TUNA corpus used in the ASGRE Challenge, implicitly assumes that human-produced referring expressions are a kind of gold standard to aspire to. Systems geared to perform well in a comparison to such a corpus are essentially mimicking human language production. It is not clear whether generated language which maximizes similarity with human language production is also optimal for human understanding of the generated language. We saw some indication in our evaluations that this may not be the case.

In the short term, it would be desirable to augment existing corpora with comprehension data from experiments involving human participants. In particular, information on identification accuracy and reaction times from task-based experiments would enable us to rank the referring expressions in corpora of initial referring expressions. For corpora concentrating on subsequent reference, self-paced reading experiments would serve a similar aim. A more ambitious project would be to add eye-tracking data to the existing corpora.

The second question that arose during the ASGRE Challenge was how to perform evaluations of components of NLG systems. The Challenge focused on content, so for the human evaluation, attribute sets were converted to a natural language representation in such away that each attribute was always realised in the same way and in the same position regardless of context, except that negated attributes contained in a list of premodifiers or postmodifiers were grouped together at the end of the list, in order to avoid ambiguity. For this relatively simple task, this “realisation” step was really tantamount to rewriting the attributes as a sequence of words, but for more complex tasks, it might not be clear to what extent people’s task performance is due to the choice of content, the surface realisation of that content, or the combination of the two.

Presenting human participants with some type of content representation that abstracts away from surface realisation is also problematic, because then there might be unwanted effects from presentation in a non-linguistic, unintuitive format. However, depending on the task, this could be considered as an option. Another option might be to use a range of realisers on the same content delivered from the submitted systems and average over the results. For some tasks it might be more appropriate to consider content determination and realisation as one system and expect fully realised referring expressions as output from the systems.

3.2.3 Scope for Variation

There are three main areas that provide scope for variation in the structure of the task(s) of an evaluation campaign:

- **Subtask:** Currently most work in GRE seems to be concerned with content determination, which therefore lends itself to become the initial task for a GRE STEC. However, other possible subtasks that could become the centre of a STEC include surface realisation and lexical choice.

- **Type of reference:** A STEC could either focus on full definite reference, as represented for example in the TUNA corpus, or on anaphoric referring expressions and pronouns, such as the references in the GREC corpus.[1] Other possibilities for variation in the task arise from choosing to produce references to groups or sets of objects rather than singular objects, from allowing relations between objects to be used, and possibly from including reference to non-physical objects.

- **Goal:** Two main goals can be distinguished that a GRE system can aim for: either the aim is to model human production, or it could be to optimise human comprehension. Of course, the goal of the task will have an important influence on the choice of evaluation technique.

3.2.4 Likely Participants

Likely participants in a shared task evaluation challenge in referring expression generation would of course be the group of people currently working in referring expression generation (e.g. all participants in the ASGRE Challenge had worked in GRE before). A number of people interested in NLG who have done work on referring expression generation in the past might feel their interest in the field rekindled by a STEC. We also imagine that members of the CoNLL/EMNLP community might also be interested in participating in such an endeavour.

[1] Both of these corpora are described further below.

3.3 Addressing the Evaluation Desiderata

3.3.1 Larger Goals

The goals of a STEC in GRE, as we see them, can be described as *community aims* on the one hand, and aims of more *scientific* nature on the other.

Community Aims

The most basic aim is simply to have fun [BK06]. Almost anyone who as participated in a STEC will describe the experience as fun (confirmed in feedback from the ASGRE Challenge participants), despite the long nights before submission deadlines and hours spent on working around unexpected input data. People like comparing their work to others', even if they don't come off too well.

Another community aim, more from the perspective of the organisers than from that of the participants, is to encourage collaboration and to consolidate the community. Having a shared task or at least a shared corpus to work on will enable much more detailed discussions, as a common ground can be presumed.

The third community aim is to broaden the community. A STEC often draws the attention of people to a certain research area that they might not otherwise have spent much time on. We hope that people from other areas of NLP might bring new approaches to the research area of GRE.

The opposite effect might of course also be the case: by promoting a STEC on one particular subproblem of a task, a large part of what is — in this case — a rather small research community starts concentrating on only this one problem. This is in fact one of the concerns most frequently articulated in the context of the current 'evaluation debate' in NLG. However, in the case of the ASGRE Challenge, described in section 3.4.2, we saw several people with only latent interest in GRE sufficiently intrigued by the challenge to actually do something in the area.

Scientific Aims

As with all shared-task evaluations, the core scientifc aim is to gain insights into which approaches work best on the given task, as measured by the given set of evaluation methods. In addition to assessing existing methods, shared-task evaluations tend to produce a hothousing effect [BK06] where research effort is for a time intensely focused on the task and new techniques are developed which can sometimes dramatically and quickly advance the state of the art.

GRE being probably the most well-circumscribed subtask of NLG, a STEC in this field can serve as a test-bed to investigate the viability of STECs in other subtasks of NLG. Initial campaigns in GRE can serve to address and hopefully overcome the difficulties inherent in shared task and automatic evaluation of NLG tasks. If we realise that these problems cannot be solved for GRE, we might have to come to the conclusion that other evaluation techniques have to be used for GRE and possibly also other NLG subtasks.

However, we hope that shared evaluation campaigns in GRE will increase the number of novel approaches to the evaluation of GRE and help overcome these difficulties. We expect this will in particular be the case, as we envisage different and possibly new evaluation methods to be applied in a GRE evaluation campaign and their usefulness to be assessed as part of the campaign. One of the foremost aims of a STEC in GRE should be to evaluate evaluation metrics and schemes.

3.3.2 Impact on Other Disciplines

As GRE is turning out to become the pilot task for NLG evaluation campaigns, there is likely to be an impact on the NLG community at large, who might want to monitor the execution and outcomes of a shared task in GRE to gauge implications for evaluation in other subtasks. The experience gathered in early campaigns in GRE will be valuable for further discussion of the pitfalls and benefits of shared task evaluation in NLG more generally. It will enable the community to avoid initial mistakes and difficulties with STECs in other subfields, should any be implemented.

The main discipline outside of Computational Linguistics a STEC in GRE draws on is psycholinguistics. Apart from providing valuable advice and expertise for setting up evaluation experiments, psycholinguists working on reference might be interested in the outcomes of a STEC in GRE to inform their psycholinguistic models of how people refer. In fact, psycholinguists present at a small workshop on GRE held at the University of Aberdeen showed great interest in in the field.[2]

[2] For more information on the *Workshop on Generating Referring Expressions that Are Optimal for Hearers*, see `http://www.csd.abdn.ac.uk/∼kvdeemte/index-workshop`.

3.3.3 Applicable Context

What kind of referring expressions are adequate and which ones from the vast set of possible descriptions are used by humans in any given task, is of course highly dependent on the context of the reference task.

For the pilot challenge, a rather artificial setting of objects "floating" on the screen was used for the task of attribute selection at a semantic level. However, more complex and natural settings are envisaged for future STECs, also including reference as part of a larger discourse.

3.3.4 Inputs and Outputs

For the basic GRE task of object identification in a one-off description used in the pilot challenge, the input needs to contain nothing more than a representation of the physical properties of the domain and the objects contained in it. The problem of GRE algorithms being designed for a specific underlying knowledge representation and vice versa can only be solved by providing one standard knowledge base for all participating systems to work on. This issue is discussed in [Vie07] and in more detail in [VD06]. Along with the underlying knowledge base, the systems take a pointer to the target referent or referents as input.

For different tasks, the input will of course differ. In more sophisticated settings information about the discourse and user models will have to be part of the input an algorithm takes for the generation of each expression.

The output expected from participating systems can vary from unordered sets of attributes to be used in a description for the target referent(s), through ordered lists indicating attribute ordering, to fully fledged surface realisations of referring expressions. These variations have an impact on the choice and implementation of evaluation methods and metrics. For the pilot challenge, we chose only to require unordered sets in a previously specified XML format to keep evaluation simple and enable more people to participate with their existing or slightly modified systems.

3.3.5 Methodology

As mentioned above, the evaluation techniques and metrics to be used should depend highly on the subtask submitted systems are expected to execute. In any case, we suggest that a number of automatic metrics should be used to compare against reference sets of example referring expressions and other criteria, such as length of the referring expressions, in parallel to a human task-based evaluation, in order to allow us to compare the outcomes for different evaluation methods (see Section 3.2.2). In this context it is important always to keep in mind that different systems might be more appropriate for different purposes, an aspect which a single evaluation technique cannot give us enough information about. Exact details of the evaluation have to be determined for each STEC depending on the task(s) that are tested. This determination of adequate evaluation techniques is in our eyes one of the main aims of a STEC (see Section 3.3.1).

If the task domain is general enough, there are no particular criteria for choosing human judges and participants both for corpus collection and for evaluation other than them being adults, as was the case for the pilot challenge. If the task domain is more specialised, for example a biological or medical domain, they need to be drawn from a more expert pool of people.

The collection or choice of a particular reference set of example referring expressions is one of the main difficulties for any NLG evaluation, as the variability of natural language means that it is impossible to guarantee completeness of a corpus (see [SM07b, VD06] for a discussion of this issue). However, as discussed in [Vie07], we believe that with appropriate care and a sufficiently exact definition of the task, evaluation against a reference set can still be very informative.

3.3.6 Expected Outcomes

Especially in the early stages of shared task evaluation in GRE we hope that we will get an insight into the usefulness of different evaluation techniques and criteria. An analysis in this direction for the pilot challenge can be found in [GvD07].

The ranking of submitted systems according to different evaluation techniques will give an indication of which approaches seem promising for a certain task in a certain domain given the criteria examined by the evaluation techniques (e.g. minimality or length of description, humanlikeness, uniqueness or effectiveness).

Hopefully, a number of new approaches will have been developed to address the task of the STEC and we will know whether they are promising or not compared to existing approaches.

3.3.7 Costs

The most expensive aspect of an evaluation campaign as described above would most likely be the evaluation involving human participants. The participants might have to be paid, and the setting up of such an evaluation, recruiting participants and compiling and analysing the data would cost a fair amount of person hours.

For the same reasons, adding comprehension data to the existing corpora might become another source of expense.

In addition, there will be organisational costs for things such as running a website, providing webspace for development data to be downloaded and systems and results to be uploaded, printing proceedings, and of course the labour of the people involved in the organisation.

For the foreseeable future GRE STECs cannot be expected to grow into very large ventures warranting individual STEC workshops. Rather, the results can be presented in the context of a special session at an established NLG conference as was the case for the ASGRE pilot challenge at UCNLG+MT. This should keep organisational costs low compared to larger STECs in other fields of NLP.

At the current stage we don't anticipate the costs to be so high as to require special funding. Most organisational costs can be met by community effort and possibly a minimal registration fee. In addition, the Aberdeen group's EPSRC Platform Grant and funds associated with the Brighton-based Prodigy project were able to cover costs resulting from human-based evaluation for the pilot challenge and similar funding sources are expected to meet such costs for the next GRE STEC, anticipated to be held in conjunction with INLG'08, and desirable corpus extensions (see Section 3.4.1).

Long-established STECs such as CoNNL and SEMEVAL (formerly SENSEVAL) have demonstrated that highly successful evaluation intitiatives can be run on a shoe-string, thanks to extensive community involvement.

3.4 Making It Happen

3.4.1 Resources

In this Section four existing corpora of referring expressions that could be used for automatic evaluation of different GRE tasks are described. Table 3.1 provides a summary.

Corpus	Count	Domain	Type of reference
TUNA	2280	human photos and furniture	initial, sing & pl
Drawers	140	filing cabinets	initial, sing
GREC	8000	Wikipedia texts	anaphoric, sing
COCONUT	393	interior design	varied

Table 3.1: A summary of existing corpora

	Subjects	Singular Trials	Plural Trials
FURNITURE	60	7	13
PEOPLE	60	6	12
TOTAL		1200	1080

Table 3.2: TUNA Corpus layout

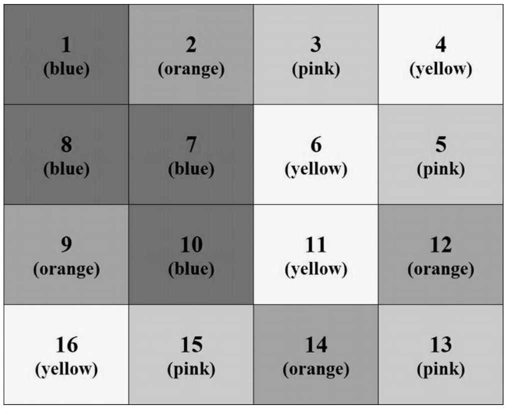
TUNA Corpus

The TUNA corpus consists of 2280 references collected through a controlled experiment that was run over the web[GvdSvD07, vGv07][3]. In the experiment, participants were shown six distractor objects and either one or two clearly marked target referents. Their task was to describe the referents unambiguously. Objects in the trials belonged to two domains: real black and white photographs of people, where distinguishing attributes for referents in these pictures included such features as whether a person had a beard, was bald, wore a suit, and so on; and artificially designed digital images of furniture and household items, such as chairs and desks. Distinguishing attributes in these pictures included the colour of an object, its size (large or small), and the direction faced.

The size and content of the corpus is summarised in Table 3.2.

The annotation of the corpus was designed to meet the requirements of semantic transparency. Thus, each description in the corpus is paired with a knowledge base representation. The latter consists of all the entities shown to a participant in the relevant trial, together with their properties (including their row and column in the grid as seen by the subject). The description itself is annotated with XML tags which indicate which segments of a noun phrase correspond to which domain properties. This makes it an ideal resource for researchers interested in GRE as a content determina-

[3] See http://www.csd.abdn.ac.uk/research/tuna/ for full details.



1 (blue)	2 (orange)	3 (pink)	4 (yellow)
8 (blue)	7 (blue)	6 (yellow)	5 (pink)
9 (orange)	10 (blue)	11 (yellow)	12 (orange)
16 (yellow)	15 (pink)	14 (orange)	13 (pink)

Figure 3.1: The Drawer Domain

tion (attribute selection) problem, as well as a potential resource for use in machine learning experiments and extensions of GRE to realisation and lexicalisation, since the annotation pairs natural language expressions with their semantic representations.

The Drawer Domain

The corpus over the Drawer Domain comprises 140 one-shot descriptions drawn from a physical experimental setting consisting of four filing cabinets, each of which is four drawers high.[4] The cabinets are positioned so that the drawers form a four-by-four grid; each drawer is labelled with a number between 1 and 16 and is coloured either blue, pink, yellow, or orange. There are four drawers of each colour which are distributed randomly over the grid, as shown in Figure 3.1.

Subjects were given a randomly generated number between 1 and 16, and asked to produce a description of the numbered drawer using any properties other than the number. Twenty people contributed to the corpus.

The data set ranges from two descriptions of Drawer 1 to 12 descriptions of Drawer 16. One of the most obvious things about the data set is that even the same person may refer to the same entity in different ways on different occasions, with the differences being semantic as well as syntactic.

[4] More information is available from http://www.ics.mq.edu.au/∼jviethen/drawers/.

Six of the descriptions are ambiguous in that it is not clear which exact drawer they refer to. None of the target referents were sets, however 16 descriptions used reference to a set of drawers to identify the referent.

Each description is annotated with a normalised form to remove superficial variations such as the distinction between relative clauses and reduced relatives, and between different lexical items that were synonymous in context, such as *column* and *cabinet*.

Four absolute properties used for describing the drawers can be identified in the corpus. These are the colour of the drawer; its row and column; and in those cases where the drawer is situated in one of the corners of the grid, its cornerhood. A number of the natural descriptions also made use of the following relational properties that hold between drawers: above, below, next to, right of, left of and between.

Here are some examples of the referring expressions produced:

- the top drawer second from the right [d_3]
- the orange drawer on the left [d_9]
- the orange drawer between two pink ones [d_{12}]
- the bottom left drawer [d_{16}]

GREC (GRE in Context)

The GREC corpus [BV07] is a corpus of 1,078 introductory sections from entries in the online encyclopaedia Wikipedia in which references to the subject of the entry have been annotated. An introductory section was defined as the part of the entry preceding the table of contents. Wikipedia mark-up, images, HTML tags etc. were removed from the entries to yield text-only versions. These were then annotated for references to the subject of the entry by five annotators, and the annotations double-checked. The inter-annotator agreement was 86% before double-checking. The texts in the corpus fall into four subdomains: rivers (83 texts), cities (248 texts), countries (255 texts) and people (492 texts). The corpus currently contains 8,000 nominal expressions which have been annotated for syntactic features. The corpus itself as well as the annotation scheme are currently being extended.[5]

[5] See http://www.nltg.brighton.ac.uk/projects/prodigy for developments.

COCONUT

The COCONUT Corpus consists of 24 computer-mediated design dialogues in which two people collaborate on a simple interior design task [EJTM00]. Nine of the 24 dialogues were annotated for 482 coded utterances by linguists and computational linguists. The annotation scheme covered problem-solving features and dialog features, including forward-looking functions, backward-looking functions, gist tags, and reference tags. The reference tags capture the relation between furniture items in the current utterance and furniture items discussed previously in the same dialogue (*SameItem, MutuallyExclusive, Subset, Reference*). The 393 nominal expressions annotated in the corpus and sets of features derived from the corpus have been used to automatically construct generation rules for an RE generation module [JW00].

3.4.2 Timeline

We propose a two-step plan for proceeding with shared evaluation in GRE. The first step is the now-complete initial pilot challenge, and following this, we envisage the first main STEC in GRE to take place in 2008 and be presented at INLG 2008.

Pilot Event: the ASGRE Challenge

As the first step, the GRE Working Group from the Arlington workshop organised the *Attribute Selection for Generating Referring Expressions (ASGRE) Challenge* much in the spirit of the task outlined in section 3.2.[6] This challenge served as a dry run to pilot the idea of a STEC in NLG with the primary purpose of assessing interest in the community and gathering experience with the organisation of shared task evaluation in GRE. The results of this challenge were presented at a special session at UCNLG+MT[7] on 11 September 2007 in Copenhagen.

The challenge was based on a subset of the TUNA corpus described in section 3.4.1 and offered participants three tracks:

1. Shared Competitive Task: Submitted systems should implement the task of mapping a given input representation to a (single) attribute

[6] For details, see http://www.csd.abdn.ac.uk/research/evaluation.

[7] Using Corpora in Natural Language Generation: Language Generation and Machine Translation. For more information and proceedings see http://www.itri.brighton.ac.uk/ucnlg/.

set that identifies the intended referent.

2. Open Category: Participants were encouraged to submit either systems that use the data for a task different from that in the Shared Task proper, or papers of an observational nature making comments on the task and evaluation.

3. Evaluation Techniques: Also invited were proposals for methods to be used in the evaluation of attribute selection for GRE.

Six teams participated in the pilot challenge and submitted a total of 22 systems. [BG07] gives an overview over the evaluation criteria applied, including DICE similiarity to the TUNA corpus, minimality and task-based evaluation of effectiveness of the generated referring expressions, as well as the results for the submitted systems. [GvD07] provides a short analysis of the evaluation techniques that were used in the pilot challenge.

An informal suggestion was made to use a comparison metric called MASI, which penalises systems more for omissions of properties than DICE does. This metric seems to be very useful and can be applied in subsequent attribute selection challenges.

Main STEC

A larger scale STEC in GRE is planned for in 2008, co-located with INLG'08. This STEC will have a larger scope in terms of the task proposed and the evaluation carried out. We envisage several shared tasks within the area of GRE, but believe an open category should always be part of such a campaign. We believe that human evaluation should become an integral part of the main event.

In the long term, a GRE evaluation campaign could take place in a virtual environment as described in Chapter 5. This would provide a more natural setting for the generation of context-sensitive referring expressions and most likely further the hothouse effect we are hoping for.

Conclusions

The enthusiastic response from GRE researchers to the ASGRE Challenge (and supportive comments from the wider NLG community) demonstrates that parts of the NLG field are willing and able to participate in comparative evaluation events.

The evaluation results of the ASGRE Challenge [BG07] do not tell us what is in general terms the best way to do attribute selection for GRE. Rather, we have directly comparable results for 22 different systems and five quality criteria. This can help guide development and selection of attribute selection systems for similar domains in the future, in particular where such systems are required to maximise specific aspects of quality.

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 many people like the buzz and energy they create, the way they draw new people in, and the hothousing of solutions they foster [BK06]. We believe that NLG should continue to organise shared-task evaluation initiatives. The risks of getting it wrong seem small: shared-task evaluations can be run on a shoe-string (as SEMEVAL and CoNLL continue to demonstrate), and if an event, task or corpus fails to inspire people, it tends to quietly go away.

Chapter 4

Text-to-Text Generation

Vasile Rus,[a] Arthur C. Graesser,[b] Amanda Stent,[c] Marilyn Walker[d] and Michael White[e]

[a]*Department of Computer Science, The University of Memphis, USA*
[b]*Department of Psychology, The University of Memphis, USA*
[c]*Department of Computer Science, Stony Brook University, USA*
[d]*Department of Computer Science, University of Sheffield, UK*
[e]*Department of Linguistics, The Ohio State University, USA*

`vrus@memphis.edu`, `a-graesser@memphis.edu`,
`amanda.stent@gmail.com`, `M.A.Walker@sheffield.ac.uk`,
`mwhite@ling.osu.edu`

4.1 Introduction

Human knowledge is encoded in texts: in books, news articles, encyclopedias, scientific papers and dictionaries. Over the last decade, there has been a dramatic increase in the availability of such knowledge sources in machine-readable form, and in research that attempts to extract knowledge from these sources and make it available in a different form. Thus Text-to-Text generation, which provides algorithms for transforming texts from one form to another, has arisen as an important component of applications such as automatic summarization, information extraction, machine translation, intelligent tutoring, and question answering.

In Text-to-Text generation (henceforth T2T), a phrase, sentence or larger unit of text is extracted from one context and re-utilized in another. At one extreme, the simplest example is extractive summarization techniques, where whole sentences from within a document, or from multiple documents, are extracted and ordered to produce a summary or abstract of a text. More sophisticated techniques may manipulate the original sentences in various ways, or extract phrases from sentences and make new sentences from them, in order to improve coherence, reduce redundancy, ensure consistent style, or fit particular communicative goals. At the other extreme, researchers have noted that texts represent not only the knowledge encoded in the text (the content), but also knowledge about how to *express* the content (the form). Thus recent work has developed techniques for learning generation dictionaries (syntactic to semantic mappings) [Rad98, BL02, Lin06, HWP07], or information presentation strategies [GS05], that can then be used in reformulating the original texts, or in producing completely new sentences.

Unlike data-to-text generation, there is no standard architecture or approach to T2T generation. However, a key difference between architectures is the level of representation of the text that is derived, and the way this representation is used in T2T generation. Thus it is important to distinguish the utterance $\mathcal{U}$ in the text and its representation(s) $\mathcal{R}(\mathcal{U})$. Previous work has used representations ranging from simple word-level features (often used in extractive techniques), to syntactic or semantic features, such as word co-occurrence, named entities, or verb types [Lap03, BL05], to syntactic structures [SM05, NO00, BM05, FW07], to semantic mappings of various types [Rad98, BL02, Lin06, HWP07]. For example, early work on NL interfaces to databases introduced a form of T2T based on the SQL representation that was derived as a representation of the user's query [McK79]. The resulting SQL was then used to paraphrase back to the user the system's understanding of the user's query, a technique still used in spoken dialogue systems today for response generation [Sen02].

Ideally, $\mathcal{R}(\mathcal{U})$ should be automatically derived from the texts, and even more ideally, it should be derived using unsupervised methods so that the techniques are domain-independent. It might be possible to make more use of related research on automatically obtaining semantic representations corresponding to particular linguistic phrases [TR02, PR04, GJ02, ECD$^+$05, Sod07]; but, to our knowledge, none of this work to date has tried to verify whether the learned phrases could be used in generation, and it has been primarily focused on a small set of semantic relations, such as *is-a* or *part-of*.

In this proposal, we argue that T2T shared tasks can stimulate progress in NLG because: the input data is already there, the applications are already there, and there are relevant generation-related research questions.

In Sections 4.2 and 4.3, we explain how T2T shared tasks enable NLG researchers to both contribute to and benefit from related areas of research including Learning Technologies (e.g. tutoring systems) and IR (e.g. question answering systems). In Section 4.4, we list some general considerations regarding evaluation metrics for T2T generation. In Section 4.5, we survey several possible T2T shared tasks, discussing specific issues regarding resources, task definitions, generation challenges and evaluation for each. Finally, in Section 4.6, we conclude with a discussion of next steps.

4.2 The NLG Angle on T2T Shared Tasks

In most T2T generation applications (e.g. summarization, question answering) existing systems are evaluated using metrics that focus on the content that is included/excluded, rather than on text quality. While research in NLG has also been concerned with content selection, methods for ensuring text quality have also been a primary focus. This means that the NLG community has an opportunity to piggyback on, and contribute to, existing applications (with their own research communities and resources). Possible NLG-related questions for T2T generation systems include:

- **Content** – Does the text contain appropriate content to satisfy the communicative goal, without containing extraneous or misleading content?
- **Organization** – Is the content in the text organized coherently?
- **Persuasiveness** – Is the text persuasive and convincing?
- **Fluency** – Is the language in the text fluent and idiomatic?

Many recent T2T generation systems have the aim of minimizing the size of a text (e.g. [GdS06, KM02]) or of improving text coherence (e.g. [BM05, FW07, Lap03, NO00]). However, T2T systems need not limit themselves to addressing only these goals; indeed, T2T generation tasks can be tailored to address a broad set of scientifically and linguistically interesting issues. The following list is deliberately designed to parallel subtasks in a data-to-text generation system:

- **Text selection** – Selection of appropriate text segments from a larger text or texts to satisfy some communicative goal.

- **Text ordering** – Ordering of selected text segments and insertion of discourse cues to form a coherent discourse.
- **Sentence aggregation** – Fusion or aggregation of text segments.
- **Referring expression regeneration** – Regeneration of referring expressions to improve coherence or to achieve other purposes (e.g. distinguish given from new, topicalization).
- **Text restructuring** – Restructuring of text segments to improve coherence, ensure consistent style, or achieve analogous goals (e.g. maximize impact, topicalization).
- **Dictionary Creation** – Learning of word-to-semantic mappings or syntactic-to-semantic mappings, rather than relying on hand-crafted dictionaries for generation.

We believe that shared tasks for T2T generation should be pursued for several reasons. It is already an area of research that includes NLG researchers as well as researchers from other fields (such as summarization, question answering, and intelligent tutoring systems), so it has a likely buy-in and could potentially draw in more NLG researchers. The data and resources for such a task are available, so start-up efforts are minimized. Moreover, there are many potential applications.

4.3 Resources for T2T Shared Tasks in NLG

The appeal of T2T as a shared task for evaluation in NLG lies in its broad applicability, and in the simplicity and naturalness of the input. Research on NLG is typically highly dependent on application-related data or processing components that provide input to the NLG system. For example, a database of airline flights or weather reports may have to be constructed, or a tutoring system or information retrieval system may have to be built. Researchers working on a shared T2T task would not have to build an application before participating in the shared task.

As mentioned above, research on T2T varies from requiring input texts in raw form (extractive summarization) to the derivation of a mapping from a semantic or conceptual representation to syntactic forms. One potential approach, already used in research on corpus-based surface realization [LG02, Bel07] is to utilize annotated resources, with the appropriate level of representation, under the assumption that, in the future, it will be possible to derive resources with such representations automatically. Below, we summarize some of the resources that could be used in a shared task on T2T generation.

1. Because of projects such as TREC, MUC, DUC, GALE, and HALO, we now have access to annotated corpora and resources for particular applications. Some T2T generation researchers have also made their data available (e.g. [BL03]).

2. Additional resources, such as the Penn Treebank [MSM93], PropBank [PGK05], NomBank [M$^+$04], FrameNet [BFL98], VerbNet [KDP00], and the Penn Discourse Treebank [PJD$^+$05, Gro06], include text annotated for syntactic and some semantic information.

3. There are now many automatic text annotation tools such as statistical parsers ([Col99, Cha00, CC07]) and semantic role labelers (e.g. ASSERT [PWH$^+$04]), as well as whole text annotation toolkits (e.g. OpenNLP, Lingpipe, GATE, nltk).

4.4 Evaluation

The evaluation of shared tasks for T2T generation can be manual, task-oriented, or even automatic. For manual evaluation, submissions from participants are judged by independent judges or by other participants [LDF05, NP04, SRH05]. The latter solution is more cost effective and could make it possible to conduct a shared task challenge on a shoestring budget, but runs the risk of criticism when participants are too tough on their colleagues in order to boost their own ranking. Task-oriented evaluation is driven by the impact of a task on user performance measures that cannot be biased by members of the NLP community ([RRO03, RSR03, FRL06]). Automatic evaluation is also possible for well-defined tasks, e.g. ones for which a tool similar to ROUGE [Lin04] could be employed. However, for more open-ended tasks, existing automatic evaluation methods are problematic because they do not take context into account and do not adequately handle desirable variation ([BR06, CBOPK06, FW07, SMS05]). The problem with these methods typically arises from their assumption that evaluation can be based on comparison with a corpus, and that the corpus represents the single *right answer* for how to generate a particular set of content. This assumption does not hold for many generation problems [WRR02, Nen06, Wal07b].

The use of ranking models for evaluation attempts to address this problem by explicitly representing the possibility for variation and its effect on text quality or utterance quality as a ranking over possible outputs. The rankings are elicited from human judges. Then models for replicating the rankings are trained, which can then be applied to unseen inputs, with ranking error rates lower than 20% on the problems applied to so far [NW06, WSMP07].

Ranking models based on the probabilities of statistical language models have also been used in an over-generate and rank architecture to rank utterances by grammaticality [LG02]. However, when grammaticality is not at issue, then it is difficult to use probabilistic language models trained on general corpora to evaluate variation that arises from differences in text or utterance quality, that depend on the perceptions of an individual user, the purpose of the text, or other types of pragmatic or contextual variation. The only way to do this would be to have many different large corpora reflecting exactly the variation that one wanted to model [IBO06]. This is why some researchers have moved to modeling human ranking judgments rather than deriving a ranking from probabilities in a corpus.

4.5 Examples of Shared Tasks for T2T Generation

T2T generation can serve to bring together and summarize shared content across multiple documents, to reformulate single documents for a different audience or purpose, or to extract content for use in dialogue applications such as interactive tutoring or information seeking dialogue. We consider examples within each of these broad categories below.

4.5.1 T2T Generation from Shared Content Units

Resources A shared content unit (SCU), or *nugget*, is short segment of text, typically a clause, containing a single fact. In recent years, several papers have described and evaluated the use of nuggets for evaluation of summarization and automatic question answering ([Nen06, H$^+$05, NP04, LDF06]). As a result of the use of nugget-based evaluation methods in DUC, TREC, and GALE, there now exist several corpora of text with nuggets annotated by hand. In some of these corpora, nuggets have been annotated by multiple annotators. Corpora annotated for nuggets provide one possible set of inputs to a T2T regeneration system [McK06].[1]

We are particularly interested in the potential of the DUC2006 corpus, with annotations to support the PYRAMID method of summary evaluation [Nen06].[2] In this framework, an SCU is similar to a collection of paraphrases in that it groups together words and phrases from distinct summaries into a single set, based on shared content. The words selected from one summary to go into an SCU are referred to as a contributor of the SCU. However, a contributor is not always strictly a paraphrase.

The annotation consists of labeling the SCU with a "concise English sentence" that expresses the shared content. The SCU has a weight corresponding to the number of model summaries that express the designated content, so it is an indicator of the "importance" of the content, according to the humans who originally produced the model summaries.

The following SCU is an example from one of the sample pyramids (D633.CDEH.pyr) and illustrates a relatively straightforward case in which the contributors are each continuous strings (i.e., no discontinuities) whose meaning corresponds fairly directly to the label. All four model summaries contribute to this SCU, so the weight is 4 (W=4). In this case, there is relatively little variation across contributors with respect to the lexicalization and syntax of the shared content.

- SCU 13 (W=4) LABEL: Plaid Cymru is the Welsh nationalist party
- C1: Plaid Cymru, the Welsh nationalist party
- C2: the Welsh nationalist party, Plaid Cymru
- C3: Plaid Cymru, the Welsh nationalist party
- C4: Wales Nationalist Party (Plaid Cymru)

Another SCU from the same sample pyramid illustrates how the contributors can sometimes be less explicit than, or slightly different from, the label expressing the shared meaning. In addition, many different lexicalizations and syntactic forms are possible. This example also illustrates an SCU where the content in the original texts is realized across multiple clauses.

- SCU 49 (W=4) LABEL: Plaid Cymru wants full independence

[1] See also `http://www.ling.ohio-state.edu/nlgeval07/presentations/McKeown-invited.ppt`.

[2] See `http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html\#examples`.

- C1: Plaid Cymru wants full independence
- C2: Plaid Cymru . . . whose policy is to . . . go for an independent Wales within the European community
- C3: calls by . . . (Plaid Cymru) . . . fully self-governing Wales within the European Community
- C4: its campaign for equal rights to Welsh self-determination, Plaid Cymru

It is possible for an SCU to have a single contributor, in the case when only one of the analyzed summaries expresses the label of the SCU. A comparison of DUC 2003 and DUC 2005 data suggests that there are a relatively large number of SCUs of weight one in the 2006 pyramids (Passonneau et al., 2005).

Task Outline We propose a set of shared tasks in this area: participants would receive a multi-document corpus with nuggets annotated, and would produce a set of output texts that re-presented the information from the input text sets to satisfy a communicative goal (e.g. summarize, explain, justify). If the labels of the SCUs in DUC2006 could serve as a representation of meaning themselves, or if the NLG community could agree on a way to represent the meaning of the labels, or if the labels could be automatically generated by some processing that could be then converted into other representations, such as dependency trees, this corpus could support a number of shared tasks for generation.

Here are some example tasks targeted at the DUC corpus or other corpora annotated for content nuggets.

- *Summary Quality* Participants are given a set of summaries and the SCU annotations, along with (potentially) human evaluations of model summary quality, and can use the content pool represented by the SCU annotations to generate alternative summaries, or alternative forms of the model summaries as an add-on task for the DUC evaluation.
- *Who Done It?* Participants may be given a set of nuggets from newspaper articles about a crime, and asked to construct a timeline for the event, or to produce a description of the main criminal's part in the event, forpresentation to a detective. (Obvious analogies to the intelligence community apply.)

- *Will You Buy It?* Participants may be given a set of nuggets from descriptions of and reviews for a product, and asked to produce a critical or enthusiastic summary of the product focusing on attributes the reader specifies (e.g. cost, size, color). (The idea for this task comes from an old hotel recommendation system [Mor89].)

- *Is It the Right Answer?* Participants may be given a set of nuggets from textbooks describing a chemical reaction or physical system, and asked to provide justifications for four possible answers to an SAT-like question regarding the reaction/system. (The idea for this task comes from the HALO project: `http://www.projecthalo.com/`.)

- *Can you converse about it?* Participants can be given a set of nuggets, for example from the DUC2006 corpus, and be asked to converse with a user through a text-based dialogue about information provided by an article or by multiple articles in the corpus.

Generation Challenges For this type of shared task, a simple system might find all the shared nuggets in the multi-document set, and order them by the most common ordering in the multi-document set. A more sophisticated system might perform improved nugget selection, insert discourse cues, or regenerate referring expressions. If the input nuggets are in multiple languages, the generation system may have to rework the output of an MT system for fluency. If the input nuggets are written from different perspectives (e.g. conflicting reviews of a product, or descriptions of a political crime from different political perspectives as in the data used by [BL03]), the generation system may have to edit the nuggets for bias (to remove, change, or add it).

Evaluation Issues For these types of shared task, the output texts could be evaluated along the multiple dimensions mentioned earlier in Section 4.2: content selection, persuasiveness, coherence, and fluency.

Semi-automatic evaluation could be used if human evaluators would also write texts for the input multi-documents that can be compared to the system-generated texts. However, manual evaluation would be preferable.

4.5.2 Text Simplification

Text simplification [CDS96, CS97, CMC$^+$98, CMP$^+$99, Sid04] aims to make text easier for a human user to comprehend by reducing the syntactic or lexical complexity of a text while also attempting to preserve its meaning and information content. Siddarthan [Sid04] reviews studies which suggest that syntactic simplification can aid comprehension of complex text by aphasics and the deaf, as well as by a much larger target group including second language learners, non-native speakers, adult literacy students and people with low reading ages. As Siddarthan explains, it is essential to take the interactions of syntax and discourse into account during text simplification, as its utility in making a text accessible to a wider audience can be undermined if the rewritten text lacks coherence and cohesion. For this reason, text simplification requires one to address generation issues such as sentence ordering, cue-word selection, referring expression generation, and determiner choice.

Resources Text simplification relies on (at least) shallow syntactic parsing and anaphora resolution. Using as input the gold standard parses in the Penn Treebank, together with the linked semantic dependencies of PropBank and discourse annotations of the Penn Discourse Treebank, would make it easier to focus a shared task on generation issues, rather than analysis ones. Another option would be to provide participants with automatic parses, together with the output of pronoun resolution algorithms (reviewed in [Sid04]), as inputs for text simplification.

Task Outline For the full version of this task, the input is a single text and a target reader (e.g. a second language learner, a child), and the output is a version of the input text adapted to the target reader. By providing participants with a baseline text simplification system, easier shared tasks could also be arranged that focus on the specific subtasks of lexical or clausal paraphrase, sentence ordering, referring expression regeneration, or discourse cue insertion. Since these subtasks interact, more advanced shared task challenges could examine multiple subtasks at once.

Generation Challenges Text simplification systems apply a variety of transformations to reduce the syntactic complexity of sentences, for example rewriting a relative clause as an independent sentence. These transformations can interfere with the coherence and cohesiveness of the original text

[Sid04], and thus one challenge is to find a way to order the larger set of shorter, transformed sentences in such a way as to minimize this interference. To illustrate, Siddarthan (p. 100) provides an example involving two simplifications of a sentence, where one of the possible orderings of the three resulting sentences is misleading:

(a) Mr. Anthony, who runs an employment agency, decries program trading, but he isn't sure it should be strictly regulated.

(b) Mr. Anthony runs an employment agency. Mr. Anthony decries program trading. But he isn't sure it should be strictly regulated.

(c) Mr. Anthony decries program trading. Mr. Anthony runs an employment agency. But he isn't sure it should be strictly regulated.

In the misleading ordering (c), the reader is apt to incorrectly interpret the pronoun *it* as referring to the employment agency, rather than program trading, as in the original.

Another challenge that arises in text simplification is to produce new referring expressions in a domain-independent way, which may include replacing pronouns in the original text with full NPs. For example, in (c) above, *it* could be replaced by *program trading*. The interaction between sentence ordering and pronoun interpretation is further illustrated below, where (b) is an amusing though misleading variant of the original; (c) avoids the problem by replacing the pronoun (adapted from Siddarthan, p. 125):

(a) Mr Blunkett has said he is "deeply concerned" by the security breach which allowed a comedian to gatecrash Prince William's 21st birthday party at Windsor Castle. He is to make a statement to the Commons on Tuesday after considering a six-page report on the incident by police.

(b) Mr Blunkett has said he is "deeply concerned" by a security breach. This breach allowed a comedian to gatecrash Prince William's 21st birthday party at Windsor Castle. He is to make a statement to the Commons on Tuesday after considering a six-page report on the incident by police.

(c) Mr Blunkett has said he is "deeply concerned" by a security breach. This breach allowed a comedian to gatecrash Prince William's 21st birthday party at Windsor Castle. Mr Blunkett is to make a statement to the Commons on Tuesday after considering a six-page report on the incident by police.

Evaluation Issues With the easier task formulations focusing on a single subtask of text simplification, automatic measures could be explored comparing the input texts to one or more reference output texts. For more complex tasks, we expect human judgments of coherence, fluency, and meaning preservation to be essential. For the full task, evaluation metrics from the medical/educational communities (e.g. Flesch, Fry graphs, SMOG) can be used to quantify the extent of simplification, while human task-based evaluations (similar to those described in [WR05, RWC05]) can be used to verify expected improvements in readability.

4.5.3 Question Generation and Answer Regeneration

The task of question generation [RCG07] is to take an input text and generate questions about it; the task of answer (re)generation is to take an input text and question and generate answers to the question.

Resources Existing resources that could be used for shared tasks in Question Generation, from the Summarization, Question Answering and Intelligent Tutoring communities, are described below.

TREC-QA Question–Answer pairs. The TREC Question Answering (TREC-QA) track [Voo01] is a good source of existing data. TREC-QA offers thousands of Question–Answer pairs that were used in Question Answering evaluations since 1999. The data can be used in a question generation and answer regeneration task. In TREC-QA data sets, for each sentence (answer) we have a single associated question. 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 that are tailored to the evaluation of lexical choice characteristics of a generation system. A subtask to generate a set of related questions for an input paragraph can be targeted for evaluating discourse related issues, such as referring expressions, the identification of central/pivotal sentences in a paragraph, and goal-oriented summarization.

ITS data. Auto-Tutor [GVR$^+$01], an Intelligent Tutoring System that participates in dialogues with the learner in natural language, offers data sets of expert-generated questions. During a typical session between AutoTutor and a student, the system guides the student on solving a concrete problem, e.g. a physics problem. For each problem in its database, AutoTutor stores a set of sentences, called expectations, that form the ideal answer to the problem. For each expectation there is an associated set of questions. Data sets from AutoTutor can be used to test question generation output.

Task Outline Question Generation is a valuable task for Learning Technologies [GVR$^+$01, LPG92], Help Systems [LPG92], Frequently Asked Question facilities [LPG92], and Question Answering [GLBJC03, Voo01]. For example, a Question Generation system could be used in intelligent tutoring to provide hints and construct questions for the tutor [GVR$^+$01, CRK$^+$06]. The input text would be a paragraph or document; the question generation component would generate a set of hint questions that the tutor would ask to encourage the student to articulate units (SCUs) that are missing from the ideal text in the dialogue history.

Question Answering systems already construct answers to questions. Current QA systems address short answer questions. They use purely extractive approaches that pay little attention to NLG issues such as coherence or fluency. Advanced QA systems would handle questions with open-ended answers (“why” and “how” questions; definitions). Such answers would be composed in the answer generation task. Just as NLG researchers could construct a shared task around automatic summarization, a shared task could be built around an existing open-source QA system.

Generation Challenges For question generation and answer regeneration, there are common NLG issues: selection of content from the input text to meet the communicative or task goals (e.g. impact student learning, answer an input question); sentence planning (e.g. to add information about an entity that is the subject of a question, to provide justification in an answer); referring expression generation (e.g. in follow-up questions, in extended answers); and surface realization (to construct fluent output text).

Additionally, question generation and answer regeneration can take the form of a multi-year shared task (as in the question-answering community). In year 1, the shared task can focus on simple factoid questions/answers (e.g. Who? What? When?). In years 2 and 3, the shared task can focus on more difficult questions and answers (e.g. involving comparisons, or timelines, or methods). In later years, the shared task can focus on complex or deep questions and answers (e.g. involving causation or justification). To focus research effort on NLG issues, subtasks can be proposed that address particular aspects of generation, such as surface realization.

Evaluation Issues Evaluation of generated questions and answers can be manual, task-oriented, or automatic. For example, in the intelligent tutoring application a task-oriented evaluation could focus on students' learning gains. A surface realization subtask can be evaluated using manual or automatic methods already used in the surface realization community. One could evaluate the proportion of SCUs that have corresponding questions generated to extract the SCU. Coverage should be an important criterion for evaluation, namely that there need to be questions generated that cover all/most of the SCUs (nouns, main verbs, propositions) in the text.

4.6 Next Steps

T2T generation represents a promising direction for shared tasks in NLG. However, to further pursue shared tasks in this area, evidence of sufficient buy-in from one or more subcommunities is still required. In particular, several volunteers will be needed to organize shared task challenges; these organizers will need to decide which particular T2T generation tasks to pursue, perhaps by asking for votes or expressions of interest from the community at large. In the meanwhile, the T2T direction can be advanced by individual research efforts that provide additional shared data resources or tools, and by presentations of demo systems at major conferences which raise interest in T2T generation.

Chapter 5

Instruction Giving in Virtual Worlds

Alexander Koller,[a] Johanna Moore,[b] Barbara di Eugenio,[c] James Lester,[d] Laura Stoia,[e] Donna Byron,[e] Jon Oberlander,[b] and Kristina Striegnitz[f]

[a]*Department of Computer Science, Columbia University, USA*
[b]*Human Communication Research Centre, University of Edinburgh, UK*
[c]*Department of Computer Science, University of Illinois at Chicago, USA*
[d]*Department of Computer Science, North Carolina State University, USA*
[e]*Dept. of Computer Science & Engineering, The Ohio State University, USA*
[f]*ArticuLab, Northwestern University, USA*

koller@cs.columbia.edu, J.Moore@ed.ac.uk,
bdieugen@cs.uic.edu, lester@csc.ncsu.edu,
stoia@cse.ohio-state.edu, dbyron@cse.ohio-state.edu,
jon@inf.ed.ac.uk, kris@northwestern.edu

5.1 Introduction

This paper reports on the results of the Virtual Environments Working Group at the Workshop on Shared Tasks and Comparative Evaluation for NLG. This working group discussed the use of virtual environments as a platform for NLG evaluation, and more specifically the generation of instructions in virtual environments as a shared task. It is based on the task proposal by [BKO$^+$07], which a variety of workshop participants expressed interest in.

The use of virtual environments (VEs) as a platform for NLG evaluation addresses the need for cheap, human-based evaluation methodologies in NLG. Using VEs, it is possible to collect data from a human experimental subject that is physically in a different place than the NLG system. This means we can leverage a huge population of potential subjects, in a way similar to "web experiments" in psycholinguistics and psychology [Rei02] or to systems that collect data by observing people playing games [vAD04]. Many existing tasks, such as the generation of referring expressions, can be implemented in a VE framework; in addition, the framework can situate the human user in a simulated physical world, allowing us to study the effects of such a setting on NLG, with potential implications for human-robot interaction. Finally, the use of virtual worlds adds a "fun" factor to the scenario which we hope will attract attention, especially from students, to NLG.

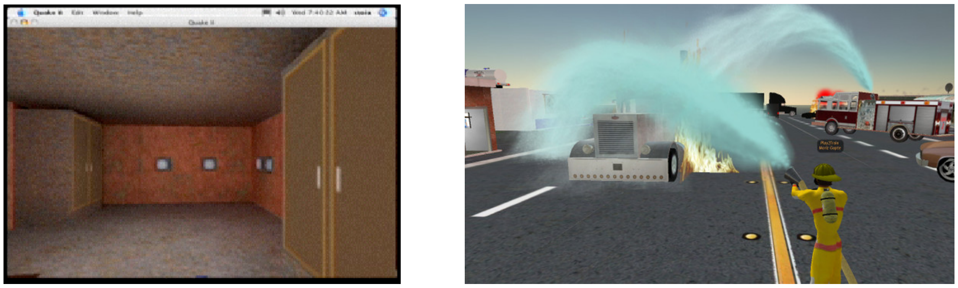
Rather than proposing a single shared task in this paper, we actually propose two different things:

1. a general "virtual environments" setting for NLG systems which can serve as a platform for many different shared tasks; and

2. a concrete shared task, in which the computer's job is to generate instructions for helping the human user solve puzzles in a virtual environment.

Moreover, we see the concrete task as scalable. We propose to start with a "baby steps" version of the task, which is perhaps less complicated than the final task but can be executed with comparatively little effort. We then propose to develop the task further based on the experiences of the first version, scale it up or down, and make it a recurring shared task in a couple of years. In doing so, we want to emphasize the collaborative rather than the competitive aspects of a shared task, and hope that the shared task would give rise to de facto standard modules for NLG.

The paper follows the standard structure for shared task proposals discussed at the workshop: We will first define the task and discuss how it can be evaluated. Then we will explain what aims we hope to achieve with this task, and what subcommunities might find it interesting. Finally, we will describe our plan for carrying out the first round of the challenge. Wherever appropriate, we will distinguish between the general VE setting and the concrete instruction giving task.



(a) (b)

Figure 5.1: Sample virtual environments: (a) the Quake 2 engine used in [SBSFL06], (b) a disaster response scenario in Second Life [AP].

5.2 Definition of the Task

The object of the instruction giving task is to assist a human user in solving a problem in a virtual environment. The user controls a character in a simulated 3D space (see Fig. 5.1); they can move and turn freely, and manipulate and pick up objects in the world. Their goal is to solve a certain problem in the virtual world, e.g. to find an object and move it to a different location. The NLG system has access to complete information about the virtual world and to a plan for achieving the user's goal. The system's job is to generate instructions that assist the user in achieving this goal. At least in the first version of the task, the user will only be able to communicate back to the system by acting in the world and perhaps by pushing buttons on a GUI to signal that they didn't understand an instruction. This will simplify the task, compared to a full-blown dialogue system.

We envision a system architecture in which the NLG server, a central game server, and the graphical 3D client can all run on separate machines and are connected over the Internet (Fig. 5.2). In this architecture, the game server is responsible for keeping track of the state of the world and mediating the communication between the NLG server and the client, and perhaps for matchmaking, i.e. the pairing of users and NLG servers. The virtual world itself can be defined by the task designer, using existing tools for designing maps for 3D computer games. Different 3D engines support different views of the scene; for example, Fig. 5.1a is a first-person view, whereas Fig. 5.1b uses a view over the avatar's shoulder. In the challenge, we will focus on a first-person view.

The NLG system is initialized with the properties of all objects in the

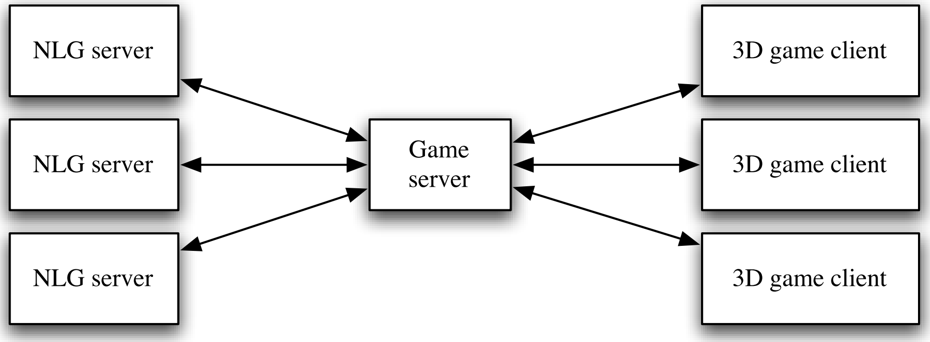


Figure 5.2: The system architecture. Note that no two subsystems need to run on the same machine.

virtual world. It is then notified every time the virtual world changes, e.g. in response to a user action. Furthermore, it receives periodic updates about the user's position and orientation, as well as about the objects in the world that the user can see. It can then decide for itself at which times it should take an action to communicate an instruction to the user, or to guide the user back into its plan, and send the instruction to the user at any time, to be displayed to the user as written text or spoken using a TTS system. The information that the system receives about the world is symbolic: All objects in the virtual world have names and properties (such as the object type, color, etc.) and three-dimensional positions. The task makes no assumptions about the linguistic formalisms or resources that the NLG system uses to generate the NL instructions.

In addition to instruction giving, virtual worlds can also be used for other concrete tasks. For instance, one could imagine an implementation of a referring expressions task in which the potential referents are all realized as objects in the virtual environment. The system could generate an RE, and the user's job would be to click on what they think is the intended referent. On the other hand, the instruction-giving task could also be scaled up in difficulty, extended to a dialogue task, or modified into a pure navigation task (such as the Map Task [ABB$^+$91]). Such tasks would still benefit from the network-based architecture.

5.3 Evaluation

One of the main strengths of the proposed task is that it can be evaluated very well. The central game server can automatically determine the task completion rate of an NLG system and the typical task completion times. In addition, because it is informed about every single mouse click of the user, it can also determine the proportion of referring expressions generated by the NLG system that were correctly resolved by the users. All these data can be collected without requiring any user intervention beyond their playing the game. The system can also collect subjective data via questionnaires presented to the user after each game round. These subjective and objective criteria could then be analyzed using a PARADISE-style framework [WLKA97].

Technically, all NLG systems participating in the shared task could be evaluated simultaneously. Each participating research group would run their system on a server at their own institution, and register it with the central game server provided by the task organizers. The game server would then accept connections from game clients (running on the machines of each experimental subject) and connect each client to a random NLG server; this run of the client would then count towards the evaluation data for this NLG system. After a certain period of time, the central game server would be stopped and the collected data aggregated and compared.

If the user is made to interact with the virtual world in a lab environment rather than over the Internet, it is also possible to collect further data through eyetracking studies. This sacrifices the size of the subject pool in favor of a more controlled experiment that allows us to collect more detailed data. Such a study of users instructed by avatars in a virtual environment is currently being piloted in Edinburgh [DJ07].

5.4 Why This Task is Interesting

The primary aim of the proposed scenario is to provide a new framework for evaluating NLG systems. By making it possible to collect experimental data over the Internet, we tap into a huge pool of potential experimental subjects: For instance, the ESP game [vAD04] has collected over 10 million labels for online images in the past three years, and the MIT Restaurant Game [Ork07], which received far less media attention and requires users to download and install a client to their own computers, still ran about 5,600 games, with an average length of ten minutes, within its first half year. This means that different systems, and different versions of the same system, can be compared in the context of a task-based human evaluation. This has advantages both over (expensive) evaluations using paid subjects, and over gold-standard based comparisons, which are problematic for NLG. These advantages apply to any task that can be evaluated in the virtual environments setting.

In addition, the instruction-giving task in virtual worlds emphasizes the role of generating referring expressions in a situated setting, and thus opens up new research perspectives. This is a very different problem than the classical non-situated Dale and Reiter–style referring expression generation task: For example, experiments have shown that human instruction givers make the instruction follower move to a different location in order to use a simpler referring expression [SBSFL06]. The task also involves such issues as aggregation and the generation of discourse cues and prosody. Overall, the virtual world setting can improve our understanding of situated communication — with potential applications to human–robot interaction, but without the need to deal with the difficulties of real robots, such as image recognition or navigation.

Because the virtual environments scenario is so open-ended, it — and specifically the instruction-giving task — can potentially be of interest to a wide range of NLG researchers. This is most obvious for research in sentence planning (GRE, aggregation, lexical choice) and realization (the real-time nature of the task imposes high demands on the system's efficiency). But as we have argued above, the task can also involve issues of prosody generation (i.e., research on text/concept-to-speech generation), discourse generation, and human–robot interaction. In addition, it touches upon a variety of neighboring research fields: In particular, the task constitutes a new application area for planning and plan recognition.

Furthermore, the virtual worlds setting could be relevant for researchers interested in dialogue systems. The instruction-giving NLG task can be extended to an instruction-giving dialogue task by allowing the user to talk back to the system, e.g. to ask clarification questions, making the virtual worlds scenario a platform for the evaluation of dialogue systems. The virtual worlds platform could also be used directly to connect two human users and observe their dialogue while solving a problem. Judicious variation of parameters (such as the familiarity of users or the visibility of an instruction-giving avatar) would allow the construction of new dialogue corpora along such lines.

It is clear that no single system participating in the proposed shared task will involve ground-breaking progress in all of these areas. However, we believe that each research team could implement a simple baseline system with limited effort, and then improve those modules they find most interesting. We hope that the teams would then make their systems (or the modules into which they put the most research effort) available to the public. These systems could then be used by other teams in the next iteration of the shared task, which would lower the barrier to entry for new NLG researchers and could lead to the development of de facto standards for such modules in the long run.

5.5 Making it Happen

5.5.1 Required Resources

The most expensive resource that is required for the proposed shared task is the computing infrastructure for the network-based evaluation. It will be necessary to develop the central game server, the 3D game client running on the experimental subjects' machines, and an API or protocol for the NLG servers. Such components don't exist today in this exact form, but there is a wealth of open-source software that can be adapted and libraries that can be used to facilitate the development. For example, Byron's research group successfully adapted the Quake 2 game engine for their human–human experiments [Byr05].

In addition, it will be necessary to develop virtual worlds and concrete tasks that the user needs to perform in these worlds. Again, there are open-source tools that support this, but of course substantial effort will be needed to define worlds that (a) people will actually want to play in, and (b) are challenging for the NLG systems we want to evaluate. One source of inspiration for the development of these worlds could be the Edinburgh Map Task [ABB+91]. In addition, experiments with human instruction givers, as started in [SBSFL06], would contribute to an understanding of the NLG-relevant phenomena in this task.

Running the evaluation itself requires a game server that has a fast network connection and is capable of keeping track of multiple instances of the virtual world simultaneously. Finally, it will be necessary to make the experiment visible to potential experimental subjects, e.g. by posting about it in online gaming forums or listing it in a directory of psycholinguistic web experiments.

5.5.2 Plan of Execution

The task of giving instructions in virtual worlds is, at this point, not yet sufficiently well-defined and the research challenges involved in it not yet sufficiently well-understood to be used as a shared task. This is why we propose to proceed in two steps, as follows.

In a first step, we propose to publicize the instruction-giving task as a challenge for teams of students. We will implement the necessary software infrastructure and some sample worlds and tasks, as well as a clear API for NLG systems. We hope to complete this step around Spring 2008. We will then publish a call for participation to student teams anywhere (which will hopefully be supported by the readers of this document), and run a first evaluation using the students' submissions late in 2008. We believe that it is feasible for a (reasonably well supervised) student team to come up with a system that can participate in the challenge within a few months, although such a system will typically not have a very high task completion or user satisfaction rate. As a side effect, we believe that the challenge, with its 3D and game-playing aspects, would attract smart students to spend time on NLG.

We will then organize a workshop to present the students' systems, compare notes, learn from the experiences in this first round, and refine the task definition into a concrete shared task to be organized in 2009. This first "real" instance of the shared task would then also be an opportunity to iron out bugs in the software infrastructure and come up with improved, more interesting, or more challenging virtual worlds and tasks. From this point on, we could then organize the shared task annually or every other year. In doing so, we will emphasize the non-competitive character of the challenge, and review our experiences from each year's challenge to make sure we are still working towards interesting research goals, rather than pursuing a local maximum, and modify or extend the shared task as needed.

5.6 Conclusion

In this document, we have presented our proposal for a shared task of generating instructions in a virtual world. This proposal has two aspects: It is simultaneously a concrete shared task proposal and a proposal for a novel framework for evaluating NLG systems.

After an initial preparation phase in which we will develop the software infrastructure necessary for carrying out this task, we will first carry out a simple version of the proposed task, targeted at student teams. We will then evaluate our experiences from this step and use them to define a more advanced version of the shared task, which we will publicize as an actual research challenge in 2009.

One interesting topic to explore will be the relationship between the shared task we propose and the Referring Expression Generation shared task. Our task properly subsumes the Referring Expression Generation task: As a tiny special case, we can position the user in front of a number of possible referents and then generate a RE without allowing the user to move. Thus our system could be used as an internet-based evaluation platform for the GRE task, but whether this is reasonable or overkill remains to be seen.

Bibliography

[ABB$^+$91] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. The HCRC Map Task corpus. *Language and Speech*, 34:351–366, 1991.

[AP] Idaho Bioterrorism Awareness and Preparedness Program. Play2train website. `http://play2train.hopto.org/`.

[Bel07] A. Belz. Probabilistic Generation of Weather Forecast Texts. *Proceedings of NAACL HLT*, pages 164–171, 2007.

[BFL98] C. Baker, C. Fillmore, and J. Lowe. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, 1998.

[BG07] Anja Belz and Albert Gatt. The attribute selection for GRE challenge: Overview and evaluation results. In *Proceedings of the MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation*, pages 75–83, 2007.

[BK06] Anja Belz and Adam Kilgarriff. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of INLG-2006*, 2006.

[BKO$^+$07] Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. Generating instructions in virtual environments (GIVE): A challenge and an evaluation testbed for NLG. In Robert Dale and Mike White, editors, *Proceedings of the Workshop for Shared Tasks and Comparative Evaluation in NLG*, Arlington, VA, 2007.

[BL02] R. Barzilay and L. Lee. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of EMNLP*, pages 164–171, 2002.

[BL03] R. Barzilay and L. Lee. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of NAACL 2003*, 2003.

[BL05] R. Barzilay and M. Lapata. Modeling local coherence: an entity-based approach. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 141–148, 2005.

[BM05] R. Barzilay and K. McKeown. Sentence fusion for multi-document news summarization. *Computational Linguistics*, 31(3), 2005.

[BR06] A. Belz and E. Reiter. Comparing automatic and human evaluation of nlg systems. In *Proceedings of EACL06*, 2006.

[BV07] Anja Belz and Sebastian Varges. Generation of repeated references to discourse entities. In *Proceedings of the 11th European Workshop on Natural Language Generation*, 2007.

[Byr05] Donna K. Byron. The OSU Quake 2004 corpus of two-party situated problem-solving dialogs. Technical Report OSU-CISRC-805-TR57, The Ohio State University Computer Science and Engineering Department, 2005. `ftp://ftp.cse.ohio-state.edu/pub/tech-report/2005/TR57.pdf`.

[CBOPK06] C. Callison-Burch, M. Osborne, and P. P. Koehn. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL-06*, pages 249–256, 2006.

[CC07] S. Clark and J. Curran. Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, 33(4), 2007.

[CDS96] Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. Motivations and methods for text simplification. In *In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, 1996.

[Cha00] E. Charniak. A maximum-entropy-inspired parser. In *Proceedings of NAACL*, 2000.

[CMC+98] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of AAAI98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1998.

[CMP+99] John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. Simplifying English text for language impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*, 1999.

[Col99] M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

[CPW07] Nathalie Colineau, Cécile Paris, and Ross Wilkinson. Bang for buck in exploratory search. Technical report, CSIRO ICT Centre, 2007.

[CRK+06] Z. Cai, V. Rus, H.J. Kim, S. Susarla, P. Karnam, and A.C. Graesser. Nlgml: A natural language generation markup language. In T.C. Reeves and S.F. Yamashita, editors, *Proceedings of E-Learning Conference*, page 27472752, Honolulu, Hawaii, 2006.

[CS97] Raman Chandrasekar and Bangalore Srinivas. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10:183–190, 1997.

[DE07] Barbara Di Eugenio. Shared task and comparative evaluation for nlg: to go ahead, or not to go head? Position paper at the Arlington workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, April 20-21, 2007, Arlington, Virginia, USA., 2007.

[DJ07] Sara Dalzel-Job. A comparison of eye tracking and self-report measures of engagement with an eca. Master's thesis, University of Edinburgh, 2007.

[ECD+05] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, , and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[EJTM00] Barbara Di Eugenio, Pamela W. Jordan, Richmond H. Thomason, and Johanna D. Moore. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies*, 53(6):1017–1076, 2000.

[FRL06] K. Forbes-Riley and D.J. Litman. Modeling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In *Proceedings of HLT/NAACL 2006*, 2006.

[FW07] M.E. Foster and M. White. Avoiding repetition in generated text. In *Proc. of the 11th European Workshop on Natural Language Generation*, 2007.

[GdS06] M. Gagnon and L. da Sylva. Text compression by syntactic pruning. In *Proceedings of the 19th Canadian Conference on Artificial Intelligence*, 2006.

[GJ02] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[GLBJC03] A.C. Graesser, M.M. Louwerse, J. Burger, and et al. J. Carroll. Question generation and answering systems: R&d for technology-enabled learning systems. research roadmap for federation of american sciences., 2003.

[Gre07] Nancy Green. Position statement for workshop on stec in nlg. Position paper at the Arlington workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, April 20-21, 2007, Arlington, Virginia, USA., 2007.

[Gro06] The PDTB Research Group. The penn discourse treebank 1.0 annotation manual. Technical Report IRCS-06-01, Institute for Research in Cognitive Science, University of Pennsylvania, 2006.

[GS05] H. Guo and A. Stent. Trainable adaptable multimedia presentation generation. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI 2005)*, 2005. Demo paper.

[GvD07] Albert Gatt and Kees van Deemter. Content determination in GRE: Evaluating the evaluator. In *Proceedings of the MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation*, pages 101–103, 2007.

[GvdSvD07] Albert Gatt, Ielka van der Sluis, and Kees van Deemter. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation*, pages 49–56, 2007.

[GVR+01] A.C. Graesser, K. VanLehn, C.P. Rose, P.W. Jordan, and D. Harter. Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22(4):3952, 2001.

[H+05] A. Harnly et al. Automation of summary evaluation by the pyramid method. In *Proceedings of RANLP*, 2005.

[HWP07] Ryuichiro Higashinaka, Marilyn A. Walker, and Rashmi Prasad. An unsupervised method for learning generation dictionaries for spoken dialogue systems by mining user reviews. *ACM Trans. Speech Lang. Process.*, 4(4):8, 2007.

[IBO06] Amy Isard, Carsten Brockmann, and Jon Oberlander. Individuality and alignment in generated dialogues. In *Proceedings of the 4th International Natural Language Generation Conference (INLG-06)*, pages 22–29, Sydney, Australia, 2006.

[JW00] Pamela Jordan and Marilyn Walker. Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000.

[KDP00] K. Kipper, H. Trang Dang, and M. Palmer. Class-based construction of a verb lexicon. In *Proceedings of AAAI*, 2000.

[KM02] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 2002.

[Lap03] M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*, 2003.

[LDF05] J. Lin and D. Demner-Fushman. Automatically evaluating answers to definition questions. In *Proceedings of HLT/EMNLP 2005*, 2005.

[LDF06] J. Lin and D. Demner-Fushman. Will pyramids built of nuggets topple over? In *Proceedings of Proceedings of HLT/NAACL 2006*, 2006.

[LG02] Irene Langkilde-Geary. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *International Natural Language Generation Conference (INLG)*, pages 17–24, 2002.

[Lin04] C. Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

[Lin06] Jing Lin. Using distributional similarity to identify individual verb choice. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 33–40, Sydney, Australia, 2006.

[LPG92] T.W. Lauer, E. Peacock, and A.C. Graesser. *Questions and Information Systems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

[M$^+$04] A. Meyers et al. Annotating noun argument structure for nombank. In *Proceedings of LREC-2004*, 2004.

[McC07] Kathleen McCoy. To share a task or not? some ramblings from a mad (i.e., crazy) nlger. Position paper at the Arlington workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, April 20-21, 2007, Arlington, Virginia, USA., 2007.

[McD93] David McDonald. Issues in the choice of a source for natural language generation. *Computational Linguistics*, 19(1):191–197, March 1993.

[McD07] David McDonald. Flexibility counts more then precision. Position paper at the Arlington workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, April 20-21, 2007, Arlington, Virginia, USA., 2007.

[McK79] K. McKeown. Paraphrasing using given and new information in a question-answer system. In *Proceedings of ACL 1979*, 1979.

[McK06] K.R. McKeown. Lessons Learned from Large Scale Evaluation of Systems that Produce Text: Nightmares and Pleasant Surprises. *Proceedings of the Fourth International Natural Language Generation Conference*, pages 3–5, 2006.

[Mor89] K. Morik. User models and conversational settings: Modeling the user's wants. In A. Kobsa and W. Wahlster, editors, *User Models in Dialog Systems*. Springer, 1989.

[MS07] Chris Mellish and Donia Scott. Nlg evaluation: Lets open up the box. Position paper at the Arlington work-shop on Shared Tasks and Comparative Evaluation in Natural Language Generation, April 20-21, 2007, Ar-lington, Virginia, USA., 2007.

[MSC$^+$06] C. Mellish, D. Scott, L. Cahill, D. Paiva, R. Evans, and M. Reape. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12(1):1–34, March 2006.

[MSM93] M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 1993.

[Nen06] A. Nenkova. *Understanding the process of multi-document summarization: content selection, rewrite and evaluation*. PhD thesis, Columbia University, 2006.

[NO00] H. Nanba and M. Okumura. Producing more readable extracts by revising them. In *Proceedings of COLING 2000*, 2000.

[NP04] A. Nenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*, 2004.

[NW06] Chrystal Nakutsu and Michael White. Learning to say it well: reranking realizations by predicted synthesis quality. In *Proceedings of ACL/COLING*, 2006.

[Ork07] Jeff Orkin. Learning plan networks in conversational video games. Master's thesis, Massachusetts Institute of Technology, 2007.

[PCW06] Cécile Paris, Nathalie Colineau, and Ross Wilkinson. Evaluations of nlg systems: common corpus and tasks or common dimensions and metrics?. In *International Conference on Natural Language Generation (INLG-06)*, pages 127–129, Sydney, Australia, July 2006. Held as a workshop on the COLING/ACL Conference, July 15-16.

[PCW07] Cécile Paris, Nathalie Colineau, and Ross Wilkinson. Nlg systems evaluation: a framework to measure impact on and cost for all stakeholders. Position paper at the Arlington workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, April 20-21, 2007, Arlington, Virginia, USA., 2007.

[PGK05] M. Palmer, D. Gildea, and P. Kingsbury. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1), 2005.

[PJD$^+$05] Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation (UCNLG-05)*, 2005.

[PR04] Patrick Pantel and Deepak Ravichandran. Automatically labeling semantic classes. In *Proc. HLT/NAACL*, pages 321–328, 2004.

[PWH$^+$04] S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky. Shallow semantic parsing using support vector machines. In *Proceedingss of HLT/NAACL*, 2004.

[Rad98] Dragomir R. Radev. Learning correlations between linguistic indicators and semantic constraints: Reuse of context-dependent descriptions of entities. In *COLING-ACL*, pages 1072–1078, 1998.

[RCG07] V. Rus, Z. Cai, and A. Graesser. Evaluation in natural language generation: The question generation task. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, Arlington, VA, April 20-21 2007.

[Rei02] Ulf-Dietrich Reips. Standards for Internet-based experimenting. *Experimental Psychology*, 49(4):243–256, 2002.

[RRO03] E. Reiter, R. Robertson, and L. Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, (144):41–58, 2003.

[RSR03] E. Reiter, S. Sripada, and R. Robertson. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research*, 18(491-516), 2003.

[RWC05] E. Reiter, S. Williams, and L. Crichton. Generating feedback reports for adults taking basic skills tests. In A. Macintosh, R. Ellis, and T. Allen, editors, *Applications and Innovations in Intelligent Systems XIII (Proceedings of ES-05)*, pages pages 50–63, 2005.

[SBSFL06] L. Stoia, D. Byron, D. Shockley, and E. Fosler-Lussier. Sentence planning for realtime navigational instruction. In *Companion Volume to Proceedings of HLT-NAACL 2006*, pages 157–160, New York City, USA, 2006. Association for Computational Linguistics.

[Sen02] Stephanie Seneff. Response planning and generation in the mercury flight reservation system. *Computer Speech and Language: Special Issue on Spoken Language Generation*, 2002.

[Sid04] A. Siddharthan. *Syntactic simplification and text cohesion*. PhD thesis, University of Cambridge, 2004.

[SM05] R. Soricut and D. Marcu. Towards developing generation algorithms for text-to-text applications. In *Proceedings of ACL 2005*, 2005.

[SM06a] Donia Scott and Johanna Moore. An NLG evaluation competition? eight reasons to be cautious. Technical Report 2006/09, Department of Computing, The Open

University, 2006. http://mcs.open.ac.uk/ds5473/publications/TR2006_09.pdf.

[SM06b] Donia Scott and Johanna Moore. An NLG evaluation competition? Eight reasons to be cautious. Technical Report 2006/09, Department of Computing, The Open University, 2006.

[SM07a] Donia Scott and Johanna Moore. An nlg evaluation competition? eight reasons to be cautious. Position paper at the Arlington workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, April 20-21, 2007, Arlington, Virginia, USA., 2007.

[SM07b] Donia Scott and Johanna Moore. An NLG evaluation competition? eight reasons to be cautious. In Robert Dale and Mike White, editors, *Proceedings of the Workshop for Shared Tasks and Comparative Evaluation in NLG*, Arlington, VA, 2007.

[SMS05] A. Stent, M. Marge, and M. Singhai. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of CICLing*, 2005.

[Sod07] Stephen Soderland. Moving from textual relations to ontologized relations. In *AAAI Spring Symposium workshop on Machine Reading*, 2007.

[SRH05] S. Sripada, E. Reiter, and L. Hawizy. Evaluating an nlg system using post-editing. In *Proceedings of IJCAI 2005*, 2005.

[TR02] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 214–221, 2002.

[vAD04] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the ACM CHI Conference*, 2004.

[VD06] Jette Viethen and Robert Dale. Towards the evaluation of referring expression generation. In *Proceedings of the 4th Australiasian Language Technology Workshop*, pages 115–122, Sydney, Australia, 2006.

[vGv07] Ielka van der Sluis, Albert Gatt, and Kees van Deemter. Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 2007.

[Vie07] Jette Viethen. Automatic evaluation of referring expression generation is possible. In Robert Dale and Mike White, editors, *Proceedings of the Workshop for Shared Tasks and Comparative Evaluation in NLG*, Arlington, VA, 2007.

[Voo01] E. Voorhees. The trec Question Answering track. *Natural Language Engineering*, 7(4), 2001.

[Wal07a] Marilyn Walker. Share and share alike: Resources for language generation. Position paper at the Arlington workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, April 20-21, 2007, Arlington, Virginia, USA., 2007.

[Wal07b] Marilyn A. Walker. Share and share alike: Resources for language generation. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, Arlington, VA, April 20-21 2007.

[WLKA97] Marilyn Walker, Diane Litman, Candace Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th ACL*, 1997.

[WR05] S. Williams and E. Reiter. Appropriate microplanning choices for low-skilled readers. In *Proceedings of IJCAI-2005*, pages 1704-1705.

[WRR02] Marilyn Walker, Owen Rambow, and Monica Rogati. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16:409–433, 2002.

[WSMP07] Marilyn A. Walker, Amanda Stent, Francois Mairesse, and Rashmi Prasad. Learning to Adapt to Individuals and Domains in Spoken Language Generation. *Journal of Artificial Intelligence Research*, 30, 2007. to appear.