Reuse and Challenges in Evaluating Language Generation Systems: Position Paper

Kalina Bontcheva
University of Sheffield
Regent Court, 211 P ortobello Street
Sheffield S1 4DP, UK
kalina@dcs.shef.ac.uk

Abstract

Although there is an increasing shift towards evaluating Natural Language Generation (NLG) systems, there are still many NLG-specific open issues that hinder effective comparative and quantitative evaluation in this field. The paper starts off by describing a *task-based*, i.e., black-box evaluation of a hypertext NLG system. Then we examine the problem of *glass-box*, i.e., module specific, evaluation in language generation, with focus on evaluating machine learning methods for text planning.

1 Introduction

Although there is an increasing shift towards evaluating Natural Language Generation (NLG) systems, there are still many NLG-specific open issues that hinder effective comparative and quantitative evaluation in this field. As discussed in (Dale and Mellish, 1998), because of the differences between language understanding and generation, most NLU evaluation techniques[1] cannot be applied to generation. The main problems come from the *lack of well-defined input and output* for NLG systems (see also (Wilks, 1992)). Different systems assume different kinds of input, depending on their domains, tasks and target media, which makes comparative evaluation particularly difficult.[2] It is also very hard to obtain a quantitative, objective, measure of the quality of output texts, especially across different domains and genres. Therefore, NLG systems are normally evaluated with respect to their usefulness for a particular (set of) task(s), which is established by measuring user performance on these tasks, i.e., *extrinsic* evaluation. This is often also referred to as *black-box* evaluation, because it does not focus on any specific module, but evaluates the system's performance as a whole. This paper presents one such evaluation experiment with focus on the issue of reusing resources such as questionnaires, and task and experiment designs. It then examines the problem of *glass-box*, i.e., module specific, evaluation in language generation, with focus on the problem of evaluating machine learning methods for text planning.

2 The System in Brief

HYLITE+ (Bontcheva and Wilks, 2001; Bontcheva, 2001b) is a dynamic hypertext system[3] that generates encyclopaedia-style explanations of terms in two specialised domains: chemistry and computers. The user interacts with the system in a Web browser by specifying a term she wants to look up. The system generates a

[1] For a comprehensive review see (Sparck Jones and Galliers, 1996).

[2] The same is not true for understanding tasks since they all operate on the same input, i.e., existing texts. So for example, two part-of-speech taggers or information extraction systems can be compared by running them on the same test corpus and measuring their relative performance.

[3] In *dynamic hypertext* page content and links are created on demand and are often adapted to the user and the previous interaction.

hypertext explanation of the term; further information can be obtained by following hypertext links or specifying another query. The system is based on applied NLG techniques, a re-usable user modelling component (VIEWGEN), and a flexible architecture with module feedback. The adaptivity is implemented on the basis of a user and a discourse models which are used to determine, for example, which concepts are unknown, so clarifying information can be included for them. The user model is updated dynamically, based on the user's interaction with the system. When a user registers with the system for the first time, her model is initialised from a set of stereotypes. The system determines which stereotypes apply on the basis of information provided by the user herself. If no such information is provided, the system assumes a novice user.

3 Extrinsic Evaluation of HYLITE+

Due to the fact that HYLITE+ generates hypertext which content and links are adapted to the user, it can be evaluated following strategies from two fields: NLG and adaptive hypertext. After reviewing the approaches, used for evaluation of the NLG and adaptive hypertext systems most similar to ours,e.g., (Cox et al., 1999), (Reiter et al., 1995), (Höök, 1998), we discovered that they were all evaluated extrinsically by measuring human performance on a set of tasks, given different versions of the system. The experiments were typically followed by an informal interview and/or questionnaire, used to gather some qualitative data, e.g., on the quality of the generated text.

Setting up and conducting such task-based experiments is costly and time-consuming, therefore we looked at opportunities for reusing materials and methodologies from previous evaluation experiments of similar systems from the two fields. This resulted in a substantial reduction of the time and effort needed to prepare the experiments. We also used the findings of some of these experiments in order to improve the design of our own evaluation. For example, (Cox et al., 1999) used pre-generated static pages as a baseline and the study reported that the difference in the two systems' response times might have influenced some of the results. Therefore, we chose instead to have both the baseline non-adaptive and the adaptive systems to generate the pages in real time, which eliminated the possible influence of the different response times.

3.1 Choosing the Main Goals of the Evaluation

The first issue that needs to be addressed when designing the extrinsic, or black-box, evaluation is to determine what are the goals of the experiment. Hypermedia applications are evaluated along three aspects: *interface look and feel*, *representation of the information structure*, and *application-specific information* (Wills et al., 1999). The information structure is concerned with the hypertext network (nodes and links) and navigation aids (e.g., site maps, links to related material, index). The application-specific information concerns the hypermedia content – text, images, audio and video. For our system there is no need to evaluate the interface, since HYLITE+ uses simple HTML and existing Web browsers (e.g. Netscape, Internet Explorer) as rendering tools. Therefore, the evaluation efforts were concentrated on the information content and navigational structure of the generated hypertext.

Information content was measured on the basis of:

- average *time to complete* each task;
- average number of *pages visited* per task;
- average number of *distinct pages* visited per task;
- percent of *correctly answered questions* per task;
- questionnaire results about *content* and *comprehension* of the generated pages;
- *user preference* for any of the systems.

The **navigational structure** was measured by the following metrics:

- average *time per page visited*;
- average *number of pages visited*;
- *total number of pages visited*;

- number of *links followed*;
- usage of the browser Back button;
- usage of the *system's topic list* to find information;
- observation and subjective *opinion on orientation*;
- subjective *opinion on navigation* and ease of finding information.

3.2 Choosing the Methodology

The experiment has a *repeated measures*, *task-based* design (also called within-subjects design), i.e., the same users interacted with the two versions of the system, in order to complete a given set of tasks. Prior to the experiment, the participants were asked to provide some *background information* (e.g., computing experience, familiarity with Web browsers, and electronic encyclopaedia) and fill in a *multiple choice pre-test*, that diagnosed their domain knowledge.

The design of the tasks follows the design used in the evaluation of two other adaptive hypermedia applications – PUSH (Höök, 1998) and (Wills et al., 1999). Each of the participants was first given a set of three tasks – each set contained one browsing, one problem-solving, and one information location task. The order was not randomised, because the browsing task was also intended as a task that would allow users to familiarise themselves with the system and the available information; it was not used for deriving the quantitative measures discussed above.

The participants performed the first set of tasks with the non-adaptive/adaptive system and then swapped systems for the second set of three tasks. The types of tasks – browsing, problem-solving, and information location – were chosen to reflect the different uses of hypermedia information.

Qualitative data and feedback were obtained using a *questionnaire* and *semi-structured interviews*, where the subjects could discuss their experience with the two systems. There were two main types of questions and statements: those related to the usability of the adaptive and baseline systems, e.g., statements like "I found the adaptive system difficult to use"; and those related to hypertext and navigation, e.g., links, text length, structure.

3.3 Results

Due to the small number of participants and the differences in their prior domain knowledge and browsing styles, the results obtained could not be used to derive a statistically reliable comparison between the measures obtained for the adaptive and the non-adaptive versions, but the quantitative results and user feedback are sufficiently encouraging to suggest that HYLITE+ adaptivity is of benefit to the user.

The most important outcome of this small-scale evaluation was that it showed the need to control not just for user's prior knowledge (e.g., novice, advanced), but also for hypertext reading style. Although previous studies of people browsing hypertext (e.g., (Nielsen, 2000)) have distinguished two types: *skimmers* and *readers*, in this experiment we did not control for that, because the tasks from which we derived the quantitative measures were concerned with locating information and problem solving, not browsing. Still, our results showed the need to control for this variable, regardless of the task type, because reading style influences some of the quantitative measures (e.g., task performance, mean time per task, number of visited pages, use of browser navigation buttons). Due to space limitations no further details can be provided in this paper, but see (Bontcheva, 2001a) for a detailed discussion.

3.4 Discussion

The methodology used for HYLITE's black-box evaluation was based on experience not only in the field of language generation, but also in the field of hypermedia, which motivated us to evaluate also the usability of the system and elicit the users' attitudes towards the intelligent behaviour of our generation system. This emphasis on usability, which comes from human-computer interaction, allowed us to obtain results which ultimately had implications for the architecture of our generation system (see (Bontcheva and Wilks, 2001) for further details) and which we would have not obtained otherwise. This leads us to believe that reuse of evaluation resources and methodologies from different, but related fields, can be beneficial for NLP systems in general.

On the other hand, even though evaluating the NLG system in a task-based fashion has had positive impact, there is still a need for glass-box evaluation on a module by module basis, especially using quantitative evaluation metrics, in order to be able to detect specific problems in the generation modules. This is the evaluation challenge that we discuss in the rest of the paper.

4 The Challenge: Automatic Quantitative Evaluation of Content Planners

Content planning, also called deep language generation, is the stage where the system needs to decide *what to say*, i.e., select some predicates encoding the semantics of the text to be generated, and then decide *when to say* them, i.e., choose an ordering of these predicates that will result in the generation of coherent discourse. Typically content plans are created manually by NLG experts in collaboration with domain specialists, using a corpus of target texts. However, this is a time consuming process, so recently researchers have started experimenting with using machine learning for content planning. This is the research area which we will investigate as part of building an NLG system for the e-science Grid project MIAKT[4]. The surface realisation module will be reused from HYLITE+, while the HYLITE+ content planner will be used as a baseline.

An integral part of the development of machine learning approaches to NLP tasks is the ability to perform automatic quantitative evaluation in order to measure differences between different configurations of the module and also allow comparative evaluation with other approaches. For example, the MUC corpora and the associated scoring tool are frequently used by researchers working on machine learning for Information Extraction both as part of the development process and also as means for comparison of the performance of different systems (see e.g., (Marsh and Perzanowski, 1998)). Similarly, automatic quantitative evaluation of content planners needs:

- an annotated corpus;
- an evaluation metric and a scoring tool, implementing this metric.

Below we will discuss each of these components and highlight the outstanding problems and challenges.

4.1 Evaluation Corpora for Content Planning

Research on content planning comes from two fields: document summarisation which uses some NLG techniques to generate the summaries; and natural language generation where the systems generate from some semantic representation, e.g., a domain knowledge base or numeric weather data. Here we review some work from these fields that has addressed the issue of evaluation corpora.

4.1.1 Previous Work

(Kan and Mckeown, 2002) have developed a corpus-trained summarisation system for indicative summaries. As part of this work they annotated manually 100 bibliography entries with indicative summaries and then used a decision tree learner to annotate automatically another 1900 entries with 24 predicates like `Audience`, `Topic`, and `Content`. For example, some annotations for the `Audience` predicate are: `For adult readers`; `This books is intended for adult readers`. The annotated texts are then used to learn the kinds of predicates present in the summaries, their ordering using bigram statistics, and surface realisation patterns.

(Barzilay et al., 2002) have taken the problem of learning sentence ordering for summarisation one step further by considering multi-document summarisation of news articles. Their experiments show that ordering is significant for text comprehension and there is no *one* ideal ordering, rather there is a set of acceptable orderings. Therefore, an annotated corpus which provides only one of the acceptable orderings is not sufficient to enable

[4] The MIAKT project is sponsored by the UK Engineering and Physical Sciences Research Council (grant GR/R85150/01) and involves the University of Southampton, University of Sheffield, the Open University, University of Oxford, and King's College London.

the system to differentiate between the many good orderings and the bad ones. To solve this problem they developed a corpus of multiple versions of the same content, each version providing an acceptable ordering. This corpus[5] consists of ten sets of news articles, two to three articles per event. Sentences were extracted manually from these sets and human subjects were asked to order them so that they form a readable text. In this way 100 orderings were acquired, 10 orderings per set. However, since this procedure involved a lot of human input, the construction of such a corpus on a larger scale is quite expensive.

The difference between the techniques used for summarisation and those used for generation is that the summarisation ones typically do not use very detailed semantic representations, unlike the full NLG systems. Consequently this means that a corpus annotated for summarisation purposes is likely to contain isufficient information for a full NLG application, while corpus with detailed semantic NLG annotation will most likely be useful for a summarisation content planner. Since the experience from building annotated corpora for learning ordering for summarisation has shown that they are expensive to build, then the creation of semantically annotated corpora for NLG is going to be even more expensive. Therefore, reuse and some automation are paramount.

So far, only very small semantically annotated corpora for NLG have been created. For example, (Duboue and McKeown, 2001) have collected an annotated corpus of 24 transcripts of medical briefings. They use 29 categories to classify the 200 tags used in their tagset. Each transcript had an average of 33 tags with some tags being much more frequent than others. Since the tags need to convey the semantics of the text units, they are highly domain specific, which means that any other NLG system or learning approach that would want to use this corpus for evaluation will have to be retargetted to this domain.

4.1.2 The Proposed Approach for MIAKT

As evident from this discussion, there are still a number of problems that need to be solved so that a semantically annotated corpus of a useful size can be created, thus enabling the comparative evaluation of different learning strategies and content planning components. Previous work has typically started from already existing texts/transcripts and then used humans to annotate them with semantic predicates, which is an expensive operation. In addition, the experience from the Information Extraction evaluations in MUC and ACE has shown that even humans find it difficult to annotate texts with deeper semantic information. For example, the interannotator variability on the scenario template task in MUC-7 was between 85.15 and 96.64 on the f-measures (Marsh and Perzanowski, 1998).

In the MIAKT project we will experiment with a different approach to creating an annotated corpus of orderings, which is similar to the approach taken by (Barzilay et al., 2002), where humans were given sentences and asked to order them in an acceptable way. Since MIAKT is a full NLG system we cannot use already existing sentences, as it was possible in their summarisation systems. Instead, we will use the HYLITE+ surface realiser to generate sentences for each of the semantic predicates and then provide users with a graphical editor, where they can re-arrange the ordering of these sentences by using drag and drop. In this way, there will be no need for the users to annotate with semantic information, because the system will have the corresponding predicates from which the sentences were generated. This idea is similar to the way in which language generation is used to support users with entering knowledge base content (Power et al., 1998). The proposed technique is called “What You See Is What You Meant” (WYSIWYM) and allows a domain expert to edit a NLG knowledge base reliably by interacting with a text, generated by the system, which presents both the knowledge already defined and the options for extending it. In MIAKT we will use instead the generator to produce the sentences, so the user only needs to enter their order. We will not need to use WYSIWYM editing for knowledge entry, because the knowledge base will already exist.

The difference between using generated sentences and sentences from human-written texts is that the human-written ones tend to be more com-

[5] Available at http://www.cs.columbia.edu/ noemie/ordering/.

plex and aggregate the content of similar predicates. This co-occurence information may be important, because, in a sense, it conveys stronger restrictions on ordering than those between two sentences. Therefore we would like to experiment with taking an already annotated corpus of human-authored texts, e.g., MUC-7 and compare the results achieved by using this corpus and a corpus of multiple orderings created by humans from the automatically generated sentences. In general, the question here is whether or not it is possible to reuse a corpus annotated for information extraction for the training of a content planning NLG component.

4.2 Evaluation Metrics

Previous work on learning order constraints has used human subjects for evaluation. For example, (Barzilay et al., 2002) asked humans to grade the summaries, while (Duboue and McKeown, 2001) manually analysed the derived constraints by comparing them to an existing text planner. However, this is not sufficient if different planners or versions of the same planner are to be compared in a quantitative fashion. In contrast, quantitative metrics for automatic evaluation of surface realisers have been developed (Bangalore et al., 2000) and they have been shown to correlate well with human judgement for quality and understandability.

These metrics are two kinds: using string edit distance and using tree-based metrics. The string edit distance ones measure the insertion, deletion, and substitution errors between the reference sentences in the corpus and the generated ones. Two different measures were evaluated and the one that treats deletions in one place and insertion in the other as a single movement error was found to be more appropriate. In the context of content planning we intend use the string edit distance metrics by comparing the proposition sequence generated by the planner against the "ideal" proposition sequence from the corpus.

The tree-based metrics were developed to reflect the intuition that not all moves are equally bad in surface realisation. Therefore these metrics use the dependency tree as a basis of calculating the string edit distances. However, it is not very clear whether this type of metrics will be applicable to the content planning problem given that we do not intend to use a planner that produces a tree-like structure of the text (as do for example RST-based planners, e.g., (Moore, 1995)).

If the reuse experiments in MIAKT are successful, we will make our evaluation tool publically available, together with the annotated corpus and the knowledge base of predicates, which we hope will encourage other researchers to use them for development and/or comparative evaluation of content planners.

5 Conclusion

In this paper we discussed the reuse of existing resouces and methodologies for extrinsic evaluation of language generation systems. We also showed that a number of challenges still exist in evaluation of NLG systems and, more specifically, evaluation of content planners. While other fields like machine translation and text summarisation already have some evaluation metrics and resources available for reuse, language generation has so far lagged behind and no comparative system evaluation has ever been done on a larger scale, e.g., text summarisation systems are compared in the DUC evaluation exercise. As a step towards comparative evaluation for NLG, we intend to make available the annotated corpus, evaluation metric(s) and tools to be developed as part of the recently started MIAKT project.

6 Acknowledgments

The work on MIAKT described here is being supported by the UK Engineering and Physical Sciences Research Council (grant GR/R85150/01). The work on HYLITE+ was supported by a PhD fellowship by the University of Sheffield and an Overseas Research Students Award. I also wish to thank Yorick Wilks and Hamish Cunningham for their comments on this work, the anonymous reviewers who helped me improve the paper, and the human evaluators who participated in the HYLITE+ experiments.

References

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation.

In *International Conference on Natural Language Generation (INLG 2000)*, Mitzpe Ramon, Israel.

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Artificial Intelligence Research*, 17:35–55.

Kalina Bontcheva and Yorick Wilks. 2001. Dealing with dependencies between content planning and surface realisation in a pipeline generation architecture. In *Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI'2001)*, Seattle, USA, August.

Kalina Bontcheva. 2001a. *Generating Adaptive Hypertext Explanations*. Ph.D. thesis, University of Sheffield.

Kalina Bontcheva. 2001b. Tailoring the content of dynamically generated explanations. In M. Bauer, P. Gmytrasiewicz, and J. Vassileva, editors, *User Modelling 2001*, volume 2109 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, Berling Heidelberg.

Richard Cox, Mick O'Donnell, and Jon Oberlander. 1999. Dynamic versus static hypermedia in museum education: an evaluation of ILEX, the intelligent labelling explorer. In Susanne P. Lajoie and Martial Vivet, editors, *Artificial Intelligence in Education: Open Learning Environment: New Computational Technologies to Support Learning, Exploration and Collaboration*, pages 181 – 188. IOS Press, Amsterdam ; Oxford. Papers from the 9th International Conference on Artificial Intelligence in Education (AI-ED 99).

Robert Dale and Chris Mellish. 1998. Towards evaluation in natural language generation. In *Proceedings of First International Conference on Language Resources and Evaluation*, pages 555 – 562, Granada, Spain, 28-30 May.

Pablo A. Duboue and Kathleen R. McKeown. 2001. Empirically estimanting order constraints for content planning in generation. In *Proceedings of ACL-EACL 2001*, Toulouse, France, July.

Kristina Höök. 1998. Evaluating the utility and usability of an adaptive hypermedia system. *Knowledge-Based Systems*, 10:311—319.

Min-Yen Kan and Kathleen R. Mckeown. 2002. Corpus-trained text generation for summarization. In *Proceedings of the Second International Conference on Natural Language Generation (INLG 2002)*.

Elaine Marsh and Dennis Perzanowski. 1998. MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. `http://www.itl.nist.gov/iaui/894.02/-related_projects/muc/index.html`.

Johanna D. Moore. 1995. *Participating in Explanatory Dialogues*. MIT Press, Cambridge, MA.

Jakob Nielsen. 2000. *Designing Web Usability: The Practice of Simplicity*. New Riders Publishing.

Richard Power, Donia Scott, and Richard Evans. 1998. What you see is what you meant: direct knowledge editings with natural language feedback. In *13th European Conference on Artificial Intelligence (ECAI'98)*, pages 677–681. John Wiley and Sons.

Ehud Reiter, Chris Mellish, and Jon Levine. 1995. Automatic generation of technical documentation. *Journal of Applied Artificial Intelligence*, 9(3):259–287.

Karen Sparck Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer Verlag, Berlin, Heidelberg.

Yorick A. Wilks. 1992. Where am I coming from: The reversibility of analysis and generation in natural language processing. In Martin Puetz, editor, *Thirty Years of Linguistic Evolution*. John Benjamins.

G. B. Wills, I. Heath, R.M. Crowder, and W. Hall. 1999. User evaluation of an industrial hypermedia application. Technical report, M99/2, University of Southampton. `http://www.bib.ecs.soton.ac.uk/data/1444/html/html/`.