Quality of service and communicative competence in NLG evaluation

Kristiina JOKINEN
University of Helsinki and University of Tampere
Finland
Kristiina.Jokinen@helsinki.fi

Abstract

The paper discusses quality of service evaluation which emphasises the user's experience in the evaluation of system functionality and efficiency. For NLG systems, an important quality feature is communicatively adequate language generation, which affects the users' perception of the system and consequently, evaluation results. The paper drafts an evaluation task that aims at measuring quality of service, taking the system's communicative competence into account.

1 Introduction

The state of the art Natural Language Generation systems can generate summaries and short texts which exhibit variation in sentence structure, anaphoric references, and amount of information included in the text, as well as some adaptation to different users. The starting point can be a structured database or a specifically designed representation, while the output can be text and also spoken language given a suitable text-to-speech component. The standard architecture (Reiter and Dale 2000) provides basis for generation technology which ranges from rule-based systems via XML transformations to statistical generators.

As the academic research extends out to industrial markets, high priority should be given to evaluation techniques. The goal is not only to provide diagnostic feedback about the system performance, but to enable researchers and developers to test and compare different techniques and approaches with respect to generation tasks. Moreover, evaluation allows self-assessment to guide and focus future research. To potential users, customers, and manufacturers evaluation offers slightly different benefits: with an increased number of applications which can integrate an NLG component, evaluation provides surveys of the available generation components and their suitability to particular practical tasks. Vivid interest has thus been shown in finding suitable evaluation tasks and methods, e.g. in the recent workshop (Dale and White 2007), resulting in the Shared Task Evaluation Campaign.

Setting up a framework that addresses (some of) the motivations and requirements for evaluation is a complex task, and to structure the goals, three fundamental questions need to be asked:

1. Definition: what is it that we are interested in and require from an NLG?
2. Measures: which specific property of the system and its performance can we identify with the goal and use in evaluation?
3. Method: how to determine the appropriate value for a given measure and a given NLG system? Can the results predict properties of future systems?

The paper seeks to answer these questions from the perspective of communicative systems. The starting point is that generation products are not generated or read in void: they are produced as communicative acts in various communicative situations. NLG evaluation thus resembles that of dialogue systems: besides task completeness, one need to measure intangible factors such as impact of the text on the user and the user's expectations and satisfaction concerning the output. Moreover, it is important to measure the quality of service, or the system's effectiveness as perceived by the users through their experience with the system.

The paper starts with the definition (Section 2), continues with a discussion about metrics (Section 3) and methods (Section 4), and concludes with a concrete evaluation proposal (Section 5).

2 Definition of evaluation

We distinguish between assessment and evaluation (Möller 2007), or performance and adequacy evaluation (Hirschman and Thompson 1997). Assessment refers to the measurement of system

performance in specific areas with respect to certain criteria, whereas evaluation refers to the determination of the fitness of a system for a specific purpose. Assessment also requires well-defined baseline performance for the comparison of alternative technologies, but adequacy evaluation is mainly determined by the user needs.

Performance of a system should be distinguished from its quality. According to Möller (2007), performance is "an ability of the module to provide the function it has been designed for". Quality, on the other hand, is determined by the perceptions of the system users. It "results from a perception and a judgment process, in which the perceiving subject (e.g. a test user of the system) establishes a relationship between the perceptive event and what he/she expects or desires from the service". Quality is thus everything that is perceived by the user with respect to what she expects from the system. User factors like attitude, emotions, experience, task/domain knowledge, etc., will influence the perception of quality.

It is also common to talk about usability of a system, referring to issues that deal with effectiveness and user satisfaction. The ISO definition of usability goes as follows:

> *The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.*

In usability testing, a test supervisor observes real users as they use the product in real tasks, and analyses the results for the purpose of learning how to improve the product's usability. Usability and quality evaluations seem similar and in practice, both use similar techniques (user questionnaires, interviews). An important difference is that in the latter, user expectations are explicitly taken into consideration by relating the perceived system properties to the expectations that the user had about the system. A system can be useful and usable, but in order to become utilised it must also provide a special value for the users. Considering NLG systems, it may be difficult to obtain usability or quality information as such: usually generators are not stand-alone systems but components of bigger systems. However, if a NL generator is a component e.g. in a dialogue system, its quality affects the quality of the whole system. Thus evaluating generators in the context of the same dialogue system, it is possible to obtain explicit quality judgements of the NLG component, too.

3 Evaluation metrics

Different measures can be used depending on the goals and the system itself. For individual components, quantifiable measures and glass-box evaluation provide useful information about how to optimize a component further, or which component to choose over another. Performance can be quantified by providing suitable metrics for:

- effectiveness for the task (the system provides the desired information),
- efficiency (time needed to complete the task, or the effort required from the user).

In NLG, such metrics deal with the time it takes to generate a text, or the length and complexity of sentences and syntactic constructions. The BLEU type metrics can be used to automatically compare generated texts with the target as in translation studies, while comprehensibility judgement tests can be used to determine if the texts effectively communicate the intended message.

However, the system is always situated in a context. Glass-box evaluation does not tell us how the system functions as a whole when used in its context, and black-box evaluation, focusing on the system's functioning and impact on users in real situations, should thus complement performance evaluation. This includes measures for:

- satisfaction of the user (experienced comfort, pleasantness, or joy-of-use),
- utility (involves cost measures),
- acceptability (whether a potential user is willing to use the system).

Accordingly, NLG systems should be evaluated with respect to the context where the generated text is meant to appear, and by the users who are likely to use the system. This brings us back to the evaluation of the quality of the system: we need to determine quality features which capture differences in the users' perception of the system and consequently, contribute to the system quality. Since language is used to communicate ideas and meanings, communicative competence is one of the most visible aspects of language- based applications. The quality of the system can thus be measured by its communicative capability: how accurately and reliably the intended message is conveyed to the user in a given context.

4 Evaluation methods

Good evaluation methods are generic in that they allow comparison of different systems and also predictions to be made about their future versions. One of the frameworks used in dialogue system evaluation is the PARADISE framework (Walker et al. 2000) which learns the evaluation parameters from the data and produces a performance function which specifies relative contributions of the various cost factors to the overall performance. The goal of the evaluation is to maximize user satisfaction by maximizing task success and minimizing task cost measured using various task and dialogue metrics. PARADISE is a rigorous framework, but the data collection and annotation cost for deriving the performance function is high. When considering development of complex systems, or the need for evaluation of prototypes with a large number of users, semi-automatic evaluation with less manual annotation would be preferable. Möller (2007) also points out that it may be too simplistic to relate interaction parameters in a linear fashion, since quality is a multi-dimensional property of the system.

As the correlation between the designers' and the users' views of the system can be weak, a comprehensive glass-box evaluation cannot be based solely on the interaction parameters, i.e. assessing how the designed system functionalities work with respect to various users, but also the way how the users experience the system should be explored. Extending the PARADISE type evaluation, Möller (2007) presents a taxonomy of quality aspects (for speech-based interactive systems but it can be applied to NLG systems, too) which includes quality evaluation of the system from the user's perspective. It aims at generic prediction power concerning quality aspects (categories of quality) and quality features (perceptual dimensions). The extra value of the system can be approximated by comparing the users' actual experience of the system with the expectations they had of the system before its evaluation (cf. case studies reported in Möller 2007; Jokinen and Hurtig 2006). The differences are indicative of the users' disappointments and satisfactions in regard to the quality aspects, and by applying more complex algorithms to calculate parameter dependencies, the model's prediction power can also be improved.

5 NLG evaluation

For a given generation task, there is usually not only one correct solution but several: the text may be constructed in more than one way. This kind of variation is typical for language-based applications in general: there is no "golden standard" to compare the results with, but the ratings about the success of a contribution depend on the situation and the evaluator's attitudes and likings. In interactive system development, the success of responses is usually related to the "contextual appropriateness", based on Grice's Cooperation Principle, and made explicit in the recommendations and best practice guidelines (see e.g. Gibbon et al., 1997). Analogously, the task in the NLG evaluation is not to measure various outputs in regard to one standard solution but rather, to provide a means to abstract away from the details of the individual outputs into the space of quality features that characterise the contextual appropriateness of the texts, i.e. the system's communicative competence with respect to the user's expectations and experience.

As mentioned, one way to organise this kind of quality evaluation is to integrate the NLG system in an interactive prototype and evaluate the output which is produced as a response to a particular communicative goal.[1] The goals can be rather straightforward information providing goals with the topic dealing with weather forecasts or traffic information (what is the weather like in X, tell about the weather in Y in general, the wind speed later in the afternoon, etc.), or more complex ones that require summaries of news texts or comparisons of database items (e.g. how the weather is expected to change tomorrow, compare air quality in X and Y, how has precipitation changed in recent years; how do I get to X). They simulate plausible "real" situations in which to evaluate one's experience of the system, and also provide discourse contexts in which to judge the appropriateness of the generated text.

The goals can be directly mapped to an interface language that enables the NLG system to be called with the selected parameter settings. A structured database can be provided as the shared

[1] The task resembles the one proposed by Walker (2007), but has been independently sketched at the ENLGW 2005 in Aberdeen with Stephan Busemann.

input, and output is a short text, possibly spoken, which can be varied by allowing the users to choose between a short or a full text, or if they wish the text to appear in a mobile phone screen (concise) or on a webpage (verbose).

The user is also instructed to evaluate each generated text(s) by answering questions that ask the user's opinion e.g. of the comprehensibility of the text, its syntactic correctness, acceptability, appropriateness, reliability, style, and the user's overall impression. The questions may also ask if the text is informative or ambiguous, if it gives too much information (what could be left out) or too little information (what is missing), and if it conforms to the user's expectations.

In the beginning of the evaluation session, before their actual experience with the system, the users are asked to estimate their familiarity with generation systems and, by going through the evaluation questions, to describe how quick, informative, fluent, and useful they expect the system to be. All the answers are given in a 5-point Likert scale, and in the analysis, evaluation answers are related to those of expectations.

Evaluation can be performed via web-based interaction (cf. the Blizzard Challenge for evaluating corpus-based speech synthesis: http://festvox.org/blizzard/). Using the web, it is possible to recruit participants from different countries, and they can also rank the texts anywhere any time. Given that several generators will take part in the evaluation, software share and installation can also be simplified. A drawback is that there is no control over the users or their environment: the users may not complete the evaluation or they may fill in random values. Web connection may also break, and degrade the system performance and speed.

6 Conclusion

The paper has discussed the quality of service evaluation which emphasises the user's perception of the system in the evaluation setup. The system's communicative competence, i.e. ability to provide reliable and useful information, is regarded as an important quality feature, and a web-based evaluation set-up is drafted in order to evaluate the quality of NLG systems with respect to their communicative capability. We finish the paper with some general questions concerning the evaluation setup.

- How realistic interactions are necessary in order to get reliable evaluation data? E.g. should the system provide meta-communication besides the factual text?
- The users should not be burdened with too many similar parameter settings. How many different features can be varied to maintain user interest and yet to guarantee systematic variation and collection of enough data?
- Even though it is not necessary to define an "ideal" text for each communicative goal, some guidelines may be useful to describe e.g. necessary/optional features of the generated texts for the participating systems.

References

Dale, R. and M. White (Eds) 2007. Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, http://www.ling.ohio-state.edu/~mwhite/nlgeval07/

Gibbon, D., R. Moore and R. Winski (Eds.) 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, New York.

Hirschman, L. and H. Thompson 1997. Overview of evaluation in speech and natural language processing. In: Cole, R., Mariani, J., Uszkoreit, H., Zaenen A., and Zue, V. (Eds.) 1997. *Survey of the State of the Art in Human Language Technology*, Cambridge University Press and Giardini Editori, Pisa.

Jokinen, K. 1996. Adequacy and Evaluation. *Procs of the ECAI-96 workshop "Gaps and Bridges: New Directions in Planning and Natural Language Generation"*. Budapest, Hungary. pp. 105-107.

Jokinen, K. and T. Hurtig 2006. User Expectations and Real Experience on a Multimodal Interactive System. *Procs of Interspeech-2006*.

Möller, S. 2007. Evaluating Speech-based Interactive Systems. In: Fang, C. and K. Jokinen (Eds.) *New Trends in Speech-based Interactive Systems*. Springer Publishers.

Reiter, E and R.Dale 2000 *Building Natural Language Generation Systems*. Cambridge University Press.

Walker, M, Kamm, C, and Litman, D. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6:363–377.

Walker, M. 2007. *Share and Share Alike: Resources for Language Generation*. In R. Dale and M. White (Eds.) pp. 28-30.