Validating the web-based evaluation of NLG systems

Alexander Koller
Saarland U.
koller@mmci.uni-saarland.de

Kristina Striegnitz
Union College
striegnk@union.edu

Donna Byron
Northeastern U.
dbyron@ccs.neu.edu

Justine Cassell
Northwestern U.
justine@northwestern.edu

Robert Dale
Macquarie U.
Robert.Dale@mq.edu.au

Sara Dalzel-Job
U. of Edinburgh
S.Dalzel-Job@sms.ed.ac.uk

Jon Oberlander
U. of Edinburgh
{J.Oberlander|J.Moore}@ed.ac.uk

Johanna Moore
U. of Edinburgh

Abstract

The GIVE Challenge is a recent shared task in which NLG systems are evaluated over the Internet. In this paper, we validate this novel NLG evaluation methodology by comparing the Internet-based results with results we collected in a lab experiment. We find that the results delivered by both methods are consistent, but the Internet-based approach offers the statistical power necessary for more fine-grained evaluations and is cheaper to carry out.

1 Introduction

Recently, there has been an increased interest in evaluating and comparing natural language generation (NLG) systems on shared tasks (Belz, 2009; Dale and White, 2007; Gatt et al., 2008). However, this is a notoriously hard problem (Scott and Moore, 2007): Task-based evaluations with human experimental subjects are time-consuming and expensive, and corpus-based evaluations of NLG systems are problematic because a mismatch between human-generated output and system-generated output does not necessarily mean that the system's output is inferior (Belz and Gatt, 2008). This lack of evaluation methods which are both effective and efficient is a serious obstacle to progress in NLG research.

The GIVE Challenge (Byron et al., 2009) is a recent shared task which takes a third approach to NLG evaluation: By connecting NLG systems to experimental subjects over the Internet, it achieves a true task-based evaluation at a much lower cost. Indeed, the first GIVE Challenge acquired data from over 1100 experimental subjects online. However, it still remains to be shown that the results that can be obtained in this way are in fact comparable to more established task-based evaluation efforts, which are based on a carefully selected subject pool and carried out in a controlled laboratory environment. By accepting connections from arbitrary subjects over the Internet, the evaluator gives up control over the subjects' behavior, level of language proficiency, cooperativeness, etc.; there is also an issue of whether demographic factors such as gender might skew the results.

In this paper, we provide the missing link by repeating the GIVE evaluation in a laboratory environment and comparing the results. It turns out that where the two experiments both find a significant difference between two NLG systems with respect to a given evaluation measure, they always agree. However, the Internet-based experiment finds considerably more such differences, perhaps because of the higher number of experimental subjects ($n = 374$ vs. $n = 91$), and offers other opportunities for more fine-grained analysis as well. We take this as an empirical validation of the Internet-based evaluation of GIVE, and propose that it can be applied to NLG more generally. Our findings are in line with studies from psychology that indicate that the results of web-based experiments are typically consistent with the results of traditional experiments (Gosling et al., 2004). Nevertheless, we do find and discuss some effects of the uncontrolled subject pool that should be addressed in future Internet-based NLG challenges.

2 The GIVE Challenge

In the GIVE scenario (Byron et al., 2009), users try to solve a treasure hunt in a virtual 3D world that they have not seen before. The computer has complete information about the virtual world. The challenge for the NLG system is to generate, in real time, natural-language instructions that will guide the users to the successful completion of their task.

From the perspective of the users, GIVE consists in playing a 3D game which they start from a website. The game displays a virtual world and allows the user to move around in the world and manipulate objects; it also displays the generated

instructions. The first room in each game is a tutorial room in which users learn how to interact with the system; they then enter one of three evaluation worlds, where instructions for solving the treasure hunt are generated by an NLG system. Players can either finish a game successfully, lose it by triggering an alarm, or cancel the game at any time.

When a user starts the game, they are randomly connected to one of the three worlds and one of the NLG systems. The GIVE-1 Challenge evaluated five NLG systems, which we abbreviate as A, M, T, U, and W below. A running GIVE NLG system has access to the current state of the world and to an automatically computed plan that tells it what actions the user should perform to solve the task. It is notified whenever the user performs some action, and can generate an instruction and send it to the client for display at any time.

3 The experiments

The web experiment. For the GIVE-1 challenge, 1143 valid games were collected over the Internet over the course of three months. These were distributed over three evaluation worlds (World 1: 374, World 2: 369, World 3: 400). A game was considered valid if the game client didn't crash, the game wasn't marked as a test run by the developers, and the player completed the tutorial.

Of these games, 80% were played by males and 10% by females (the remaining 10% of the participants did not specify their gender). The players were widely distributed over countries: 37% connected from IP addresses in the US, 33% from Germany, and 17% from China; the rest connected from 45 further countries. About 34% of the participants self-reported as native English speakers, and 62% specified a language proficiency level of at least "expert" (3 on a 5-point scale).

The lab experiment. We repeated the GIVE-1 evaluation in a traditional laboratory setting with 91 participants recruited from a college campus. In the lab, each participant played the GIVE game once with each of the five NLG systems. To avoid learning effects, we only used the first game run from each subject in the comparison with the web experiment; as a consequence, subjects were distributed evenly over the NLG systems. To accommodate for the much lower number of participants, the laboratory experiment only used a single game world – World 1, which was known from the online version to be the easiest world.

Among this group of subjects, 93% self-rated their English proficiency as "expert" or better; 81% were native speakers. In contrast to the online experiment, 31% of participants were male and 65% were female (4% did not specify their gender).

Results: Objective measures. The GIVE software automatically recorded data for five objective measures: the percentage of successfully completed games and, for the successfully completed games, the number of instructions generated by the NLG system, of actions performed by the user (such as pushing buttons), of steps taken by the user (i.e., actions plus movements), and the task completion time (in seconds).

Fig. 1 shows the results for the objective measures collected in both experiments. To make the results comparable, the table for the Internet experiment only includes data for World 1. The task success rate is only evaluated on games that were completed successfully or lost, not cancelled, as laboratory subjects were asked not to cancel. This brings the number of Internet subjects to 322 for the success rate, and to 227 (only successful games) for the other measures.

Task success is the percentage of successfully completed games; the other measures are reported as means. The chart assigns systems to groups A through C or D for each evaluation measure. Systems in group A are better than systems in group B, and so on; if two systems have no letter in common, the difference between them is significant with $p < 0.05$. Significance was tested using a χ^2-test for task success and ANOVAs for instructions, steps, actions, and seconds. These were followed by post hoc tests (pairwise χ^2 and Tukey) to compare the NLG systems pairwise.

Results: Subjective measures. Users were asked to fill in a questionnaire collecting subjective ratings of various aspects of the instructions. For example, users were asked to rate the overall quality of the direction giving system (on a 7-point scale), the choice of words and the referring expressions (on 5-point scales), and they were asked whether they thought the instructions came at the right time. Overall, there were twelve subjective measures (see (Byron et al., 2009)), of which we only present four typical ones for space reasons.

For each question, the user could choose not to answer. On the Internet, subjects made considerable use of this option: for instance, 32% of users

	Objective Measures										Subjective Measures							
	task success		instructions		steps		actions		seconds		overall		choice of words		referring expressions		timing	
A	91%	A	83.4	B	99.8	A	9.4	A	123.9	A	4.7	A	4.7	A	4.7	A	81%	A
M	76%	B	68.1	A	145.1	B	10.0	AB	195.4	BC	3.8	AB	3.8	B	4.0	B	70%	ABC
T	85%	AB	97.8	C	142.1	B	9.7	AB	174.4	B	4.4	B	4.4	AB	4.3	AB	73%	AB
U	93%	AB	99.8	C	142.6	B	10.3	B	194.0	BC	4.0	B	4.0	B	4.0	B	51%	C
W	24%	C	159.7	D	256.0	C	9.6	AB	234.1	C	3.8	AB	3.8	B	4.2	AB	50%	BC
A	100%	A	78.2	AB	93.4	A	9.9	A	143.9	A	5.7	A	4.7	A	4.8	A	92%	A B
M	95%	A	66.3	A	141.8	B	10.5	A	211.8	B	5.4	A	3.8	B	4.3	A	95%	A B
T	93%	A	107.2	CD	134.6	B	9.6	A	205.6	B	4.9	A	4.5	A B	4.4	A	64%	A B
U	100%	A	88.8	BC	128.8	B	9.8	A	195.1	AB	5.7	A	4.7	A	4.3	A	100%	A
W	17%	B	134.5	D	213.5	C	10.0	A	252.5	B	5.0	A	4.5	A B	4.0	A	100%	B

Figure 1: Objective and selected subjective measures on the web (top) and in the lab (bottom).

didn't fill in the "overall evaluation" field of the questionnaire. In the laboratory experiment, the subjects were asked to fill in the complete questionnaire and the response rate is close to 100%.

The results for the four selected subjective measures are summarized in Fig. 1 in the same way as the objective measures. Also as above, the table is based only on successfully completed games in World 1. We will justify this latter choice below.

	Web		
	games	reported	mean
success	227 = 61%	93%	4.9
lost	92 = 24%	48%	3.4
cancelled	55 = 15%	16%	3.3
	Lab		
	# games	reported	mean
success	73 = 80%	100%	5.4
lost	18 = 20%	94%	3.3
cancelled	0	–	–

Figure 2: Skewed results for "overall evaluation".

4 Discussion

The primary question that interests us in a comparative evaluation is which NLG systems performed significantly better or worse on any given evaluation measure. In the experiments above, we find that of the 170 possible significant differences (= 17 measures $\times$ 10 pairs of NLG systems), the laboratory experiment only found six that the Internet-based experiment didn't find. Conversely, there are 26 significant differences that only the Internet-based experiment found. But even more importantly, all pairwise rankings are consistent across the two evaluations: Where both systems found a significant difference between two systems, they always ranked them in the same order. We conclude that the Internet experiment provides significance judgments that are comparable to, and in fact more precise than, the laboratory experiment.

Nevertheless, there are important differences between the laboratory and Internet-based results. For instance, the success rates in the laboratory tend to be higher, but so are the completion times. We believe that these differences can be attributed to the demographic characteristics of the participants. To substantiate this claim, we looked in some detail at differences in gender, language proficiency, and questionnaire response rates.

First, the gender distribution differed greatly between the Internet experiment (10% female) and the laboratory experiment (65% female). This is relevant because gender had a significant effect on task completion time (women took longer) and on six subjective measures including "overall evaluation" in the laboratory. We speculate that the difference in task completion time may be related to well-known gender differences in processing navigation instructions (Moffat et al., 1998).

Second, the two experiments collected data from subjects of different language proficiencies. While 93% of the participants in the laboratory experiment self-rated their English proficiency as "expert" or better, only 62% of the Internet participants did. This partially explains the lower task success rates on the Internet, as Internet subjects with English proficiencies of 3–5 performed significantly better on "task success" than the group with proficiencies 1–2. If we only look at the results of high-English-proficiency subjects on the Internet, the success rates for all NLG systems except W rise to at least 86%, and are thus close to the laboratory results.

Finally, the Internet data are skewed by the tendency of unsuccessful participants to not fill in the questionnaire. Fig. 2 summarizes some data about the "overall evaluation" question. Users who didn't complete the task successfully tended to judge the

systems much lower than successful users, but at the same time tended not to answer the question at all. This skew causes the mean subjective judgments across all Internet subjects to be artificially high. To avoid differences between the laboratory and the Internet experiment due to this skew, Fig. 1 includes only judgments from successful games.

In summary, we find that while the two experiments made consistent significance judgments, and the Internet-based evaluation methodology thus produces meaningful results, the absolute values they find for the individual evaluation measures differ due to the demographic characteristics of the participants in the two studies. This could be taken as a possible deficit of the Internet-based evaluation. However, we believe that the opposite is true. In many ways, an online user is in a much more natural communicative situation than a laboratory subject who is being discouraged from cancelling a frustrating task. In addition, every experiment – whether in the laboratory or on the Internet – suffers from some skew in the subject population due to sampling bias; for instance, one could argue that an evaluation that is based almost exclusively on native speakers in universities leads to overly benign judgments about the quality of NLG systems.

One advantage of the Internet-based approach to data collection over the laboratory-based one is that, due to the sheer number of subjects, we can detect such skews and deal with them appropriately. For instance, we might decide that we are only interested in the results from proficient English speakers and ignore the rest of the data; but we retain the option to run the analysis over all participants, and to analyze how much each system relies on the user's language proficiency. The amount of data also means that we can obtain much more fine-grained comparisons between NLG systems. For instance, the second and third evaluation world specifically exercised an NLG system's abilities to generate referring expressions and navigation instructions, respectively, and there were significant differences in the performance of some systems across different worlds. Such data, which is highly valuable for pinpointing specific weaknesses of a system, would have been prohibitively costly and time-consuming to collect with laboratory subjects.

5 Conclusion

In this paper, we have argued that carrying out task-based evaluations of NLG systems over the Internet is a valid alternative to more traditional laboratory-based evaluations. Specifically, we have shown that an Internet-based evaluation of systems in the GIVE Challenge finds consistent significant differences as a lab-based evaluation. While the Internet-based evaluation suffers from certain skews caused by the lack of control over the subject pool, it does find more differences than the lab-based evaluation because much more data is available. The increased amount of data also makes it possible to compare the quality of NLG systems across different evaluation worlds and users' language proficiency levels.

We believe that this type of evaluation effort can be applied to other NLG and dialogue tasks beyond GIVE. Nevertheless, our results also show that an Internet-based evaluation risks certain kinds of skew in the data. It is an interesting question for the future how this skew can be reduced.

References

A. Belz and A. Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08:HLT, Short Papers*, pages 197–200, Columbus, Ohio.

A. Belz. 2009. That's nice ... what can you do with it? *Computational Linguistics*, 35(1):111–118.

D. Byron, A. Koller, K. Striegnitz, J. Cassell, R. Dale, J. Moore, and J. Oberlander. 2009. Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation (Special session on Generation Challenges)*.

R. Dale and M. White, editors. 2007. *Proceedings of the NSF/SIGGEN Workshop for Shared Tasks and Comparative Evaluation in NLG*, Arlington, VA.

A. Gatt, A. Belz, and E. Kow. 2008. The TUNA challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Natural Language Generation Conference (INLG'08)*, pages 198–206.

S. D. Gosling, S. Vazire, S. Srivastava, and O. P. John. 2004. Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, 59:93–104.

S. Moffat, E. Hampson, and M. Hatzipantelis. 1998. Navigation in a "virtual" maze: Sex differences and correlation with psychometric measures of spatial ability in humans. *Evolution and Human Behavior*, 19(2):73–87.

D. Scott and J. Moore. 2007. An NLG evaluation competition? Eight reasons to be cautious. In *(Dale and White, 2007)*.