Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk

F. Jurčíček, S. Keizer, M. Gašić, F. Mairesse, B. Thomson, K. Yu, and S. Young

Engineering Department, Cambridge University, CB2 1PZ, UK

{fj228, sk561, mg436, f.mairesse, brmt2, ky219, sjy}@eng.cam.ac.uk

Abstract

This paper describes a framework for evaluation of spoken dialogue systems. Typically, evaluation of dialogue systems is performed in a controlled test environment with carefully selected and instructed users. However, this approach is very demanding. An alternative is to recruit a large group of users who evaluate the dialogue systems in a remote setting under virtually no supervision. Crowdsourcing technology, for example Amazon Mechanical Turk (AMT), provides an efficient way of recruiting subjects. This paper describes an evaluation framework for spoken dialogue systems using AMT users and compares the obtained results with a recent trial in which the systems were tested by locally recruited users. The results suggest that the use of crowdsourcing technology is feasible and it can provide reliable results.

Index Terms: crowdsourcing, spoken dialogue systems, evaluation

1. Introduction

Despite recent progress in spoken dialogue systems development, there is a re-occurring problem with evaluating new ideas in this area. Ideally, the proposed techniques should be evaluated with real users, requiring the recruitment of a group of carefully selected subjects. Prior to the evaluation, the subjects have to be instructed on how to rate the dialogue systems. During the evaluation, the subjects have to be supervised to achieve consistent ratings. Because this process is time-consuming and costly, dialogue systems are very often evaluated in interaction with a simulated user, rather than real users [1, 2, 3]. However, this raises a question regarding the potential discrepancy between simulated and real user behaviour [4]. Hence, there is a need for a methodology for efficient evaluation on real users, allowing such evaluations to be held more frequently and at modest cost.

This work describes an evaluation framework for spoken dialogue systems which uses crowdsourcing technology for recruiting and managing large groups of users/subjects. The main benefits of this approach are that it has access to a vast base of potential users, the evaluation starts practically immediately when requested, the automation of paying the subjects eases the management of the evaluation, and finally the cost of the evaluation is greatly reduced by offering the work in a highly competitive market place.

The paper is organised as follows. Section 2 describes a typical evaluation in a controlled test environment. Section 3 details the proposed framework for remote evaluation of dialogue systems using Amazon Mechanical Turk. The results obtained by the proposed framework are evaluated and discussed in Section 4. Finally, the paper is concluded in Section 5.

2. Controlled environment evaluation

In the dialogue system evaluations described in [5, 6], the subjects were recruited among students or employees of the university via advertisements distributed on local mailing lists and bulletins. The subjects that agreed to participate in the evaluation were invited to the lab on a particular date and time to do a series of tasks in a one-hour time slot. The full schedule consisted of several time slots in several parallel sessions and could therefore accommodate a large group of subjects in a few days. The subjects were expected to complete at least 20 dialogues during the session, under continuous supervision of a research team member. For each of these dialogues, subjects were provided with a specific scenario in the tourist information domain, describing what kind of venue the user should ask for (for example, a cheap Chinese restaurant in the city centre).

At the beginning of each session, the supervisor gave some general instructions to a subject, including how to use the microphone and headset, how to interact with the dialogue systems, and how to answer the questions from the questionnaire used for the subjective evaluation. During the session, the supervisor advised the subject on how to consistently rate the systems and checked whether they followed the task descriptions. For each completed dialogue, the subject was asked to provide feedback about that dialogue via a questionnaire. This included questions about the perceived performance of different aspects of the system as well as of the system as a whole. The question “Did you find all the information you were looking for?” aimed to evaluate the overall performance of the dialogue system.

3. Remote evaluation

The evaluation of dialogue systems with users in a controlled environment as described in Section 2 is very demanding. An alternative approach is to use crowdsourcing technology to recruit a large number of users which work under virtually no supervision. The first platform developed for large scale crowdsourcing is Amazon Mechanical Turk (AMT) [7]. The AMT’s workforce allows a large number of small tasks to be completed in a very short period of time. The platform has been used for a great variety of tasks, including for example image labelling [8], semantic labelling [9], and audio transcription [10, 11]. More recently, the technology has been used in the context of dialogue system evaluation, where user judgements for a set of pre-collected dialogues were obtained [12].

In the AMT terminology, the tasks are referred to as HITs (Human Intelligence Tasks), the users performing work on HITs are called workers, and those publishing HITs are called requesters. This terminology will be followed in the remainder of this paper. AMT provides infrastructure for presenting HITs to workers, collecting results, and eventually paying the workers. The interaction with AMT is performed via the Internet and requesters have to design a web interface which enables the workers to complete the published HITs.

In order to use AMT for dialogue system evaluation, initially, a fully web based framework was developed. The framework used a Java applet implementing a software phone to carry voice over the Internet. It was assumed that such a system would be cheap to operate and easy to use. However, there were numerous problems with this approach. First, many workers had problems connecting to the evaluated dialogue systems as their Internet service providers were blocking the voice connection. Second, it was observed that many users had difficulties to connect and operate their headset and external microphone, which they were required to use in order to prevent echo in the audio recordings. Finally, most of the AMT users calling the evaluated systems turned out to be non-native speakers of English. The analysis of the IP addresses of these workers revealed that they were not from the USA. The English proficiency of these workers was very low which resulted in ungrammatical and unnatural sentences. Also, their speech was heavily accented.

Therefore, a telephone based evaluation framework was implemented. Workers were provided with a phone number to call, using either a land line or a mobile phone and the web interface was only used to present the tasks and collect feedback.

3.1. Telephone framework

In the telephone framework, the web interface contains an introduction, a task description, a telephone number which should be called, and a feedback form. The introduction briefly describes the evaluated dialogue systems, emphasises that the workers should be native English speakers, and provides an example of a typical conversation. The task description presents one randomly selected task from a set of pre-generated tasks. The set of tasks is produced automatically by a template-based natural language generator using randomly sampled constraints from a domain ontology.

In order to distribute incoming calls evenly across the systems, a private branch exchange (PBX)[1] was used, which randomly routed each call to one of the systems. With this setup, workers were able to use the redial functionality on their phone without affecting the distribution of the systems called.

To prevent users from submitting feedback without calling any of the systems, the feedback form is only enabled after a unique code has been entered. This code is given to a worker by the dialogue system at the end of a call, after the actual dialogue, and is also stored in the call log. The code is issued only if some non-trivial interaction with the dialogue system has taken place. When the code is entered and successfully verified, the feedback form is enabled and a worker can submit their feedback. This same code is also used to match the feedback information with the correct call log containing the system ID.

To make the dialogue system evaluation attractive to USA workers, a toll-free phone number was purchased. As long as workers use land phones, the calls are free. From the cost perspective, the price associated with renting a toll-free phone number and paying for incoming calls is modest. It is relatively easy to find providers which offer toll-free USA phone numbers for less than $15 a month and $0.02 per minute of incoming calls.

When the telephone framework was tested, the number of calls was about 120 per day. This is significantly higher compared to the software phone framework. Although no measures were taken to prevent calls from outside the USA, all the workers were native English speakers based on manual inspection of the data. Note that it is possible to call a toll-free phone number for free with an Internet based service such as Skype from any country.

4. Evaluation of reliability of AMT workers

As the AMT workers complete the tasks without supervision, a question therefore arises as to what extent the results obtained with the AMT workers are accurate. To evaluate the quality of the results obtained from AMT workers, a test trial was held with three variants of the HIS dialogue system [6]. The results obtained with AMT workers were then compared to results from a recent trial of the same systems with users from Cambridge. The assumption was that the ranking of the trialled systems should be the same or the changes in the ranking should not be statistically significant.

The trialled dialogue systems were designed for the Cambridge Restaurant Information domain, in which users can obtain information about venues such as restaurants, pubs, and coffee shops. The users may provide information about what kind of venue they are looking for, for example the area, the price range, or type of food they prefer. Once a suitable venue has been offered by the system, the user may request additional information such as the address, phone number, postcode, or the prices at that venue. An example of a typical dialogue is provided in Table 1.

Table 1: *An example of a dialogue from the Cambridge Restaurant Information domain.*

System	Thank you for calling the Cambridge Information system.
User	Hi, I'm looking for an expensive Chinese restaurant.
System	There is no place to eat serving Chinese food in the expensive price range.
User	I want something cheap then.
System	Yu Garden serves Chinese food. It is in the cheap price range.
User	Ok, give me the address.
System	Yu Garden is located on A/529 Newmarket Road.
User	Thank you. Goodbye.

Typically, evaluation of dialogue systems is based on success rate, representing the proportion of successful dialogues in the trial. In this work, a dialogue is considered to be successful if a dialogue system offers a venue matching user's constraints and it provides all of the information that the users requested.

A success rate can be either subjective or objective. A subjective success rate is computed from the users' feedback information, more specifically from their answers to the question "Did you find all the information you were looking for?", as described in Section 2. However, this metric relies on the ability of the users to accurately rate the performance of the systems. For the purpose of objective evaluation, two objective scoring measures were derived: (1) an objective success rate based on the assigned user goals, (2) an objective success rate based on inferred user goals.

The objective measure based on the assigned goal assumes that the users exactly follow the task description. Given the task description of the goal that was given to the user, it is possible to determine from the system responses whether the offered venue is matching the constraints defined in the goal and whether all required information about the venue has been provided. For this purpose, a simple scoring algorithm was implemented. The problem with this approach however, is that in practice, some users tend to divert from the assigned task description, causing some dialogues to be scored negatively as a result. Typically,

[1] The used PBX was ASTERISK PBX which is available at `http://www.asterisk.org`

the user forgets to specify a constraint mentioned in the scenario, or forgets to ask for some additional information about an offered venue, for example the postcode. In such cases, the system is penalised, even when its responses to the user were appropriate. Hence, the objective success rates based on assigned goals might not always be accurate, especially when there is no direct supervision of the subjects.

In order to address the above problem, a method for computing objective success rates based on inferred user goals was developed. A finely tuned heuristic algorithm is used to infer the goals from the log of the system and user dialogue acts. In this way, the scoring is based on what the user actually asked for. For example, if the user asks about a cheap coffee shop than it becomes part of the constraints of the inferred goal. If a user changes his mind during the dialogue, this is accounted for by identifying a new goal that has to be satisfied.

Ideally, the goal inference algorithm should use the user acts in the true semantic annotations. However, obtaining manual semantic annotations is rather tedious. Instead, AMT workers are used to transcribe the recorded audio [2] and a semantic parser is used to obtain automatic semantic annotations of the resulting audio transcriptions. This approach is justified by the fact that the performance of the used semantic parser on transcribed sentences is comparable to the accuracy of human annotations, based on the inter-annotator agreement scores in a similarly complex tourist information domain [13].

4.1. Results

The evaluated systems were three variants of the HIS dialogue system [6]. The HIS dialogue system is based on a Partially Observable Markov Decision Processes (POMDP) framework which aims to handle inherent uncertainty in spoken dialogue systems in a principled way. The system consists of an automatic speech recogniser (ASR), a semantic parser, a POMDP dialogue manager, a natural language generator (NLG), and a HMM speech synthesiser. As described in [6], the speech recogniser and the semantic parser process the user's speech into an N-best list of dialogue acts. The N-best list of dialogue acts is then used by the dialogue manager to produce a system action. The system action is then passed to the natural language generator that converts it to text, which is finally synthesised.

The aim of the original trial with Cambridge users was to contrast effects of different N-best list sizes and different NLG modules. The first trialled system used a full N-best list with a template-based NLG (NBT). The second system used a 1-best list with a template-based NLG (1BT). The third system used a full N-best list with a reinforcement-learning based NLG (NBRL). The results for both trials are given in Table 2. The systems evaluated with AMT workers have prefix A while the systems evaluated by Cambridge users have prefix C. For each success rate, a 95% confidence interval is given in brackets. The subjective success rates are shown in the column "SubSucc" while the objective success rates based on assigned goals and inferred goals are shown in the columns "ObjSucc AG" and "ObjSucc IG."

Table 2: *Results for the AMT and CAM trials. The success rates are followed by their 95% confidence intervals.*

System	# calls	SubSucc	ObjSucc AG	ObjSucc IG
A/NBT	403	64.3% (4.7)	28.8% (4.4)	44.8% (4.6)
A/1BT	390	67.4% (4.7)	36.7% (4.7)	51.5% (4.7)
A/NBRL	130	56.2% (8.5)	37.7% (8.3)	55.9% (7.7)
C/NBT	199	65.3% (6.6)	46.7% (6.9)	42.0% (6.5)
C/1BT	108	62.0% (9.2)	38.9% (9.2)	42.5% (8.6)
C/NBRL	101	60.4% (9.5)	49.5% (9.7)	55.7% (8.8)

To verify the quality of the subjective rating, inspection of 30 user rated dialogues was carried out. It was observed that all dialogues rated as failed were indeed unsuccessful dialogues; however, many dialogues rated as successful should have been rated as unsuccessful. This suggests that users are overly optimistic in their ratings and rate dialogues as unsuccessful only if something went seriously wrong.

Overall, the objective success rates may seem rather low. This can be partly explained by the high WER (see below) found in the data, and also by the fact that the criterion used for success is very strict. Most reports on system evaluations present success rates based on much softer criteria, such as the dialogue being closed without the user hanging up, or the system merely offering a touristic venue, a bus time, or flight information, without actually checking if that was really what the user was looking for. When comparing "ObjSucc AG" to "ObjSucc IG", one can see that the "ObjSucc AG" values are significantly lower. This is in line with the expectations, because the "ObjSucc AG" ratings are computed under the assumption that users exactly follow the task descriptions, making them overly pessimistic. The objective success rates based on inferred goals ("ObjSucc IG") offers a more accurate insight into the performance of the tested systems, as it correctly rates dialogues in which users diverted from the assigned task description.

The results in terms of the "ObjSucc IG" ratings show that the rankings of the evaluated systems in the AMT and Cambridge trials are consistent. In both trials, the NBRL system is significantly better than the NBT system. On the other hand, neither difference between the NBT and 1BT systems is statistically significant. Two tailed z-tests at the 95% confidence level were used to test the statistical significance of the difference between the success rates.

When dialogue systems are compared on two different user populations and the same speech recogniser is used, the speech recognition performance should be analysed as the ASR performance can differ significantly. In the AMT trial, the users were only native English speakers with North American accent. On the other hand, in the Cambridge trial the users were from a more diverse population. They were either native British speakers or non-native speakers, representing a variety of nationalities. However, it appears that the differences in the user population did not significantly affect the speech recognition performance. The WER for the AMT trial was 53.9% while the WER for the Cambridge trial was 56.6%. This unusually high WER was caused by using mismatched ASR acoustic and language models. In particular, the acoustic model was trained on high quality wide-band audio while in the evaluation the audio signal was a narrow-band telephone speech. Also it was observed that the speaking style of the AMT users was significantly more casual than had previously been encountered with lab-based users.[3]

The impact of the different number of users in the AMT and Cambridge trials was also analysed. As can be seen from Table 3, the number of users in the AMT trial was significantly larger compared to the Cambridge trial while the average number of calls per user was significantly lower. This can be explained by a difference in the recruiting process in which the Cambridge

[2] A separate set of HITs is produced to obtain audio transcriptions of the trial data. To control the quality of the transcriptions, each AMT worker has to transcribe a small amount of speech for which a gold standard annotation is available. If any worker's accuracy is lower than some threshold then its transcriptions are rejected.

[3] When the acoustic model and language model were later retrained using the data collected in this trail, the resulting WER decreased to approximately 20%.

Table 3: *Number of users per trial, average number of calls per user, median of calls per user.*

Trial	# users	average # calls	median # calls
AMT	140	6.5	2
Cambridge	17	24.4	20

users committed themselves to making between 15 to 40 calls in total. On the other hand, the AMT workers can freely decide whether they want to continue contributing to the trial or not. Based on the median of the number of calls per user, about 50% of the AMT users made less than 3 calls. When the data were inspected, it was found that 26 AMT workers out of 140 made more than 10 calls and three users made more than 40 calls. This disparity in the number of calls made by different users is unfortunate as it can bias the subjective rating. What can happen is that a small number of AMT workers like this type of HIT so much that they try to complete almost all available HITs. Consequently, the number of submitted (and potentially unreliable) feedbacks by these few workers can make up a very large portion of all the collected responses, and so the average subjective scores represent mostly these few AMT users. However, this should not play any role in the objective scoring as the scores are determined automatically. In the future, more attention should be paid to encouraging workers to make more calls on average. This could be done, for example, by some form of bonus for reaching a certain number of calls.

In addition, the random call routing algorithm should be improved. It was observed that some workers called some systems significantly more often than others especially when the workers made only a few calls, say, less than six calls. To alleviate the problem, a future routing algorithm should route an incoming call to the system which was called least times by the user. However, such routing is not trivial to set up in the ASTERISK PBX, and therefore, a special routing application would have to be implemented.

5. Conclusion

This paper has discussed a framework for evaluating dialogue systems using Amazon Mechanical Turk for recruiting a large group of users. The framework combines a telephone infrastructure and web interface where the telephone infrastructure is used for connecting users with the dialogue systems and the web interface is used to present tasks and collect user feedback. When the results obtained with the AMT workers were compared to the results collected with Cambridge users, it was found that the ranking in both trials was consistent between these two populations. This suggests that the results obtained from the AMT trial and the more controlled Cambridge trial are indistinguishable.

The use of the AMT workers is substantially more efficient when compared to a controlled test. First, no effort had to be put into recruiting users on AMT. On the other hand, it took several weeks to recruit the users for the Cambridge trial. Second, the evaluation was much cheaper. For example, the AMT workers were paid $0.20 per call while Cambridge users had to be paid $0.80 per call.

The use of AMT workers has the potential to scale to a point at which all differences in performance would be statistically significant. Note that in order to achieve statistical significance, a large number of dialogues has to be collected. For example, if the z-test was used to test a difference of 5% in success rate between two systems where the first system has 45% success rate, one would have to collect about 810 calls per system to identify a statistically significant difference at 95% confidence level. If the tested difference was two times smaller, one would need approximately 4 times more dialogues. While collecting such large numbers of calls in controlled tests has numerous practical problems, it would be achievable with AMT workers by simply extending the trial period.

Overall, it appears that crowdsourcing provides an effective method of rapidly testing spoken dialogue systems at modest cost. Nevertheless, one should keep in mind that the workers are not the "true real users", as the workers are paid. The ultimate goal should be to design and deploy such dialogue systems that they would be used by people genuinely interested in the service and to reach a volume of calls that would allow to experiment with different techniques.

6. Acknowledgment

This research was partly funded by the EU FP7 Programme under grant agreements 216594 (CLASSIC project: www.classic-project.org) and 270435 (JAMES project).

7. References

[1] F. Jurčíček, B. Thomson, and S. Young, "Natural Actor and Belief Critic: Reinforcement algorithm for learning parameters of dialogue systems modelled as POMDPs," *ACM TSLP Special issue: Machine Learning for Robust and Adaptive Spoken Dialogue Systems*, 2011.

[2] M. Gašić, F. Lefèvre, F. Jurčíček, S. Keizer, F. Mairesse, B. Thomson, K. Yu, and S. Young, "Back-off Action Selection in Summary Space-Based POMDP-based Dialogue Systems," in *Proc. ASRU*, 2009.

[3] M. Geist and O. Pietquin, "Managing Uncertainty within the KTD Framework," in *Proceedings of the Workshop on Active Learning and Experimental Design*, Sardinia, Italy, 2011.

[4] J. Schatzmann, "Statistical user modeling for dialogue systems," Ph.D. dissertation, University of Cambridge, 2008.

[5] B. Thomson and S. Young, "Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems." *Computer Speech and Language*, vol. 24, no. 4, pp. 562 – 588, 2010.

[6] S. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu, "The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management," *Computer Speech and Language*, vol. 24, no. 2, pp. 150–174, 2010.

[7] Amazon, "Amazon Mechanical Turk," 2011. [Online]. Available: https://www.mturk.com/mturk/welcome

[8] A. Sorokin and D. Forsyth, "Utility data annotation with Amazon Mechanical Turk," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, 2008.

[9] F. Mairesse, M. Gašić, F. Jurčíček, S. Keizer, B. Thomson, K. Yu, and S. Young, "Phrase-based statistical language generation using graphical models and active learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2010.

[10] M. Marge, S. Banerjee, and A. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *ICASSP*, Dallas, TX, March 2010.

[11] G. Parent and M. Eskenazi, "Toward better crowdsourced transcription: transcription of a year of the LetsGo bus information system data," in *SLT*, Berkeley, CA, December 2010.

[12] Z. Yang, B. Li, Y. Zhu, I. King, G. Levow, and H. Meng, "Collection of user judgements on spoken dialog system with crowdsourcing," in *SLT*, Berkeley, CA, December 2010.

[13] S. Keizer, M. Gašić, F. Jurčíček, F. Mairesse, B. Thomson, K. Yu, and S. Young, "Parameter estimation for agenda-based user simulation," in *SIGDIAL '10: Proc. of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2010.