THE REAL CHALLENGE 2014: PROGRESS AND PROSPECTS

Maxine Eskenazi, Alan W Black, Sungjin Lee
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh PA 1521
{max,awb}@cs.cmu.edu, junion@yahoo-inc.com

David Traum
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA 90094
traum@ict.usc.edu

Abstract

The REAL Challenge took place for the first time in 2014, with a long term goal of creating streams of real data that the research community can use, by fostering the creation of systems that are capable of attracting real users. A novel approach is to have high school and undergraduate students devise the types of applications that would attract many real users and that need spoken interaction. The projects are presented to researchers from the spoken dialog research community and the researchers and students work together to refine and develop the ideas. Eleven projects were presented at the first workshop. Many of them have found mentors to help in the next stages of the projects. The students have also brought out issues in the use of speech for real applications. Those issues involve privacy and significant personalization of the applications. While long-term impact of the challenge remains to be seen, the challenge has already been a success at its immediate aims of bringing new ideas and new researchers into the community, and serves as a model for related outreach efforts.

1 Introduction

This paper describes the REAL Challenge (REAL), including the motivations for the challenge and preliminary results from the first year and prospects for the near future. The ultimate goal of REAL is to bring about a steady stream of data from real users talking to spoken dialogue systems, that can be used for academic research. The immediate goal of the first year of REAL is to bring together high school and undergraduate students, who have fresh ideas of how people will talk to things in the future and what the constraints may be, and seasoned researchers, who know how to create the systems and could work with the students to realize a Wizard of Oz (WOZ) study or a proof-of-concept prototype to try out the idea.

At SLT 2012, panelists stated that there was no publicly available, significant stream of spoken dialog data coming from real users other than the Lets Go data (Raux et al., 2006). Although Lets Go can be used to create statistical models for some information-giving systems, with the wide variety of community needs, it cannot satisfy applications that are not two-way and information giving. In answer to this, REAL was created to spark ideas for speech applications that are needed on a regular basis (fulfilling some real need) by real users. Observing present applications in the commercial and academic community and how little use that they are getting, it was apparent, at least to the authors of this paper, that new minds were needed to devise the right kind of applications. This led the REAL organizers to reach out to high school and undergraduate students.

From announcements in late summer 2013 to the REAL workshop on June 21, 2014, and beyond, this paper traces how REAL was managed, the proposals we received, what happened at the workshop, what follow up we have had and how we measure success.

2 Motivation

Speech and spoken dialog researchers often note that whereas industry has access to a wealth of ecologically valid speech data, the academic community lags far behind. The lag in quantity of data can impede research on system evaluation and in training the machine learning (ML) system components. This chasm can be filled by using recruited subjects. But studies (Ai et al., 2007) have found that the resulting data does not resemble real user data. Paid users follow the rules, but are usu-

Proceedings of the SIGDIAL 2015 Conference, pages 209–216,
Prague, Czech Republic, 2-4 September 2015.

ally just going through the motions. They do not create and follow their own personal goals. Without personal goals, they are not overly concerned about satisfying the problem they were asked to solve. For example, if they asked for a specific flight booking, they won't change their mind opportunistically when a better plan becomes available. Yet this ability to find alternative ways to accomplish a goal is present in real user behavior and poses interesting challenges to spoken dialog systems. Paid users are not bothered by system results that are not what they had requested. They often want to finish the task as rapidly as possible while real users will usually take a little more time to get what they want. And, they don't quit or curse the system at same rate if things are not going well. Thus, at evaluation time, the feedback from the paid user does not reflect the quality of system performance on real users.

Although simulated users can be another data-generating possibility, there are still several good reasons to pursue direct learning from human users. Usually conventional methods to build a user simulator follow a cycle of operations: data collection; annotation; model training and evaluation; and deployment for policy training. The whole development cycle takes quite a long time, and so user behavior can change by the time it is done. Moreover, it is highly likely that the new dialog policy, trained with the user simulator, will cause different user behavior patterns. Additionally, there are always discrepancies between real and simulated user behavior due to many simplifying assumptions in the user model. Thus, training on data from a simulated user can make dialog policies lag behind the ones that are optimal for real users.

While there are significant real user speech databases in industry, that data and the platforms that collected it are not available for release to researchers due to a variety of issues including intellectual property (IP), monetization, customer loyalty and information privacy concerns. So while industry can forge ahead (Halevy et al., 2009), academia is unable to show comparable performances, not due to poor research quality, but simply because of the lack of data.

Thus the community needs new streams of speech data that are available to academia. For this, we must find new applications that real users actually need and will use often. Although assistant applications like Siri, Cortana et al. have sparked the interest and imagination of the public, many people dont use them. The speech and spoken dialog communities must find something else, embracing novel interfaces and applications. And the research community may not be the place where these new ideas should come from. They might better originate with people who are: completely comfortable with the new technologies; not influenced by rigid ideas of what can and can't be done; and not limited by an agenda of what they need to do next. This leads us to believe that the community needs the input of young students who have always lived with the technology and know how they would use it in the future. Biased as the research community is by its knowledge of the science behind the systems, researchers also sometimes overlook some of the basic issues that must be dealt with, going forward. Younger students may also be able to identify the red flags that are keeping speech from being an interface of choice. An important side-benefit of this approach is that this challenge serves as an additional vehicle to bring new practitioners into the spoken dialogue community, by having early access to top researchers and training materials.

3 THE REAL CHALLENGE PROCESS

There is a significant leap from a young student's idea to a data-generating system. The process that REAL put in place breaks this leap into small, achievable steps. First, the organizers of REAL formed an international scientific committee, shown in Table 1. The scientific committee consisted of people who had espoused the spirit of REAL and were willing to work to make it a success.

A webpage (https://dialrc.org/realchallenge/) was created, including a timeline through the June 21st, 2014 workshop, a separate page with details of REAL for students and their teachers, contact information and an application form. Researchers around the world were contacted and asked to recruit students. Six countries began recruitment and four, China, Ireland, Korea and the US, had applicants for the 2014 challenge. One experienced researcher headed each country's efforts and was responsible for recruiting and organizing their students and for sending them to the workshop. The international Coordination Committee members are shown in Table 2.

Table 1: The REAL Scientific Committee

Alan W. Black	Carnegie Mellon University, USA
Maxine Eskenazi	Carnegie Mellon University, USA
Helen Hastie	Heriot-Watt University, Scotland
Gary Geunbae Lee	POSTECH South Korea
Sungjin Lee	Carnegie Mellon University, USA
Satoshi Nakamura	Nara Advanced Institute of Science and Technology, Japan
Elmar Noeth	Fredrich-Alexander University, Erlangen-Nuremberg, Germany
Antoine Raux	Lenovo, USA
David Traum	University of Southern California, USA
Jason Williams	Microsoft Research, USA

The students were encouraged to contact the organizers at any time for more information and/or for guidance in proposal writing. The proposals were submitted by April 1, 2014. They were sent to the scientific committee for review, with two reviewers per proposal. The reviewers, taking into account the age of the participants (from 13 to 23 years old), were asked to evaluate the proposals according the following criteria:

novelty: the proposal could not be exactly the same as an existing application. While existing applications could have the same subject, like cooking, the user interaction and/or function had to be novel.

speech is clearly necessary: the students needed to show that the application solves an issue thanks to its use of speech communication.

practical: this idea could be implemented either with current technology or with clearly definable extensions.

viable: this application is likely to attract real users — while it is not evident at present how best to measure viability, at this stage we could poll potential users. We also believe that the students are well aware of their peers habits and needs.

Table 2: International Coordination Committee

USA	Alan W. Black & Sungjin Lee	Carnegie Mellon University
China	Kai Yu	Shanghai Jiaotong University
Ireland	Emer Gilmartin	Trinity College Dublin
Korea	Gary Geunbae Lee	POSTECH
Scotland	Helen Hastie	Heriot-Watt University
Sweden	Samer Al Moubayed & Jose David Lopes	KTH

The reviews were edited to take into account the age of the students. They included feedback on shaping the ideas (focusing the application, getting rid of spurious activities) and requiring more details about the application (how would someone use it, under what conditions would someone use it. who would want to use it). After the students received their feedback, they were told what they would need to prepare for the workshop: a one-minute presentation of their idea, a poster and a presentation in front of the poster. Some students (China, Ireland) had exams at the time of the workshop and participated via Skype. These students were asked to record their in-front-of-poster presentations in case Skype was not working (in the end it worked very well!). Then the students were given some training:

- a class on speech and spoken dialog for the high school students (undergrads had had this in one of their regular classes);
- a video on how to make a poster – ensuring smooth communication between students and researchers on the day of the workshop: the poster included the goal, a comparison to what presently exists, why their idea was better, and an illustration of the use of their idea showing why it is needed, how someone would use it and how it solves the problem.

The workshop was held on June 21, 2014. After the one-minute presentations, the students stood in front of their posters for 90 minutes. In the following 30 minutes they could go around to see one another's posters. Then groups of researchers and

students formed to discuss the ideas. All of the students found at least two researchers interested in having a discussion with them. Each group created a few slides summarizing their discussion and reported back to everyone. Most of the reports contained ways to focus ideas, to make them doable and most importantly, to define the next steps.

After the workshop, the organizers followed up with the researcher participants to find out their plans going forward. They were also asked whether they would be encouraging high school or undergraduates to join REAL in the next round.

Going forward, the organizers plan to have yearly REAL meetings. While the first workshop saw only proposals, the second and following ones should see both new proposals and results of WOZ studies and proof-of-concept demos from the proposals presented the previous year. This rolling participation enables new students and researchers to join at any time and puts less pressure on past participants – the successful projects will have something to show, but aren't expected to have a fully working system, within just one year. The intended cycle for successful proposals is the following:

1. find technical partners
2. for limitations that must be dealt with: work on why this is a limitation and what the possible fixes are
3. for applications or systems: work on the design then on the prototype or WOZ system
4. conduct a study (testing the prototype or WOZ system)
5. show study results (and possibly demo of system or propose a major design change for speech systems)
6. write a proposal for future funding to continue the work

4 Year One Winning Proposals

The first year of REAL enabled the organizers to assess how well its goals were fulfilled, what outcomes there were and what lessons were learned. The main outcome of REAL can first be shown in the quality of the proposals. Here are summaries of the 11 successful proposals from 2014 (note that all participants from outside the US are undergrads, the US participants are high school students):

Bocal (Jude Rosen, Joe Flot, US)
How can we protect the privacy of the user at the same time as offering a high quality of speech commanding and response? Bone-conducting devices can answer this question by capturing sounds emanating through skulls. The next step includes finding out a specific set of scenarios where the device will be useful and conducting Wizard of Oz experiments to collect data about how users would behave with the device on.

Daily Journaling (Keun Woo Park, Jungkook Park, Korea)
This system will help users record events in their everyday life. Lightweight and multimodal, it uses many sensors to determine what is going on around the user. To interpret what it captures, it asks the user questions. With the information gleaned from the questions, it updates its information about the user.

Fashion Advisor (Jung-eun Kim, Korea)
This advisor knows what clothing a person possesses and carries on a dialog in the morning to help the user choose what to wear. It would have a camera to capture the user and show them how they would look when wearing its suggestions (like a mirror). It also knows what the weather will be and will suggest appropriate clothing. It can also search sources such as Pinterest for clothes to purchase that would work with what the use has and their body type.

Gourmet (Jaichen Shi, China)
The Gourmet helps people choose a restaurant. Many people have dietary restrictions and the Gourmet would suggest restaurants where the user can be assured of finding something they can eat. It also tells the user what other diners have thought of a restaurant and can find specific feedback from diners who were at the restaurant on the present day. When a choice is made, it can call the restaurant for reservations.

Human Chatting System (Yunqi Guo, China)
This is a system that allows people to chat

with it. It is aimed at helping people rehearse discussions they would have with real people, either helping in how to deal with a difficult social situation (asking a girl for a date, for example), or speaking a new language (with a tutor that detects speaking errors and tells the student how to correct them).

Lecture Trainer (Qizhe Xie, China)
This application would listen to a user preparing a presentation and help them out. It could help with word choice, but also with grammar, intonation, and fluency. The user could choose a topic and also listen to recorded speeches from famous people so that the user could imitate the latter.

Mobile Cooking App (BongJin Sohn, Jong-Woo Choi, DongHyun Kim, Korea)
Modern-day appliances continue to evolve based on communication with users to identify and meet their needs. The cooking app will offer a cooking guide in the form of audio or video, voice control for oven and alarm setting, and provide a grocery list, etc. This app traces interaction history and each step of a recipe to make a dialog intelligent and efficient by being context-aware.

Neeloid (Neeloy Chakraborty, US)
The invention connects people with their surroundings. Camera and other sensors can also work together to create an accurate description of the audience's surroundings. It also understands gestures pointing at certain things for inquiry and looks into connected wiki to retrieve relevant information. This invention may give the visually impaired the confidence of knowing what is around them without the use of a white cane, hoople, guide, etc. Another application of this idea is as an educational tool that can be used by a wide variety of people, in particular, children full of curiosity.

Sam the Kitchen Assistant (Enno Hermann, Ireland)
Sam comes to the aid of the cook who has hands occupied and full of food and eyes also busy. Sam can tell a cook what to do next in a recipe, but also has information about how to adapt a recipe to any one of many dietary restrictions. Sam can suggest a recipe, on the way home, given what is in the house and list what needs to be bought.

SmartCID (Zachary McAlexander, David Donehue, US)
Millions of consumers today use smart technology in everyday life, including smartphones, tablets, and desktop computers. However, none of these technologies are truly easy-to-use. The user must always issue some command before the aid begins to operate. SmartCID solves this problem by automatically detecting external activity and instantaneously capturing content. For example, SmartCID can detect things like people posing for a picture, the word cheese said by a group, or a laugh from the user, to prompt the device to begin recording a video or audio file, allowing the user to review the funny moment at a later date.

Smart Watch (So Hyeon Jung, Korea)
This is a patient health care system. Elderly users (some with poor eyesight) can be told when to take their medication. They can also find out when their supply of medication is about to run out and get help ordering more. The system can also guide its users in healthy eating choices for the specific nutrients that the individual needs. And since it can suggest good foods, it can also help with calorie counts.

5 Outcomes from the First Year

The first outcome of the workshop was the proposals for new ideas, described in the previous section. All of them met the desired criteria of novelty, use of speech, with potential for practicality and viability. One of the ideas has already led to a peer-reviewed publication (Jung et al., 2015).

Another outcome of REAL is the set of issues in the ubiquitous use of speech that the students raised. First, the Bocal proposal raised the issue of privacy. Although we generally think that speech should be used in any setting, it is possible that privacy may restrict its frequent use in environments where there are other people in close proximity to the speaker. In this situation, it may indeed be necessary to either whisper or use a bone-conducting microphone. Second, several proposals, such as Mobile Cooking, Lecture Trainer, and Human Chatting System show that the most com-

Table 3: Next steps resulting from the Workshop

type - country	student	type of help	action
Academic -US	any	provide system components	Webinar on virtual human toolkit
Academic - US	KAIST	offer of mentorship	lectures to high school students, participation in next round
Academic - Germany	any present students	could mentor	students in next round
Academic/industry - Germany	-	-	students in next round
Academic - US	2 students from China	offer of internships	none
Industry - US	2 students from US	offer of internships	none
Academic - US	Three projects from US	mentorship	students in next round
Academic - Ireland	Student from Ireland	mentorship	Creating prototype of proposed system also young high school students in next round
Academic - Scotland	-	-	students in next round
Academic - Korea	One student from POSTECH	mentorship	Students will continue to participate next year
Academic - China	-	-	Students will continue to participate next year
Academic - Sweden			students in the next round

pelling applications for a user may not be for general use, but rather suites of applications that are important to individuals. Finally, we see that many of the proposals, without being prompted by organizers or teachers, were in a context of busy hands and eyes.

A third outcome of REAL is what took place the day of the workshop (described in Section 3). Students described their ideas to technologists/researchers. The participants met with students in the afternoon. The breakout reports from these meetings were given by both the researchers and the students. All had made slides and the one common element was the next steps points that all displayed. For many of the projects, the students got help in:

focus: concentrating on just one thing, deciding which thing was worth it, not trying to solve all of the worlds problems.

deciding what to do next: e.g., Is there hardware to concentrate on? Should a scenario be defined? What software is involved? What software modules exist and which ones must be built?

Finally, there is the promise of what is to come. Table 3 shows the post-meeting feedback from participants concerning their plans. For example, one academic participant is proposing internships to two of the students (from two different proposals).

6 Assessing REAL

The first year of REAL can be assessed using several metrics. But before the metrics are used, some perspective is needed. It is very difficult in one year to get a large part of the speech and spoken dialog community actively interested. It is hard to plan the venue of the workshop so that it coincides with a major meeting, while not taking place at the same time. It is also hard to organize students

in many different countries, including the funding for the students. And finding support for REAL is also difficult. Industry is not yet sure what a company can get from this meeting. One measure is researcher participation. There were 21 researchers at the Workshop, 17 people were from academia, and the rest were from industry. Another metric is the depth and breadth of what is being proposed to the students to take their work forward. Yet another metric is whether colleagues plan to get more students involved in the coming year. This is also shown in Table 3. Three colleagues from three different countries proposed either to:

- increase the numbers of their participants next year
- bring in a new high school class
- bring in new undergraduate students

The use of Skype is considered to have been very helpful this year. If a student worked on a proposal during the year and could not, for some reason, attend the workshop (including exams, lack of travel funding, etc), then they were still able to make a presentation and get feedback. Another way to assess REAL is to observe the results of the interaction between the students and the researchers at the workshop breakout sessions. Some examples of the changes in the projects:

- Smart Watch project: there were four functions proposed: calorie-store, alarm, food recommendation, exercise recommendation. Issues that arose: hardware could become multiple devices; calorie store might be difficult for users; it should be multimodal, combining both spoken dialog and images for the users. Plan of action: break project into individual functions; examine existing apps to get a sense of range of interaction; do a WOZ data collection with diet expert function to observe dialogs and users reactions; use WOZ data to finalize design and train ASR/NLU. Subsequent to the workshop, this action plan was followed, and the food and exercise recommendation functions were implemented and tested, resulting in a peer review publication (Jung et al., 2015).
- Bocal project: focusing ideas into a platform for allowing system-user communication when privacy is important. Noting that the key technology will be transferring input from skull microphones to text, the main challenges were gaining an understanding of the differences between speech through skull and standard microphones and understanding how this technology will influence users' behavior. The action items were: choosing application domains that will necessitate privacy, like banking; collecting data with a WOZ setup; analyzing the data to find features for encoding the output of the skull microphone; developing models for transforming the output of the skull microphone to text; developing a spoken dialog system for exhibiting the feasibility of the approach.

Thus, the students got a considerable amount of help in focusing their ideas, in breaking down the steps that they need to take in the upcoming year to find out how feasible their projects are, and in understanding what the hardware and usage issues were. As seen on Table 3, several of the students have found mentors and they will be going forward with their projects.

7 CONCLUSIONS AND FUTURE DIRECTIONS

Although we have had a successful first year we are interested in the long term continued success of this challenge. As it grows in stability year to year it will be easier to get students to be aware of and take part in it. Even since out first year we have seen more standardized SDKs for developing speech based systems on more platforms. Microsoft's Cortana, and Amazon's Echo offer SDKs that we would like to utilize to aid student's proposals and eventual development.

The REAL Challenge is a bold step for researchers. Its stated goal was to find new applications that would create streams of spoken dialog data from real users. It has achieved this goal — students have proposed novel systems that have the potential to be very useful and thus to attract real users. Beyond the stated goals of the Challenge, the students have brought to the forefront issues that must be dealt with:

- The issue of privacy must be addressed. For example, real users would not dictate email or text messages if they feel that their messages are not secure.

- It is probable that the most successful speech applications will not be the general ones (like SIRI and Cortana), but may be the ones that are highly personalized to specific tasks.

Plans going forward concern both this year's projects and those to come in the future. REAL is seen as a regularly occurring event where there are multiple levels of presentation. There will be students who have proposed an idea (like all of the 2014 participants) who are looking for feedback and mentorship. There will be students who proposed their ideas the preceding year and are presenting either WOZ study results or a prototype. And ultimately there will be students (and researchers) who proposed one year, presented preliminary results the next year and are presenting a working system and real user data.

The REAL Challenge continues in its second year with renewed support from the National Science Foundation. Year two proposals for the REAL challenge are under development, with an intended participant workshop in Fall 2015. So far there are six proposals for 2015: three undergraduates and three high school students. The undergrad proposals are all new, while two of the ones from the high school students are updates of last years proposal and one is new. Table 4 shows this years proposals.

Table 4: REAL Challenge 2015 Entries

institution	**level**	**year**	**subject**
Heriot Watt	ugrad	Y1	Table talk - to order food at a restaurant
Heriot Watt	ugrad	Y1	BuddyBot - a companion for sick children in hospital
Pittsburgh Sci	high	Y2	next stage for Smart Content Interaction Device project
Pittsburgh Sci	high	Y1	multilinguistic conference meeting
Pittsburgh Sci	high	Y2	next stage, uses for bone conduction
Sogang U	ugrad	Y1	home chat system to dialog with home devices

Due to the differences in academic schedules around the world, to the success of virtual participation and to cost, the second year will see the students all participate remotely. Experts in the field will be brought in to the Workshop in person. Individual presentations will be given and group breakouts will be organized. Given that last year this Challenge not only proposed novel applications, but also unearthed interesting issues, part of the Workshop will address some of the issues (such as privacy) that are being brought to light.

ACKNOWLEDGEMENTS

The REAL Challenge has been sponsored by NSF grant no. CNS-1405644 and 1406000. The student participants in REAL were mentioned in the description of their projects. We would like to thank Ann Gollapudi, the teacher of the US high-school students. The persons who ran the Challenge in their own countries were:

- Gary Lee, Korea
- Kai Yu, China
- Emer Gilmartin, Ireland.

The authors would like to thank the researchers who took part in the Workshop, many of whom have made plans to follow up on projects and/or future versions of the Challenge.

References

Hua Ai, Antoine Raux, Dan Bohus, Maxine Eskenazi, and Diane Litman. 2007. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *In Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, March.

Sohyeon Jung, SeonghanRyu, Sangdo Han, and Gary Geunbae Lee. 2015. Diettalk: Diet and health assistant based onspoken dialog system. In *Proceedings of Sixth International Workshop on Spoken Dialogue systems (IWSDS 2015)*, January.

Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: one year of let's go! experience. In *INTERSPEECH*. ISCA.