Differences in User Responses to a Wizard-of-Oz versus Automated System

Jesse Thomason
University of Pittsburgh
Pittsburgh, PA 15260, USA
jdt34@pitt.edu

Diane Litman
University of Pittsburgh
Pittsburgh, PA 15260, USA
litman@cs.pitt.edu

Abstract

Wizard-of-Oz experimental setup in a dialogue system is commonly used to gather data for informing an automated version of that system. Previous work has exposed dependencies between user behavior towards systems and user belief about whether the system is automated or human-controlled. This work examines whether user behavior changes when user belief is held constant and the system's operator is varied. We perform a post-hoc experiment using generalizable prosodic and lexical features of user responses to a dialogue system backed with and without a human wizard. Our results suggest that user responses are different when communicating with a wizarded and an automated system, indicating that wizard data may be less reliable for informing automated systems than generally assumed.

1 Introduction

In a Wizard-of-Oz (WOZ) experimental setup, some or all of the automated portions of a dialogue system are replaced with a hidden, human evaluator. This setup is often used to gather data from users who believe they are interacting with an automated system (Wolska et al., 2004; Andrews et al., 2008; Becker et al., 2011). This data can inform a downstream, real automated system. A WOZ experimental protocol calls for holding "all other input and output ... constant so that the only unknown variable is who does the internal processing" (Paek, 2001). Thus, hiding the human wizard's input by layers of system interface can render that system believably automated.

An assumption of this WOZ data-gathering strategy is that user behavior will not vary substantially between the WOZ and automated (AUT) experimental setups. However, it was shown in a dialogue system that training with a small set of data from an automated system gave rise to better performance than training with a large set of data from an analogous wizarded system (Drummond and Litman, 2011). There, it was suggested that differences in system automation may be responsible for the performance gap. It is possible that user responses to these dialogue systems differed substantially.

This paper aims to investigate this possibility by comparing data between a wizarded and automated version of a tutoring dialogue system. We hypothesize that what users say and how they say it will differ when the only change is whether the system's speech recognition and correctness evaluation components are wizarded or automated.

2 Dialogue System

The data for this study is provided by the baseline conditions (one wizarded (WOZ) and one automated (AUT)) of two prior experiments with a spoken tutorial dialogue system. Users of this system were students recruited at our university, and each was a native speaker of American English. Users were novices and were tutored in basic Newtonian physics by the system. Each was engaged by a set of dialogues that illustrated one or more basic physics concepts. Those dialogues included remedial subdialogues that were accessed when the users pro-

Proceedings of NAACL-HLT 2013, pages 796–801,
Atlanta, Georgia, 9–14 June 2013.

Tutor: So what are the forces acting on the packet after it's dropped from the plane?
Student: um gravity then well air resistance is negligible just gravity
Tutor: Fine. So what's the direction of the force of gravity on the packet?
Student: vertically down

Figure 1: Tutor text is shown on a screen and read aloud via text-to-speech. The user responds verbally to the tutor's queries.

vided incorrect or off-topic answers. These prior experiments were examining the effects of system adaptation in response to detected student uncertainty (Forbes-Riley and Litman(a), 2011; Forbes-Riley and Litman(b), 2011). However, in this study we consider only the baseline, non-adaptive conditions of those experiments. Figure 1 shows a sample dialogue excerpt between the student and tutor.

In the baseline conditions of the WOZ and AUT system past experiments, as shown in Figure 2, the setups varied only by the system component responsible for understanding and evaluating a user's verbal response. Each student participated in only one of the two setups, and students were not informed when the system was wizarded. In the WOZ setup a human wizard marked student responses to prompts as correct or incorrect. In the AUT setup, automatic speech recognition was performed on student responses[1], and (in)correctness of answers was determined using natural language understanding models trained from the WOZ experiment's data.

3 Post-Hoc Experiment

Using both lexical and prosodic features, we aimed to determine whether there exist significant differences in users' turn-level responses to the WOZ and AUT systems.

It was suspected that the imperfect accuracy[2] (87%) of the AUT system's evaluations of the (in)correctness of user responses may have led to remedial sub-dialogues being accessed by the AUT system more often, since false-negatives accounted

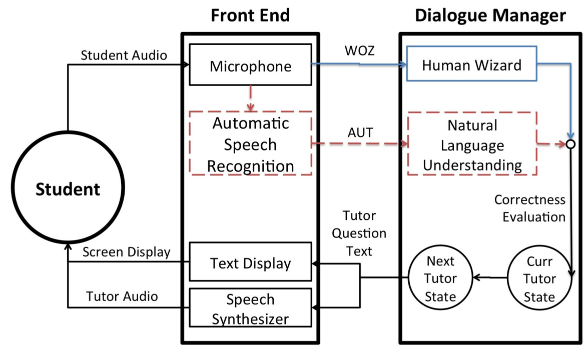

Figure 2: The workflow of the tutoring dialogue system with the WOZ setup component shown in solid, blue and the AUT setup component shown in dashed, red.

System	#Users	#Qu	#Turns
WOZ	21	111	1542
AUT	25	111	2034

Table 1: Counts for users, unique questions, and user turns in each data set.

for 72% of inaccurate evaluations. To correct for this imbalance, rather than comparing user responses to all questions, we compared the features of user responses (turns) to each question individually. We omitted questions which were presented in only one setup[3] as well as turns for which a human transcriber found no user speech. Table 1 gives the numbers of users, number of unique questions asked, and total number of user responses contained in the remaining data and used in our investigations.

For prosodic features, we considered duration, pitch, and energy (RMS), each extracted using openSMILE (Eyben et al., 2010). From pitch and energy, which contain many samples during a single turn, we extracted features for maximum, minimum, mean, and standard deviation of these readings. We also considered speech duration and the length of the pause before speech began. This gave us a total of 10 prosodic features. To account for any differences in recording environment and users' voices, we normalized each prosodic feature by dividing its value on each turn by its value in the first turn of the current problem dialogue for that user. This normal-

[1]The average word-error rate for these AUT responses was 19%.

[2]Agreement of $\kappa = 0.7$ between the system and human.

[3]There were 3 such questions containing 6 user responses; each question was a remedial sub-dialogue accessed in the AUT but not WOZ setup.

ization scheme was chosen for our analysis because it is used in the live system, though we note that alternative methods considering more user responses could be explored in the future.

For lexical features, we used the *Linguistic Inquiry and Word Count* (LIWC). LIWC (Pennebaker et al., 2001), a word-count dictionary, provides features representing the percentage of words in an utterance falling under particular categories. Though still a counting strategy, these categories capture higher-level concepts than would simple unigrams. For example, one category is *Tentative(T)*, which includes words such as "maybe", "perhaps", and "guess". Less abstract categories, such as *Prepositions(P)*, with words such as "to", "with", and "above", are also generated by LIWC. Using these example categories, the utterance "Maybe above" would receive feature vector:

$$\langle 0, \ldots, 0, T = 50, 0, \ldots, 0, P = 50, 0, \ldots, 0 \rangle \quad (1)$$

Human transcriptions of users' speech were made available post-hoc for both system versions. We extracted 69 LIWC categories as lexical features from these human transcriptions of each user turn.

Between the WOZ and AUT setups, we looked for user response feature differences in two ways. First, a Welch's two-tailed t-test was used to compare the distributions of each feature's values between WOZ and AUT user responses per question. We noted the features found to be significantly different. Second, we built classification models to distinguish between user responses per question from the WOZ and AUT experiments. For each question, a J48[4] decision tree model was trained and tested using 10-fold cross validation via the Weka[5] toolkit. Only questions with at least 10 responses between both setups were considered. Each model was compared against majority-class baseline for its respective question by checking for statistically significant differences in the model's accuracy.

[4]We tried Logistic Regression and Support Vector classifiers but these were consistently outperformed by J48.

[5]http://www.cs.waikato.ac.nz/ml/weka

4 Results

4.1 Statistical Comparison of Features

The number of questions for which at least one feature differed statistically significantly was calculated. Since distinct sets of students were involved in the WOZ and AUT setups, it is possible that some of these differences are inherent between the students and not resulting from the presence or absence of a human wizard. To control for this possibility, we assigned students randomly into two new groups (preserving the original class distribution in each new group) and tested for feature differences between these new groups. Table 2 summarizes the differences found by each feature set. We report only questions for which at least one feature differed between WOZ and AUT but not between these two random groups[6]. Table 2 also shows the percentage of turns that those questions comprised in the corpus. Prosodic and lexical features each differ for a substantial portion of the corpus of turns, and when both sets are considered about 67% of the corpus is captured.

Feature set	#Qu	% Corpus by Turns
Prosodic	42	46.22%
Lexical	33	35.46%
Either	61	66.86%

Table 2: Number of questions for which at least one feature from the feature set was found to differ with significance $p < 0.05$ between WOZ and AUT responses and the percentage the corpus represented by those questions, weighted by the speech turns they comprise.

After controlling for possible between-student differences, all 10 prosodic features and 29 out of 69 lexical features differed significantly ($p < 0.05$) for at least one question. Table 3 gives the features which were able to differentiate at least 10% of the corpus by turns.

These t-tests show there exist features which differ for a substantial number of questions between the two experimental setups. Examination of Table

[6]We repeated this random split procedure 10 times and found, after omitting features found significant in any of the 10 splits, that 58.08% of the corpus was still captured. Less than 2% of the turns belonged to questions with at least one feature different through all 10 splits.

Feature	% CbT	#Qu	#W>A
Duration	22.15%	19	1
RMS Min	16.86%	15	14
Dictionary Words	15.13%	13	11
pronoun	12.56%	10	10
social	11.35%	9	8
funct	10.99%	9	9
Six Letter Words	10.91%	9	0

Table 3: Features shown to differ with significance $p < 0.05$ between WOZ and AUT responses in questions comprising at least 10% of the corpus by turns (CbT). The numbers of questions these turns comprised and of questions with greater (W)OZ than (A)UT mean are also given.

Tutor: So how do these two forces' directions compare?
Top two most common responses:
WOZ(9),AUT(2): they are opposite
WOZ(3),AUT(8): opposite
Longest responses per tutor setup:
WOZ Student: the relationship between the two forces' directions are towards each other since the sun is pulling the gravitational force of the earth
AUT Student: they are opposite directions

Figure 3: The tutor question and select user responses to a question for which the *Dictionary Words* feature was greater for WOZ responses.

3 in addition suggests that users used more words with the wizarded system. For example, the fourth row shows that all of the questions showing differences for the LIWC category *pronoun* (the words "they", "he", and "it" are popular in this corpus) exposed higher percentage of pronouns in the WOZ utterances. The usual dominance of the third row, *Dictionary Words*, by the WOZ utterances also reflects this trend. Figure 3 gives common and characteristic student responses for each setup on a question for which *Dictionary Words* differed significantly. We next applied machine learning to classify the experiment-of-origin of responses based on these features.

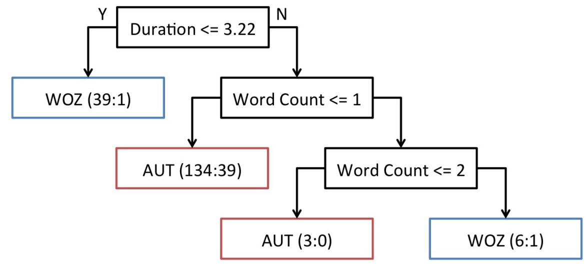

Figure 4: The J48 tree for the question "Would you like to do another problem?". Classification nodes are marked in blue and red for WOZ and AUT, respectively, and specify (#Instances:#Incorrect).

4.2 Response Classification Experiments

After removing questions with less than 10 responses between the two setups, there remained 97 questions totaling 2980 turns. Of the J48 models built and tested on each question, 21 of 97 outperformed the majority-class baseline accuracies for those questions with significance $p < 0.05$. These 21 questions represented 32.79% of the corpus by turns. We present in detail the two of these 21 questions with the most turns.

The question "Would you like to do another problem?" represented 6.11% of the corpus by turns and the J48 model built for it, shown in Figure 4, outperformed the baseline accuracy with $p < 0.001$. While the *Duration* feature was the root node, a bigger decision was made by *Word Count* ≤ 1, for which most responses were from AUT data. This result is consistent with literature (Schechtman and Horowitz, 2003; Rosé and Torrey, 2005) that suggests that users interacting with automated systems will be more curt.

The question "Now let's find the forces exerted on the car in the vertical direction during the collision. First, what vertical force is always exerted on an object near the surface of the earth?" represented 1.54% of the corpus by turns and the J48 model built for it, shown in Figure 5, outperformed the baseline accuracy with $p < 0.01$. Again, *Duration* emerged as the tree root, but here the biggest decision fell to *RMS mean*. Student responses approximately louder than the initial response to the tutor in this question dialogue were marked, almost entirely accurately, as AUT.

Since both trees were rooted at *Duration*, we sam-

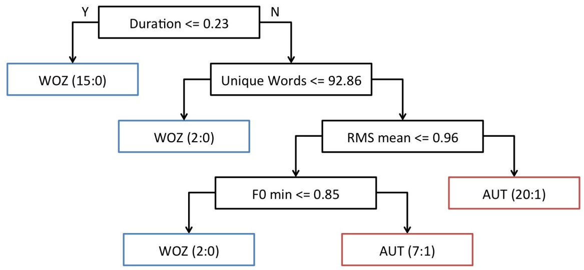

Figure 5: The J48 tree for the question "Now let's find the forces exerted on the car in the vertical direction during the collision. First, what vertical force is always exerted on an object near the surface of the earth?". Classification nodes are marked in blue and red for WOZ and AUT, respectively, and specify (#Instances:#Incorrect).

pled common responses from each experiment for both problems. We noticed that hyper-articulation (speaking slowly, loudly, and enunciating each syllable) was more common in the AUT responses. For example, one user answering "Would you like to do another problem?" took almost 4 seconds to clearly and slowly pronounce the word "yes". We suspect that these hyper-articulations may have contributed to the classifiers' ability to detect WOZ responses based on their brevity.

The performance of the per-question J48 models shows, for a non-trivial portion of the turns, that the experiment-of-origin can be classified based on generalizable prosodic and lexical features alone. The two trees discussed above demonstrate the simplicity of the models needed to perform this separation.

5 Discussion and Future Work

We demonstrate that there exist significant differences between user responses to a wizarded and an automatic dialogue system's questions, even when the contribution of the wizard is as atomic as speech recognition and correctness evaluation. Our generalizable features are derived exclusively from the recordings of the users' responses and human transcriptions of their speech.

Because the role of the wizard in the WOZ setup was limited to evaluating users' spoken response to a prompt, our results suggest that user speech changes as a result of user confidence in the system's accuracy. For example, Figure 3 demonstrates that users in the WOZ setup used complete sentences and gave long responses, where AUT users, possibly anticipating system error, used shorter (sometimes one word) responses. This relationship between user confidence and user speech may be analogous to observed differences like users' longer speech and typed responses to systems when told those systems are human-operated (Schechtman and Horowitz, 2003; Rosé and Torrey, 2005). Our results suggest ways in which raw wizarded data may fall short of ideal for training an automated system.

Having established that differences exist, our future work will focus on deeper exploration of the nature of these differences in users' responses. We suspect users become less confident in the automated system over time, so one direction of study will be to measure how the observed differences change over the course of the dialogue. We expect that they are minimal early on and become more pronounced in the automated setup as users' confidence is shaken. Additionally, some technical aspects of our methodology may impact these and future results: using different methods of normalization for user speech values than the one from this paper may affect visibility of observed differences between the setups.

Future work may also attempt to address these differences directly. Intentional wizard error could be introduced to frustrate the user into responding as she would to a less accurate system, analogous to intentional errors produced in user simulation in spoken dialogue systems (Lee and Eskenazi, 2012). This strategy would be further informed by studies of the relationship between the system's evaluation accuracy and student responses' deviation from wizarded responses. Alternatively, post-hoc domain adaptation could be used to adjust the WOZ data. Generalizable statistical classification domain adaptation (Daumé and Marcu, 2006) and adaptation demonstrated to work well in NLP-specific domains (Jiang and Zhai, 2007) both have the potential to adjust WOZ data to better match that seen by automated systems.

Acknowledgments

This work is funded by NSF award 0914615. We thank Scott Silliman for his support and Pamela Jordan, Wenting Xiong, and the anonymous reviewers for their helpful suggestions and commentary.

References

Pierre Andrews, Suresh Manandhar, and Marco De Boni. 2008. Argumentative Human Computer Dialogue for Automated Persuasion. *Proceedings of the 9th SIGDIAL Workshop on Discourse and Dialogue*, pages 138-147, Columbus, June 2008. Association for Computational Linguistics.

Lee Becker, Wayne Ward, Sarel van Vuuren, Martha Palmer. 2011. DISCUSS: A dialogue move taxonomy layered over semantic representations. *International Workshop on Computational Semantics (IWCS)*, Main Conference. Association for Computational Linguistics.

Hal Daumé III and Daniel Marcu. 2006. Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research*, 26:101-126. AI Access Foundation.

Joanna Drummond and Diane Litman. 2011. Examining the Impacts of Dialogue Content and System Automation on Affect Models in a Spoken Tutorial Dialogue System. *Proceedings of the SIGDIAL 2011 Conference*, Portland, Oregon, June. Association for Computational Linguistics.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. *MM '10 Proceedings of the International Conference on Multimedia*, 1459-1462.

Kate Forbes-Riley and Diane Litman. 2011. Designing and Evaluating a Wizarded Uncertainty-Adaptive Spoken Dialogue Tutoring System. *Computer Speech and Language*, 25(1): 105-126.

Kate Forbes-Riley and Diane Litman. 2011. Benefits and Challenges of Real-Time Uncertainty Detection and Adaptation in a Spoken Dialogue Computer Tutor. *Speech Communication 2011*, 53(9-10): 1115-1136.

Jing Jiang and ChengXiang Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264-271, Prague, Czech Republic, June 2007.

Sungjin Lee and Maxine Eskenazi. 2012. An Unsupervised Approach to User Simulation: toward Self-Improving Dialog Systems. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 50-59, Seoul, South Korea, 5-6 July 2012. Association for Computational Linguistics.

Tim Paek. 2001. Empirical Methods for Evaluating Dialog Systems. *SIGDIAL '01 Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 16:1-9.

James Pennebaker, Martha Francis, and Roger Booth. 2001. Linguistic Inquiry and Word Count (LIWC): LIWC2001. Lawrence Erlbaum Associates, Mahwah, NJ.

Carolyn P. Rosé and Cristen Torrey. 2005. Interactivity and Expectation: Eliciting Learning Oriented Behavior with Tutorial Dialogue Systems. *Human-Computer Interaction-INTERACT 2005*, 323-336. Springer Berlin/Heidelberg.

Nicole Schechtman and Leonard M. Horowitz. 2003. Media Inequality in Conversation: How People Behave Differently When Interacting with Computers and People. *CHI '03 Proceedings of the SIGCHI conference on Human factors in computing systems*, 281-288.

Magdalena Wolska, Ivana Kruijff-Korbayová, Helmut Horacek. 2004. Lexical-Semantic Interpretation of Language Input in Mathematical Dialogs. *Proceedings of the ACL 2nd Workshop on Text Meaning and Interpretation*, 81-88.