Examining the Impacts of Dialogue Content and System Automation on Affect Models in a Spoken Tutorial Dialogue System

Joanna Drummond
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
jmd73@cs.pitt.edu

Diane Litman
Department of Computer Science
Learning Research & Development Ctr.
University of Pittsburgh
Pittsburgh, PA 15260
litman@cs.pitt.edu

Abstract

Many dialogue system developers use data gathered from previous versions of the dialogue system to build models which enable the system to detect and respond to users' affect. Previous work in the dialogue systems community for domain adaptation has shown that large differences between versions of dialogue systems affect performance of ported models. Thus, we wish to investigate how more minor differences, like small dialogue content changes and switching from a wizarded system to a fully automated system, influence the performance of our affect detection models. We perform a post-hoc experiment where we use various data sets to train multiple models, and compare against a test set from the most recent version of our dialogue system. Analyzing these results strongly suggests that these differences do impact these models' performance.

1 Introduction

Many dialogue system developers use data gathered from previous versions of a system to train models for analyzing users' interactions with later versions of the system in new ways, e.g. detecting users' affect enables the system to respond more appropriately. However, this training data does not always accurately reflect the current version of the system. In particular, differences in the levels of automation and the presentation of dialogue content commonly vary between versions. For example, Raux et al (2006) changed dialogue strategies for their Let's Go bus information system after real-world testing.

Previous work in dialogue systems with regards to analyzing the impact of using differing training data has primarily been in the domain adaptation field, and has focused on two areas. First, previous work empirically analyzed the *need* for domain adaptation, i.e. methods for porting existing classifiers to unrelated domains. For example, Webb and Liu (2008) developed a cue-phrase-based dialogue act classifier using the Switchboard corpus, and tested on call center data. While this performed reasonably, training on the call center corpus and testing on Switchboard performed poorly.

The second research direction involves proposing *methods* for domain adaptation. Margolis et al. (2010) observed similar poor performance when porting their dialogue act classifier between three corpora: Switchboard, the Meeting Recorder Dialog Act corpus, and a machine-translated version of the Spanish Callhome corpus. They report promising results through varying their feature set. Blitzer et al. (2007) also observed poor performance and the *need* for adaptation when porting product review sentiment classifiers. They used four review corpora from Amazon (books, DVDs, electronics, and small appliances), which yielded 12 cross-domain training/testing pairs. Their algorithmic adaptation methods showed promising results.

Our work is in the first direction, as we also empirically analyze the impact of differences in training and testing corpora to demonstrate the *need* for adaptation methods. However, our work differs from domain adaptation, as the corpora in this experiment all come from one intelligent spoken physics tutor. Instead, we analyze differences resulting from vary-

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 312–318,
Portland, Oregon, June 17-18, 2011.

ing levels of **automation** and small changes in dialogue **content** between versions of our system.

With respect to analyzing **automation**, we empirically compare the impact of differences in training on data from wizarded (WOZ) versus fully automated systems. Though many systems use data from a WOZ version of the system to train models which are then used in fully automated versions of the system, the effectiveness of this method of dialogue system development has not been tested. We hypothesize that models built with automated data will outperform models built with wizarded data.

Additionally, minor dialogue **content** changes typically exist between versions of systems. While large changes, like changing domains, have been shown to affect model performance, no work has investigated the impact of these more minute changes. We hypothesize that these differences in dialogue **content** presentation will also affect the models.

Finally, the amount of training data is a well known factor which affects performance of models built using supervised machine learning. We hypothesize that combining some, but not all, types of training corpora will improve the performance of the trained models, e.g. adding automated data to WOZ data will improve performance, as this provides fully automated examples. We hypothesize only providing more WOZ data will not be as useful.

2 Data

The data used for this work comes from two prior experiments using ITSPOKE, a spoken tutorial dialogue system, which tutors physics novices. Table 1 describes all data used, displaying the number of users per data set, the number of dialogues between the system and each user, the total number of user turns per corpus, and the percentage of turns labeled uncertain. See Appendix A for more information.

The first experiment, in 2007, compared two dialogue-based strategies for remediating user uncertainty over and above correctness (Forbes-Riley and Litman, 2011b). The goal of this work was to not only test the hypothesis that this uncertainty remediation would improve users' learning, but to investigate what types of dialogue remediation would improve users' learning the most. Since this experiment, WOZ-07, was designed to be a gold-standard case of uncertainty remediation, all natural language understanding and uncertainty annotation was performed by a human wizard, in real time (WOZ). All annotations were made at the turn-level.

For WOZ-07, users' dialogue interactions with the system would change based on which remediation strategy they were assigned to. There were two different dialogue-based remediation strategies. In addition to varying the strategies, the two control conditions in this experiment also varied when the remediation strategy was applied.

The *simple* remediation dialogue strategy provided additional information about the physics concept the user was struggling with, or asked them further questions about the concept. Both control conditions used the *simple* remediation strategy; one only applied the strategy when the user was incorrect, the other applied it if the user was incorrect and randomly when the user was correct. The *simple* remediation experimental condition applied the remediation when the user was incorrect, or correct but uncertain about their answer. The fourth condition in WOZ-07 used the second dialogue strategy, *complex* remediation. This strategy changed the way the remediation was presented, depending on a combination of the user's correctness and certainty in their answer. Only users in the *simple* remediation experimental condition learned more than users in other conditions. Figure 1 shows an example of *simple* remediation; the tutor acknowledges that the user is incorrect, saying “Well...”, and then explains the concept the previous question tested. Appendix B compares *simple* and *complex* remediation strategies.

Another experiment was performed in 2008, where users interacted with either a fully automated (ASR) version of ITSPOKE or a wizarded version. The goal of this experiment was to see if the learning gains found in the 2007 experiment would hold in the ASR version of the system. To mimic the WOZ-07 experiment, the wizarded version (WOZ-08) only used the *simple* remediation experimental condition found in WOZ-07, while the ASR version contained the *simple* remediation experimental condition and both *simple* remediation control conditions. The *complex* remediation strategy was not included due to its poor performance in WOZ-07. Thus, WOZ-08 and ASR-08 used identical dialogue strategies, with minor differences in where the reme-

TUTOR$_{p5}$: *(Response to an incorrect answer)* Well... We just discussed that by Newton's Third law, when two objects collide, the forces they exert on each other are equal in magnitude and opposite in direction. This is true regardless of the objects' differing masses. So the first question's answer is that the impact forces on the truck and the car will have the same magnitude but opposite direction. Now, the second question asks about the vehicles' change in motion. We can use Newton's Second law to answer this. What does this law say?

TUTOR$_{p6}$: *(Response to a correct, certain answer)* Fine. So the first question's answer is that the impact forces on the bus and the motorbike will have the same magnitude, but opposite direction. Now, the second question asks about the vehicles' change in motion. We can use Newton's second law to answer this. What does this law say?

Figure 1: Corpus Excerpt: Remediation in Dialogue 5, and No Remediation in Isomorphic Dialogue 6

diation would be applied. For the ASR conditions, all models were trained on WOZ-07 data; users were randomly assigned to the WOZ-08 or ASR-08 condition as they participated.

In addition to eliminating the *complex* remediation condition, a sixth dialogue, completely isomorphic to the fifth dialogue, was added to all conditions. See Appendix B dialogue examples, highlighting their content differences. Figure 1 displays two ASR-08 tutor turns with the same user. These turns are from the fifth problem, and the isomorphic sixth problem. Note that two things change between these two answers. First, the system responds to the user's incorrectness in the first example. Had the user been correct and uncertain, this is also the dialogue s/he would have seen. Second, notice that problem five discusses a car, while problem six discusses a motorcycle. To create a completely isomorphic problem, the scenario for the dialogue was changed from a car to a motorcycle.

For both the 2007 and 2008 corpora, all gold-standard uncertainty annotations were performed by a trained human annotator. Development and previous testing of the annotation scheme between this annotator and another trained annotator resulted in *kappa* $= 0.62$. All wizarded conditions were annotated in real-time; all ASR conditions were annotated in a post-hoc manner.

Data Set	**#Usr**	**#Dia**	**#Turn**	**%Unc**
WOZ-07	81	5	6561	22.73
WOZ-08	19	6	1812	21.85
ASR-08	72	6	7216	20.55
ASR-08-Train	19	6	1911	21.51
ASR-08-Test	53	6	5305	20.21

Table 1: Description of data sets

In sum, the main differences between the two systems' data are differences in **automation** (i.e. WOZ and ASR) and **content** (i.e. presentation of content, reflected by differing dialogue strategies, and number of physics dialogues).

3 Post-Hoc Experiment

In this post-hoc analysis, we will analyze the impact of **content** differences by comparing the performance of models built with WOZ-07 and WOZ-08, and **automation** differences by comparing models built with WOZ-08 and ASR-08 data. Instead of the original study design, where WOZ-08 and ASR-08 subjects were run in parallel, we could have gathered the WOZ data first, and used the WOZ data and the first few ASR users for system evaluation and development purposes. Thus, for the post-hoc analysis, we mimic this by using WOZ-08 as a training set, and splitting ASR-08 into two data sets–ASR-08-Train (the first few users), and ASR-08-Test. (Please see the last two rows of Table 1.) We held out the first 19 users for ASR-08-Train, since this approximates the amount of data used to train the model built with WOZ-08. For our post-hoc study, the remaining 53 ASR users were used as a test set for all training sets, to mimic an authentic development lifestyle for a dialogue system. Additionally, this guaranteed that no users appear in both the training and testing set given any training set.

As all uncertainty remediation happens at the turn-level, we classified uncertainty at the turn-level, and compared these automated results with the gold-standard annotations. We used all the features that were designed for the original model. Since previous experiments with our data showed little variance between different machine learning algorithms, we chose a J48 decision tree, implemented by WEKA,[1]

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

for all experiments due to its easy readability. Since our class distribution is skewed (see Table 1), we also used a cost matrix which heavily penalizes classifying an uncertain instance as certain.

We use simple lexical, prosodic and system-specific features described in (Forbes-Riley and Litman, 2011a) to build our models. These features were kept constant through all experiments, so the results could be directly comparable. For all lexical features for all data sets, ASR text was used.[2] For all WOZ conditions, we gathered ASR text post-hoc.

We trained models on individual training sets, to inspect the impact of **content** and **automation** differences. We then trained new models on combinations of these original training sets, to investigate possible interactions. To allow for direct comparison, we used ASR-08-Test to evaluate all models.

Since detecting uncertainty is related to detecting affective user states, we use the evaluation measures Unweighted Average (UA) Recall and UA Precision, presented in (Schuller et al., 2009).We also use UA F-measure. Note that because only one hold-out evaluation set was used, rather than using multiple sets for cross-fold validation, we do not test for statistical significance between models' results.

4 Results

The first three rows of Table 2 present the results of training a model on each possible training set individually. Note that the number of instances per training set varies. WOZ-07 simply has more users in the training set than WOZ-08 or ASR-08-Train. While WOZ-08 and ASR-08-Train have the same number of users, the number of turns slightly varies, since dialogues vary depending on users' answers.

When comparing WOZ-08 to WOZ-07, first notice that WOZ-08 outperforms WOZ-07 with a much smaller amount of data. Both are wizarded versions, but **content** differences exist between these experiments; WOZ-08 only used the *simple* remediation strategy, and added a dialogue.

When comparing ASR-08-Train to the other two individual training sets, note that it best approximates the test set. This training condition outperforms all others, while using less data than WOZ-07. While WOZ-08 and ASR-08 have the same **content**, the system changes from wizarded to automated language recognition. This allows us to directly compare how differences due to **automation** (e.g. errors in detecting correct answers) can affect performance of the models. Note that even though we used ASR transcriptions of WOZ-08 turns, the effects of ASR errors on later utterances are only propagated in ASR-08-Train. As ASR-08-Train noticeably outperforms WOZ-08, with approximately the same amount of training data, we conclude that using automated data for training better prepares the model for the data it will be classifying.

As we also wish to investigate how incorporating more diverse training data would alter the performance of the model, we combined ASR-08-Train and WOZ-08 with the WOZ-07 training set, shown in Table 2. We combined these sets practically, as we wish to test how our model could have performed if we had used our first few 2008 users to train the model in the actual 2008 experiment.

First, note that all combination training sets outperform individual training sets. As ASR-08-Train outperformed WOZ-08 for individual training sets, it is not surprising that WOZ-07+ASR-08-Train outperforms WOZ-07+WOZ-08.

However, we could have used WOZ-07 for feature development only, and trained on WOZ-08 + ASR-08-Train. Since the training and testing sets contain identical **content**, it is unsurprising that the precision for this classifier is high. This classifier does not perform as well with respect to recall, perhaps since its training data is not as varied. Also note, while this model trained on few data points, we used additional data for feature development purposes.

Combining all three possible training sets does not outperform WOZ-07+ASR-08-Train; it performs equivalently, and uses much more data. We hypothesize that, since WOZ-07 constitutes the majority of the training set, the benefit of including WOZ-08 may be mitigated. Downsampling WOZ-07 could test this hypothesis. Alternatively, the benefit of combining WOZ-07+ASR-08-Train could be that we provide many varied examples in this combined training set. Since WOZ-07 already accounts for differences in both **content** and **automation**, WOZ-08 doesn't introduce novel examples for the classifier, and adding it may not be beneficial.

[2] We used ASR instead of manual transcriptions, to better approximate automated data.

Training Set	***n***	**UA Rec.**	**UA Prec.**	**UA F1**
WOZ-07	6561	54.6%	53.0%	53.79%
WOZ-08	1812	58.0%	55.4%	56.67%
ASR-08-Train	1911	60.5%	57.2%	58.80%
WOZ-07 + WOZ-08	8373	66.1%	61.0%	63.45%
WOZ-07 + ASR-08-Train	8472	**68.3%**	63.5%	65.81%
WOZ-08 + ASR-08-Train	3723	64.0%	**73.4%**	**68.38%**
WOZ-07 + WOZ-08 + ASR-08-Train	10284	**68.3%**	63.6%	65.86%

Table 2: Results; Testing on ASR-08-Test ($n = 5305$). Bold denotes best performance per metric.

In sum, different training set combinations provide different benefits. With respect to UA F1 and UA Precision, WOZ-08 + ASR-08-Train outperforms all other training sets. Using only 3723 turns to train the model, this configuration uses the least amount of training data. However, this requires previously collected data, such as WOZ-07, for feature development purposes. Alternatively, WOZ-07 + ASR-08-Train performs better than WOZ-08 + ASR-08-Train with respect to UA Recall, and does not require a separate feature development set. Thus, the 'best' training set would depend on both the experimental design, and the preferred metric.

5 Discussion and Future Work

In this paper, we provided evidence that the degree of **automation** of a system used to collect training data can impact the performance of a model when used in a fully automated system. Since one common technique of building fully automated dialogue systems uses a semi-automated wizarded version, this result suggests incorporating a small amount of automated data could greatly improve performance of the models. Our results also suggest that the type of data is more important than the quantity when building these models, since well-performing models were built with small amounts of data. We also investigated the impact of building models trained with different dialogue **content**, another common method of developing dialogue systems. As the WOZ-08 model outperforms the WOZ-07 model, it appears that this has a noticeable impact.

However, the WOZ-08 and WOZ-07 experiments may not have had identical user population, due to the timing differences between studies. We wish to perform further post hoc-experiments to analyze the impact of population differences in our data. To do so, we will eliminate all dialogue strategy differences between WOZ-07 and WOZ-08. To further support our results regarding **content** differences, we wish to split WOZ-08 into two training sets, one including the sixth problem, and one excluding it. After controlling for differences in quantity of data, we will analyze the resulting models. To further strength our results regarding **automation** differences, we will eliminate all differences in when the remediation dialogue strategy was applied between the WOZ-08 and ASR-08-Test corpus, and try to replicate the results found in this paper.

As our results suggest the *need* for applying domain adaptation methods to improve models' performance when there are differences in **automation** and **content**, future work could investigate applying already existing *methods* for domain adaptation, and developing new ones for this problem. In particular, the results we presented suggest a method for building a dialogue system that could mitigate the effects of changes in automation and content. A small wizarded condition, with changes in dialogue content, could be used for feature development. This data, or data from another small wizarded condition, could then be used to train a preliminary model. This preliminary model could be tested with a small number of users using an automated version. Then, the data from the preliminary conditions could be used to build the final model, which would be used for the current, fully automated version of the system.

Acknowledgments

We thank Michal Valko, Michael Lipschultz, Wenting Xiong, and the ITSPOKE group for helpful comments and suggestions, and an REU supplement to NSF Grant #0631930 for funding this work.

References

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 440.

K. Forbes-Riley and D. Litman. 2011a. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*.

K. Forbes-Riley and D. Litman. 2011b. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language*, 25(1):105–126.

A. Margolis, K. Livescu, and M. Ostendorf. 2010. Domain adaptation with unlabeled data for dialog act tagging. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 45–52. Association for Computational Linguistics.

A. Raux, D. Bohus, B. Langner, A.W. Black, and M. Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of Lets Go! experience. In *Proc. Interspeech*, pages 65–68. Citeseer.

B. Schuller, S. Steidl, and A. Batliner. 2009. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

N. Webb and T. Liu. 2008. Investigating the portability of corpus-derived cue phrases for dialogue act classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 977–984. Association for Computational Linguistics.

Data Set	Dialogue #5		Dialogue #6	
	Avg. #Turn	Avg. %Unc	Avg. #Turn	Avg. %Unc
WOZ-07	15.21	26.87	N/A	N/A
WOZ-08	14.37	28.21	13.42	9.02
ASR-08	16.26	19.90	16.28	9.73
ASR-08 -Train	16.05	19.67	16.00	11.84
ASR-08 Test	16.38	19.98	16.38	8.99

Table 3: Dialogue-level description of corpora

Appendix A Differences between corpora

We note that one possible difficulty in using dialogues with differing **content** could be differing levels of users' uncertainty between those dialogues. Thus, Table 3 depicts the average percent of uncertain turns over all users, per dialogue. Only Dialogues #5 and #6 are presented; average number of turns and average percent uncertainty do change between all problems. While the average number of turns are similar, the percentage of uncertain turns drops greatly between the two isomorphic problems.

Appendix B Dialogue Examples

We also present dialogue examples from our corpora to illustrate various **content** differences between versions of ITSPOKE. The changes between Dialogues #5 and #6 are as follows: every instance of "car" in Dialogue #5 with "motorbike," and every instance of "truck" is replaced with "PAT bus." Since users discuss Problem #6 with ITSPOKE immediately after discussing Problem #5, the drop in percentage of user uncertainty between Dialogue #5 and Dialogue #6 show in Table 3 might be explained by this. However, all of the domain-specific words and all of the knowledge concepts will remain the same between these two problems.

Figures 1 (in the main paper) and 2 give examples of the two types of remediation found in the 2007 corpus. ITSPOKE can change its dialogue based on not only how it presents the content, but also what feedback it gives. *Complex* remediation changes both of these attributes. Table 4 displays possible feedbacks given to the user, depending on their levels of correctness and certainty. In Figure 2, the segment of the tutor's turn is labeled after that segment is completed (e.g. the Feedback is "That's not correct... resolve it."). The type of remediation can also change. While Figure 1 depicts the normal remediation path as if the user had answered incorrectly or correct but uncertain, *complex* remediation, shown in Figure 2, first gives the user a short version of the answer that they should have given, before moving down the normal remediation path.

Tutor: What's the overall net force on the truck equal to?
User: Zero?? *[Incorrect & Uncertain]*
Tutor: That's not correct, but don't worry. You seem to be aware of your mistake. Let's resolve it. *[FEEDBACK]* The net force on the truck is equal to the impact force on it. Let's walk through this answer step by step. *[NEW SHORT ANSWER]* We can derive the net force on the truck by summing the individual forces on it, just like we did for the car. First, what horizontal force is exerted on the truck during the collision? *[EXISTING SUBDIALOGUE]*

Figure 2: Example of *Complex* uncertainty remediation.

User Answer	Examples of Feedback Phrases	
	Simple	*Complex*
Correct & Certain	That's right.	That's right.
Correct & Uncertain	That's right.	That's right, but you don't sound very certain, so let's recap.
Incorrect & Uncertain	Well...	Good try, but that's not right. It sounds like you knew there might be an error in your answer. Let's fix it.
Incorrect & Certain	Well...	I'm sorry, but there's a mistake in your answer that we need to work out.

Table 4: Example Feedback Phrases used in *Simple* and *Complex* Remediation