Simulating the Behaviour of Older versus Younger Users when Interacting with Spoken Dialogue Systems

Kallirroi Georgila, Maria Wolters and Johanna D. Moore
Human Communication Research Centre
University of Edinburgh
kgeorgil|mwolters|jmoore@inf.ed.ac.uk

Abstract

In this paper we build user simulations of older and younger adults using a corpus of interactions with a Wizard-of-Oz appointment scheduling system. We measure the quality of these models with standard metrics proposed in the literature. Our results agree with predictions based on statistical analysis of the corpus and previous findings about the diversity of older people's behaviour. Furthermore, our results show that these metrics can be a good predictor of the behaviour of different types of users, which provides evidence for the validity of current user simulation evaluation metrics.

1 Introduction

Using machine learning to induce dialogue management policies requires large amounts of training data, and thus it is typically not feasible to build such models solely with data from real users. Instead, data from real users is used to build simulated users (SUs), who then interact with the system as often as needed. In order to learn good policies, the behaviour of the SUs needs to cover the range of variation seen in real users (Schatzmann et al., 2005; Georgila et al., 2006). Furthermore, SUs are critical for evaluating candidate dialogue policies.

To date, several techniques for building SUs have been investigated and metrics for evaluating their quality have been proposed (Schatzmann et al., 2005; Georgila et al., 2006). However, to our knowledge, no one has tried to build user simulations for different populations of real users and measure whether results from evaluating the quality of those simulations agree with what is known about those particular types of real users, extracted from other studies of those populations. This is presumably due to the lack of corpora for different types of users.

In this paper we focus on the behaviour of older vs. younger adults. Most of the work to date on dialogue systems focuses on young users. However, as average life expectancy increases, it becomes increasingly important to design dialogue systems in such a way that they can accommodate older people's behaviour. Older people are a user group with distinct needs and abilities (Czaja and Lee, 2007) that present challenges for user modelling. To our knowledge no one so far has built statistical user simulation models for older people. The only statistical spoken dialogue system for older people we are aware of is Nursebot, an early application of statistical methods (POMDPs) within the context of a medication reminder system (Roy et al., 2000).

In this study, we build SUs for both younger and older adults using n-grams. Our data comes from a fully annotated corpus of 447 interactions of older and younger users with a Wizard-of-Oz (WoZ) appointment scheduling system (Georgila et al., 2008). We then evaluate these models using standard metrics (Schatzmann et al., 2005; Georgila et al., 2006) and compare our findings with the results of statistical corpus analysis.

The novelty of our work lies in two areas. First, to the best of our knowledge this is the first time that statistical SUs have been built for the increasingly important population of older users.

Secondly, a general (but as yet untested) assumption in this field is that current SUs are "enough like" real users for training good policies, and that testing system performance in simulated dialogues is an accurate indication of how a system will perform with human users. The validity of these assumptions is

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 49–52,
Columbus, Ohio, USA, June 2008.

a critically important open research question. Currently one of the standard methods for evaluating the quality of a SU is to run a user simulation on a real corpus and measure how often the action generated by the SU agrees with the action observed in the corpus (Schatzmann et al., 2005; Georgila et al., 2006). This method can certainly give us some insight into how strongly a SU resembles a real user, but the validity of the metrics used remains an open research problem. In this paper, we take this a step further. We measure the quality of user simulation models for both older and younger users, and show that these metrics are a good predictor of the behaviour of those two user types.

The structure of the paper is as follows: In section 2 we describe our data set. In section 3 we discuss the differences between older and younger users as measured in our corpus using standard statistical techniques. Then in section 4 we present our user simulations. Finally in section 5 we present our conclusions and propose future work.

2 The Corpus

The dialogue corpus which our simulations are based on was collected during a controlled experiment where we systematically varied: (1) the number of options that users were presented with (one option, two options, four options); (2) the confirmation strategy employed (explicit confirmation, implicit confirmation, no confirmation). The combination of these 3 × 3 design choices yielded 9 different dialogue systems.

Participants were asked to schedule a health care appointment with each of the 9 systems, yielding a total of 9 dialogues per participant. System utterances were generated using a simple template-based algorithm and synthesised using the speech synthesis system Cerevoice (Aylett et al., 2006), which has been shown to be intelligible to older users (Wolters et al., 2007). The human wizard took over the function of the speech recognition, language understanding, and dialogue management components.

Each dialogue corresponded to a fixed schema: First, users arranged to see a specific health care professional, then they arranged a specific half-day, and finally, a specific half-hour time slot on that half-day was agreed. In a final step, the wizard confirmed the appointment.

The full corpus consists of 447 dialogues; 3 dialogues were not recorded. A total of 50 participants were recruited, of which 26 were older (50–85) and 24 were younger (20–30). The older users contributed 232 dialogues, the younger ones 215. Older and younger users were matched for level of education and gender.

All dialogues were transcribed orthographically and annotated with dialogue acts and dialogue context information. Using a unique mapping, we associate each dialogue act with a ⟨speech act, task⟩ pair, where the speech act is task independent and the task corresponds to the slot in focus (health professional, half-day or time slot). For each dialogue, five measures of dialogue quality were recorded: objective task completion, perceived task completion, appointment recall, length (in turns), and detailed user satisfaction ratings. A detailed description of the corpus design, statistics, and annotation scheme is provided in (Georgila et al., 2008).

Our analysis of the corpus shows that there are clear differences in the way users interact with the systems. Since it is these differences that good user simulations need to capture, the most relevant findings for the present study are summarised in the next section.

3 Older vs. Younger Users

Since the user simulations (see section 4) are based mainly on dialogue act annotations, we will use speech act statistics to illustrate some key differences in behaviour between older and younger users. User speech acts were grouped into four categories that are relevant to dialogue management: speech acts that result in grounding (`ground`), speech acts that result in confirmations (`confirm`) (note, this category overlaps with `ground` and occurs after the system has explicitly or implicitly attempted to confirm the user's response), speech acts that indicate user initiative (`init`), and speech acts that indicate social interaction with the system (`social`). We also computed the average number of different speech act types used, the average number of speech act tokens, and the average token/type ratio per user. Results are given in Table 1.

There are 28 distinct user speech acts (Georgila et al., 2008). Older users not only produce more individual speech acts, they also use a far richer variety of speech acts, on average 14 out of 28 as opposed to 9 out of 28. The token/type ratio remains the same, however. Although the absolute frequency of confirmation and grounding speech acts is approximately

Variable	Older	Younger	Sig.
# speech act types	14	9	***
# speech act tokens	126	73	***
Sp. act tokens/types	8.7	8.5	n.s.
# Confirm	31	30	n.s.
% Confirm	28.3	41.5	***
# Ground	33	30	n.s.
% Ground	29.4	41.7	***
# Social	26	5	***
% Social	17.9	5.3	***
# Init	15	3	***
% Init	9.0	3.4	**

Table 1: Behaviour of older vs. younger users. Numbers are summed over all dialogues and divided by the number of users. *: p<0.01, **: p<0.005, ***: p<0.001 or better.

the same for younger and older users, the relative frequency of these types of speech acts is far lower for older than for younger users, because older users are far more likely to take initiative by providing additional information to the system and speech acts indicating social interaction. Based on this analysis alone, we would predict that user simulations trained on younger users only will not fare well when tested on older users, because the behaviour of older users is richer and more complex.

But do older and younger users constitute two separate groups, or are there older users that behave like younger ones? In the first case, we cannot use data from older people to create simulations of younger users' behaviour. In the second case, data from older users might be sufficient to approximately cover the full range of behaviour we see in the data. The boxplots given in Fig. 1 indicate that the latter is in fact true. Even though the means differ considerably between the two groups, older users' behaviour shows much greater variation than that of younger users. For example, for user initiative, the main range of values seen for older users includes the majority of values observed for younger users.

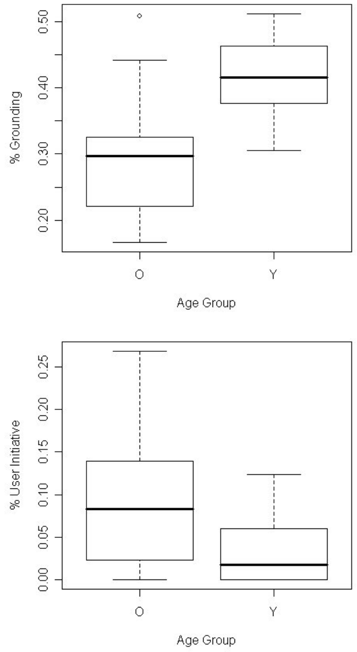

Figure 1: Relative frequency of (a) grounding and (b) user initiative.

4 User Simulations

We performed 5-fold cross validation ensuring that there was no overlap in speakers between different folds. Each user utterance corresponds to a user action annotated as a list of ⟨speech act, task⟩ pairs. For example, the utterance "I'd like to see the diabetes nurse on Thursday morning" could be annotated as [(accept_info, hp), (provide_info, halfday)] or similarly, depending on the previous system prompt. There are 389 distinct actions for older people and 125 for younger people. The actions of the younger people are a subset of the actions of the older people.

We built n-grams of system and user actions with n varying from 2 to 5. Given a history of system and user actions (n-1 actions) the SU generates an action based on a probability distribution learned from the training data (Georgila et al., 2006). We tested four values of n, 2, 3, 4, and 5. For reasons of space, we only report results from 3-grams because they suffer less from data sparsity than 4- and 5-grams and take into account larger contexts than 2-grams. However, results are similar for all values of n.

The actions generated by our SUs were compared to the actions observed in the corpus using five metrics proposed in the literature (Schatzmann et al., 2005; Georgila et al., 2006): perplexity (PP), precision, recall, expected precision and expected recall. While precision and recall are calculated based on the most likely action at a given state, expected precision and expected recall take into account all possible user actions at a given state. Details are given in (Georgila et al., 2006). In our cross-validation experiments, we used three different sources for the training and test sets: data from older users (O), data

	PP	**Prec**	**Rec**	**ExpPrec**	**ExpRec**
O-O	18.1	**42.8**	**39.8**	**56.0**	**49.4**
Y-O	19.6	**34.2**	**25.1**	**53.4**	**40.7**
A-O	18.7	**41.1**	**35.9**	**58.9**	**49.0**
O-Y	5.7	44.8	60.6	66.3	73.4
Y-Y	3.7	50.5	54.1	73.1	70.4
A-Y	3.8	45.8	58.5	70.5	73.0
O-A	10.3	**43.7**	**47.2**	60.3	**58.0**
Y-A	9.3	**40.3**	**33.3**	62.0	**51.5**
A-A	9.3	**43.2**	**43.4**	63.9	**57.9**

Table 2: Results for 3-grams and different combinations of training and test data. O: older users, Y: younger users, A: all users.

from younger users (Y), and data from all users (A). Our results are summarised in Table 2.

We find that models trained on younger users, but tested on older users (Y-O) perform worse than models trained on older users / all users and tested on older users (O-O, A-O). Thus, models of the behaviour of younger users cannot be used to simulate older users. In addition, models which are trained on older users tend to generalise better to the whole data set (O-A) than models trained only on younger users (Y-A). These results are in line with our statistical analysis, which showed that the behaviour of younger users appears to be a subset of the behaviour of older users. All results are statistically significant at $p<0.05$ or better.

5 Conclusions

In this paper we built user simulations for older and younger adults and evaluated them using standard metrics. Our results suggest that SUs trained on older people may also cover the behaviour of younger users, but not vice versa. This finding supports the principle of "inclusive design" (Keates and Clarkson, 2004): designers should consider a wide range of users when developing a product for general use. Furthermore, our results agree with predictions based on statistical analysis of our corpus. They are also in line with findings of tests of deployed Interactive Voice Response systems with younger and older users (Dulude, 2002), which show the diversity of older people's behaviour. Therefore, we have shown that standard metrics for evaluating SUs are a good predictor of the behaviour of our two user types. Overall, the metrics we used yielded a clear and consistent picture. Although our result needs to be verified on similar corpora, it has an important implication for corpus design. In order to yield realistic models of user behaviour, we need to gather less data from students, and more data from older and middle-aged users.

In our future work, we will perform more detailed statistical analyses of user behaviour. In particular, we will analyse the effect of dialogue strategies on behaviour, experiment with different Bayesian network structures, and use the resulting user simulations to learn dialogue strategies for both older and younger users as another way for testing the accuracy of our user models and validating our results.

Acknowledgements

This research was supported by the Wellcome Trust VIP grant and the Scottish Funding Council grant MATCH (HR04016). We would like to thank Robert Logie and Sarah MacPherson for contributing to the design of the original experiment, Neil Mayo and Joe Eddy for coding the Wizard-of-Oz interface, Vasilis Karaiskos and Matt Watson for collecting the data, and Melissa Kronenthal for transcribing the dialogues.

References

M. Aylett, C. Pidcock, and M.E. Fraser. 2006. The Cerevoice Blizzard Entry 2006: A prototype database unit selection engine. In *Proc. BLIZZARD Challenge*.

S. Czaja and C. Lee. 2007. The impact of aging on access to technology. *Universal Access in the Information Society (UAIS)*, 5:341–349.

L. Dulude. 2002. Automated telephone answering systems and aging. *Behaviour Information Technology*, 21:171–184.

K. Georgila, J. Henderson, and O. Lemon. 2006. User simulation for spoken dialogue systems: Learning and evaluation. In *Proc. Interspeech/ICSLP*.

K. Georgila, M. Wolters, V. Karaiskos, M. Kronenthal, R. Logie, N. Mayo, J. Moore, and M. Watson. 2008. A fully annotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems. In *Proc. LREC*.

S. Keates and J. Clarkson. 2004. *Inclusive Design*. Springer, London.

N. Roy, J. Pineau, and S. Thrun. 2000. Spoken dialog management for robots. In *Proc. ACL*.

J. Schatzmann, K. Georgila, and S. Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proc. SIGdial*.

M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens. 2007. Making synthetic speech accessible to older people. In *Proc. Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany*.