Wizard of Oz Experiments for a Companion Dialogue System: Eliciting Companionable Conversation

Nick Webb[1], David Benyon[2], Jay Bradley[2], Preben Hansen[3], Oil Mival[2]

(1) ILS Institute, SUNY Albany, Albany, NY, USA
(2) School of Computing, Edinburgh Napier University, Edinburgh, UK
(3) Swedish Institute for Computer Science, Stockholm, SE
nwebb@albany.edu, {benyon,j.bradley,mival}@napier.ac.uk, preben@sics.se

Abstract

Within the EU-funded COMPANIONS project, we are working to evaluate new collaborative conversational models of dialogue. Such an evaluation requires us to benchmark approaches to companionable dialogue. In order to determine the impact of system strategies on our evaluation paradigm, we need to generate a range of companionable conversations, using dialogue strategies such as 'empathy' and 'positivity'. By companionable dialogue, we mean interactions that take user input of some scenario, and respond in a manner appropriate to the emotional content of the user utterance. In this paper, we describe our working Wizard of Oz (WoZ) system for systematically creating dialogues that fulfil these potential strategies, and enables us to deploy a range of potential techniques for selecting which parts of user input to address is which order, to inform the wizard response to the user based on a manual, on-the-fly assessment of the polarity of the user input.

1. Introduction

Conversational agent technologies, such as those under development in the EU-funded COMPANIONS program[1] require new models of human-machine interaction. Companions are targeted as persistent, collaborative, conversational partners, where the user may have a wide degree of initiative in the resulting dialogue. Rather than singular, focused tasks, as seen in other deployed dialogue systems, fully developed Companions can have a range of tasks, and will be expected to switch between them on demand. Some of the tasks are not defined in such a way that an automatic system can know *a-priori* when the task is complete (such as annotating photographs) or even that the task itself is one of maintaining a relationship. In order to develop models for these agents, and test new methods of evaluation, we need a series of Wizard of Oz experiments to test Companion conversational behaviours.

In order to explore dialogue strategies for Companions, we are pursuing a series of Wizard of Oz (WoZ) experiments (Kelley, 1984). Utilising a WoZ approach allows us both to collect sample dialogues and interactions with our companion prototype systems (useful for developmental purposes), and to test final evaluation metrics on a complex, realistic set of interactions, so as to tune our evaluation parameters. This is important, as evaluation of companion systems cannot rely on known dialogue metrics such as task completion and user satisfaction. If the goal of the dialogue is to build and maintain a relationship, or to effect (positively) the mood of the user, we need to work with a series of realistic human-machine interactions to determine the impact of utterances on the user over the course of an interaction. Using our current WoZ system, we are investigating the impact of parameters such as content or timing of a companion utterance in a controlled environment, where the wizard has a strict series of guidelines to control the interaction to identify and/or react to certain user driven situations. These experiments will enable us to further refine our initial evaluation paradigm for companion technologies (Webb et al., 2010; Benyon et al., 2008). For example we are developing our scheme to include sub-categories of appropriate behaviour, such as appropriate uses of empathy or humour. For example, if the interaction proceeds as follows (with numbers added to identify individual utterances within a turn):

(1) `User:` *"I was having a good day (i). I had a pretty good meeting (ii). Later I was told I would be let go at the end of next week (iii)"*
(2) `System:` *"Well, it's good you had an overall positive day. I'm glad the meeting went well..."*

Here, the user turn contains three individual units or utterances. Of those, two ((i) and (ii)) are positive, and one (iii) is negative. Depending on the strategy the Companion deploys, it may behave appropriately or not. In our example, the system has determined that the overall tone of the user statement is positive, and is responding accordingly. However, this overlooks the vital importance of the negative information in utterance (iii). We need to annotate this utterance as being inappropriate to some strategy. To be able to reason about this in a meaningful way, we require benchmarking dialogues.

2. Companions Concept

The Companions vision is that of a personalised conversational, multimodal interface, one that knows its owner and is implemented on a range of platforms. Companions are advanced spoken language dialogue systems, that attempt to go beyond the limited functionality of current task-oriented systems, to be cooperative, collaborative dialogue partners, that form long term relationships with the user. Companions draw upon speech recognition, multimodal

[1] `www.companion-project.org`

human-computer interfaces, intelligent agents, knowledge representation and inference and human language technology all presented through an intuitive, natural interaction.
Benyon and Mival (2008) characterize Companions as an example of 'personification technology'. These are technologies designed so that people form relationships with them. The aim is to move from human-computer interaction to human-technology relationship design. Benyon and Mival (2008) identify five key features of technologies that need to be considered; utility, form, emotion, personality and social attitudes. In the case of Companions, conversation is the central part of the interaction, and it is thus primarily through conversation that relationships will be formed. We believe that human-computer dialogues can be evaluated in terms of the quality of speech, the dialogue itself, the tasks, the users and the appropriateness of the dialogue for the context in which it takes place.
In particular we are interested in developing behaviours and attitudes in people that demonstrate movement towards relationship forming. Several authors have shown the importance of recognising that people are quite keen to have relationships with technologies. Lester et al. (1997) discuss the persona effect and how having a character at the interface helped people to learn in an educational environment. Reeves and Nass (1996) discuss the 'media equation' and how people will attempt to form relationships with just about any technology. Bickmore and Picard (2005) argue that maintaining relationships involves managing expectations, attitudes and intentions. They emphasise that relationships are long-term built up over time through many interactions. Relationships are fundamentally social and emotional, persistent and personalised.
For the current Companions demonstrator, we are focusing on a system that performs around a "How was your day" scenario. The basic construct of the scenario is that, following a day at work, the user engages in dialogue with the system, evaluating their day, recounting events (such as meetings, presentations and appointments) and interactions with people (having coffee with Bob, for example). The Companion will respond in an emotionally appropriate manner to the input. Emotional appropriateness can be defined in many ways, and in this paper we explore two possible options, *empathy* and *positivity*.
In order to move from our initial demonstrators toward systems that are more 'companionable' in nature, we need to identify those behaviours required from Companion systems, and develop an evaluation strategy for these features.

3. Wizard Strategies

In order to test our evaluation paradigm on companionable dialogues, we devised a small series of WoZ experiments using a system developed at Napier University (Bradley et al., 2009), which is functionally similar if less developed than the Sensitive Artificial Listener (SAL) system[2], that allows us to explore issues surrounding the "How Was Your Day" (HWYD) scenario. We decided to focus on three potential strategies for the Companion to adopt in response to the HWYD scenario: *empathy*, *positivity* and *adaptive*. These are of course not the only strategies to employ, but they reflect the choices made by the Companion system designers for the advanced prototype. Focussing on these strategies allows us to make assumptions about the techniques a system could use to achieve companionable behaviour.
Empathy is defined by us to mean that the Companion always tries to mirror the overall mood of the user. Positivity is a more proactive approach, where the Companion tries to move the mood of the user toward the positive. Adaptive is the most complex strategy, where the Companion deploys both empathy and positivity in response to context, the user model, and interactions with the user. To achieve these strategies, there are a number of possible techniques, and for WoZ purposes, we looked at selection of these, given below.
First of all, we need some way of assessing the overall mood of the user, based on their contribution to the dialogue. At this stage in the Companions project, we are using only text based mood classification, ignoring for instance facial expressions or measurable affective indicators outside of prosody. Each user can (and is expected to) utter long utterances, containing many pieces of information. Take for example turn (1) from our previous example. These turn contains three utterances. We indicate that the overall mood of the utterance is positive, but that there is important specific negative information presented in the final utterance (iii). We need to keep a track of the current 'mood score' given the input from a user, and compare that to instances where we try to pay attention to specific pieces of information in the turn. We calculate a cumulative score for the entire user turn and use this as a shorthand for determining the overall mood. To obtain the cumulative score, the wizard analyses each utterance in the user turn, and calculates the approximate sentiment of each on a coarse scale. The wizard then sums the sentiments (for our examples, we use coarse grain positive (+1), negative (-1) and neutral (0) representations), and uses this score to reflect the mood of the user in the current turn. This is currently performed by hand by the wizard in real time, although in future WoZ experiments, we wish to use the same mechanisms as the Companions demonstrator directly, such as textual sentiment analysis (Strapparava and Mihalcea, 2008) and EmoVoice (Vogt et al., 2008) for identifying emotion in speech.
For example, the user turn *"I had a terrible meeting, where my presentation went really badly. But then I had a great conversation over coffee with my boss"* could be scored as (-1 + -1 + 1 = -1, or negative overall). This is a very fast approximation of the overall mood of the user. Once this has been established, the wizard must choose a *technique* to identify how to respond to the user. These techniques are independent of the overall strategy, but present a structured way in which a Companion, employing a given strategy, might choose to respond to a user. We investigated four techniques for managing this response, two based on order of occurrence of concepts in the user turn, one based on the polarity of the individual utterances with respect to the overall strategy, and one based on specific utterance content. These choice of these techniques is arbitrary, but we

[2] www.vf.utwente.nl/ hofs/sal/

choose them only to motivate a range of possible interactions. We give a short overview of each technique, below.

- TEMPORAL ORDER: Deal with each user utterance in the order in which it occurs. Example:

  `User:` *"I had a terrible meeting, where my presentation went really badly. But then I had a great conversation over coffee with my boss"*
  `System (empathetic):` *"I'm sorry to hear about your meeting, that's terrible. And your presentation too, I know you worked so hard on it. But it is great that you got a chance to talk with Bill"*
  `System (positivity):` *Oh, that's a shame about the meeting, but you shouldn't worry. You said that the meeting would be bad, and that it didn't count for much. And people usually like your presentations, so it must have just been an off day...."*

- RECENCY ORDER: Deal with each user utterance in reverse order (i.e. the most recent thing is the most important). Example:

  `User:` *"I had a terrible meeting, where my presentation went really badly. But then I had a great conversation over coffee with my boss"*
  `System (empathetic):` *"well, it's great that you had some one to one time with your boss, but what a shame about the other things that happened"*

- TARGET (positive or negative): Rather than address all (or an arbitrary) utterance in the user turn, target a particular utterance with respect to a strategy. For example, if the strategy is positivity, the Companion could either emphasise (and therefore talk about) the positive utterances in the turn (reinforcing the good things) or address (and repair) the negative utterances in the turn. Example:

  `User:` *"I had a terrible meeting, where my presentation went really badly. But then I had a great conversation over coffee with my boss"*
  `System (positivity, target +ve):` *"well, it's great that you had some one to one time with your boss, that's a useful thing to have done"*
  `System (positivity, target -ve):` *Wow, I can't believe the presentation went badly, you worked so hard, maybe they didn't have enough time to understand the contents...."*

- PRIORITISE: Using information from the context or the user model, identify which of the utterances in the turn is most important at that time. We include this option as we believe it is clearly important to drive the behaviour of the Companion, but perform minimal evaluation at the moment, until integration of the user model is completed. Example:

  `User:` *"I was having a good day. I had a pretty good meeting. Later I was told I would be let go at the end of next week"*
  `System (empathetic):` *Oh my, you've been let go. That's terrible. How awful. You must feel really down"*
  `System (positivity):` *You've been let go? Well, that's never good news, but in the last two months you've been really unhappy there. And you've often said they don't pay well, and the benefits are poor. So I suppose here is your chance to make a fresh start"*

This list does not represent an exhaustive set of techniques for a Companion to adopt. Instead it represents a small set of possible techniques used to guide the wizard in our user sessions, in an attempt to generate both appropriate and inappropriate responses to user turns. We explored each of these techniques through a series of small, in-house WoZ sessions, and identified that in order to best achieve our desired strategies (of empathy and positivity), a simple Wizard Strategy Matrix (shown in Figure 1) could be used. The first step of such a matrix is to use the CUMULATIVE method to determine the overall score of the user turn, by looking at the individual concepts and their sentiment. Then, depending on the current strategy of the wizard, the matrix indicates which concept (or concepts) in the user utterance to address, either the positive or negative concepts. Finally, the wizard has to implement one of the techniques listed above. Although this appears to be a lot of processing on the part of the wizard, most of these choices are made a-priori. For example, for one particular session, it will be pre-determined that the wizard is adopting the strategy of "positivity' '. The wizard calculates the cumulative score on the fly, and is reminded which concepts overall to target for the given strategy. The wizard then applies one of the techniques (again, chosen a-priori), such as "recency order", to determine which specific concept (in the case of multiple options) to address.

This use of the matrix and the techniques only ensures that we generate a range of possible interactions, which can be assessed at both a global level (did the interactions seem reasonable) and at the utterance level (using appropriateness annotation, as described in Webb et al. (2010)). Note, this matrix says nothing about how the wizard should address the utterance, leaving that to the discretion of the wizard. The matrix was created using a combination of wizard intuition, in combination with early results from WoZ experiments.

The underlying assumptions here are as follows. When the system is being empathetic, it should mirror the overall sentiment of the user. To do so, when the overall sentiment is positive, the wizard will ignore any negative input in the user turn. Similarly, when the overall input is negative, the wizard issues platitudes with respect to only the negative concepts in the user turn. Note we don't include anything in this table about overall neutral statements. When there is, for example, one negative and one positive concept in the turn (i.e. a cumulative score that is neutral). We decided that the wizard should select one of the techniques (such as recency or temporal order) and apply them to the input, irrespective of the polarity of the selected concept.

System Strategy	Cumulative Score		
	Overall Positive		Overall Negative
EMPATHY	Identify with positive concepts		Identify with negative concepts
POSITIVITY	Talk about positive concepts	If there is a negative concept, try to make it better	Choose negative concept, try to make it better

Figure 1: Example implementations of two WoZ strategies, empathy and positivity

Using this combination of strategies and techniques, we have so far collected 20 dialogues with the wizard, fulling ten possible combinations of concepts in the user input. We selected a maximum of three concepts per input, from scenarios given to users matching the 'HWYD' scenario. The fourteen possible combinations of concepts (represented by (+ve) for positive concepts and (-ve) for negative concepts) are shown in Table 3., where our example of earlier:

`User:` *"I was having a good day (i). I had a pretty good meeting (ii). Later I was told I would be let go at the end of next week (iii)"*

would match the condition (+ve) (+ve) (-ve) . Conditions (1), (2), (8) and (9) are relatively straightforward choices (in (1) and (8) there is only a single concept to address. In (2) and (9), whilst there are more concepts, they share polarity, and will have the same issues as conditions (4) and (11)) and so were excluded from this first experiment. Instead we concentrated on the 10 remaining conditions, and asked a wizard to role play these scenarios using each of the two strategies, positivity and empathy. Note that whilst combinations such as (6) and (13) are equivalent on a cumulative scale (both are a net negative), there is a difference with respect to the techniques used to address them. If we are using temporally based techniques, it will alter which of the concepts we address first.

(1) (+ve)	(8) (-ve)
(2) (+ve) (+ve)	(9) (-ve) (-ve)
(3) (+ve) (-ve)	(10) (-ve) (+ve)
(4) (+ve) (+ve) (+ve)	(11) (-ve) (-ve) (-ve)
(5) (+ve) (+ve) (-ve)	(12) (-ve) (+ve) (+ve)
(6) (+ve) (-ve) (-ve)	(13) (-ve) (-ve) (+ve)
(7) (+ve) (-ve) (+ve)	(14) (-ve) (+ve) (-ve)

Table 1: Possible concept combinations for user input

Having collected initial interactions, we are now in the process of refining our wizard intructions for a larger scale data collection.

4. Future Work

The next step for our WoZ experiments is to focus more clearly on what the Companions consortium terms 'interruptions' (although others might consider them 'feedback', or in some cases 'back channels'). We will look at two issues, the timing and the content. For the timing, we see that the user could interrupt at the beginning of system output, somewhere in a notional mid-point of an utterance, or at the end (in which case, it's not really an interruption). From a content perspective, we see that there are generally positive interruptions (*"yes, you're right"*), generally negative interruptions (*"no, that's not the problem"*) and interruptions that can be seen as somewhat neutral (*"well, maybe"*). The impact of these interruptions is intuitively clear, and for the wizard with respect to our interaction strategies, could be the key to a third strategy, that combines parts of an empathetic and positive approach into an 'adaptive' strategy. Consider the following example, where ellipsis (*"..."*) at the end of the system utterance indicates that the user interrupts at this point:

`User:` *"I had a terrible day, the meeting really didn't go well, and Sarah couldn't help"*
`System:` *"Well, that's a shame about Sarah, you think a friend would do more..."*
`User:` *"No, no, it's not Sarah's fault, it's that my boss didn't have clear expectations"*

In this example, the user has interrupted the system early in the output (the system has talked about just one concept, in this case the last one introduced by the user) and has done so negatively (using the cue phrase *"no"* repeatedly, for example). We can infer that this interruption should change the nature of the system output (and probably update the user model) to reflect that the meeting (and the boss) have negative polarity. Contrast this with the following example:

User: *"I had a terrible day, the meeting really didn't go well, and Sarah couldn't help"*
System: *"Well, it was nice of Sarah to try..."*
User: *"Well, I suppose so"*

Here we read that although the interruption is early, it is somewhat positive. We hypothesise that under this condition the Companion will continue with the turn as previously planned. We intend to experiment with three values for the interruption timing (early, middle, late) and three values for the interruption content (positive, negative, neutral) and determine how the system response to these interruptions could be annotated to reflect appropriate or inappropriate system behaviour. Finally, we will experiment with issues surrounding system delay. It seems clear that the level of acceptable delay is correlated with user experience of companion-like technologies. Experienced users are more tolerant to extended system delay (under the hypothesis that we know the system to be working). The same is not true for naive users, and we wish to explore the parameters of delay (what is appropriate time to assume end of user turn, what delay is acceptable before system response, should time be left to allow or facilitate interruption).

5. Discussion

So far, we have conducted a small number of WoZ dialogues in total using the interaction strategies of 'empathy' and 'positivity'. Our simplified early wizard experiments indicate that the most challenging areas (i.e. those in which the best strategy is really not clear) are when the mood is generally positive, and the system is deploying a positive strategy (what can a wizard say to someone that doesn't appear patronising), and for both strategies when the overall mood of the utterance is neutral. It seems clear already that the only way to correctly address some situations is to have a deeper understanding of the content of the concepts, as opposed to the shallow analysis we currently perform. We are addressing these situations with new WoZ experiments, and are in the process of exploring the wizard matrix more thoroughly, by deploying the WoZ experiment with a handful of naive users (users not familiar with the Companion concept), to determine if the data we generate for short interactions is useful for testing our evaluation paradigm using appropriateness annotation (Webb et al., 2010).

6. Acknowledgements

This work was funded by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

7. References

D. Benyon and O. Mival. 2008. Landscaping Personification Technologies: From Interactions to Relationships. In *Proceedings of CHI2008, Extended Abstracts, ACM*.

D. Benyon, P. Hansen, and N. Webb. 2008. Evaluating Human-Computer Conversation in Companions. In *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy.

T. Bickmore and R. Picard. 2005. Establishing and Maintaining Long-Term Human-Computer Relationships. *Transactions on Computer-Human Interaction*, 12(2).

J. Bradley, O. Mival, and D. Benyon. 2009. Wizard of Oz Experiments for Companions. In *Proceedings of HCI'09*.

J. F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. *Transactions on Office Information Systems*, 2(1).

J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal. 1997. The Persona Effect. In *Proceedings of CHI1997, Extended Abstracts, ACM*.

B. Reeves and C. Nass, editors. 1996. *The Media Equation*. CSLI Publications, Stanford, CA.

C. Strapparava and R. Mihalcea. 2008. Learning to Identify Emotions in Text. In *Proceedings of the ACM Conference on Applied Computing ACM-SAC 2008*.

T. Vogt, E. André, and N. Bee. 2008. Emovoice – a framework for online recognition of emotions from voice. In *PIT '08: Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 188–199, Berlin, Heidelberg. Springer-Verlag.

N. Webb, D. Benyon, P. Hansen, and O. Mival. 2010. Evaluating Human-Machine Conversation for Appropriateness. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.