

Evaluating Human-Computer Conversation in Companions

David Benyon¹, Preben Hansen² and Nick Webb³

Abstract. We report on the first evaluation of the Companions project prototypes. We give preliminary results from our phase one evaluation, using known and well-understood dialogue metrics. We also give a first indication of the directions we plan to take to evaluate increasingly sophisticated conversational systems, using measures of coherence and appropriateness.

1 INTRODUCTION

The Companions vision is that of a personalised conversational, multimodal interface, one that knows its owner and is implemented on a range of platforms, both static and mobile. 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 human-computer interfaces, intelligent agents, knowledge representation and inference and human language technology — all presented through an intuitive, natural interaction.

Wilks [11] argues that Companions are intelligent, personalized, persistent, multimodal interfaces to the Internet, that are intended to be humane conversational partners. The project is now into its second year, and the two initial, first phase demonstrators have been completed. There is a Senior Companion (SC) that engages the user in a conversation relating to a set of photographs – seen as a first step toward eliciting memories from a user, and a Health and Fitness Companion (HFC), which attempts to promote exercise and a healthy diet in the user. The HFC has both a static ‘home’ part where people talk to a Nabaztag character (a physical plastic rabbit-like artefact that has moving ears and flashing lights) and a mobile part where they interact through a mobile phone or device. In order to move from these 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. We begin by evaluating the current prototypes using existing, well-understood evaluation mechanisms and techniques.

Benyon and Mival [1] 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 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. For measures of appropriateness and coherence we will look to adapt current measures developed for complex speech systems [8].

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 [4] discuss the persona effect and how having a character at the interface helped people to learn in an educational environment. Reeves and Nass [6] discuss the ‘media equation’ (media = real life) and how people will attempt to form relationships with just about any technology. Bickmore and Picard [3] 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.

In this article, we outline the approach we have taken for the initial evaluation of the phase one Companion prototypes – and report the preliminary results of this evaluation phase. Subsequently, we will show the evaluation goals for later phases of development, and identify how we hope to build on existing evaluation metrics to measure companion-ness. The paper is organized as follows. First, we summarize related work in this area. Next, we describe the phase one evaluation taking place now with the initial Companions prototypes. In Section 4, we talk about the next iteration of the evaluation programme, including our intended evaluation paradigm. Section 5 presents our preliminary conclusions and provides directions for future work.

2 PHASE ONE EVALUATION STRATEGY

In June and July 2008 we evaluated three manifestations of the Companions concept – the Senior Companion (SC), the Health and Fitness Companion (HFC) [8] and the mobile HFC [7]. The purpose of the evaluations was as much to refine the evaluation protocol and data gathering method as it was to evaluate the products. However, some useful base-line data has been gathered.

The evaluations cover a variety of situations and looked at a variety of measures. Some of these seek to evaluate improvement over the state-of-the-art, others seek to evaluate the concepts behind the prototypes, or to measure some specific aspect such as the level of social presence experienced with certain design choices. Future evaluations will be a mixture of

¹ School of Computing, University of Napier, EH14 1LJ, UK. Email: d.benyon@napier.ac.uk.

² Swedish Institute of Computer Science, Kista, SWEDEN. Email: preben@sics.se.

³ ILS Institute, University at Albany, SUNY, NY, 12222, USA. Email: nwebb@albany.edu.

lab-based using students as surrogate Companion ‘owners’, some will be more ‘ecologically valid’ and will make use of real, non-technical people from different groups (e.g. the elderly).

The mechanisms for evaluation are two-fold. Qualitative surveys are used to acquire subjective opinions from the users of the Companions prototypes, in conjunction with quantitative measures, a summary of which are given below. We will analyse the resultant dialogues between users and companions to calculate measures relating specifically to the speech component, the dialogue performance, users experience and task completion as a whole.

Table 1: Types of Metrics

| Metrics | Examples |
|------------------|--|
| Speech metrics | WER, CER, etc |
| Dialogue metrics | Dialogue duration, number of turns, word per turn, etc |
| Task metrics | Task completion, etc |
| User metrics | Satisfaction, etc |

Eight people completed the whole protocol. Each participant had to complete four distinct tasks; introductory tutorials, using prototypes, on-line surveys and interviews. The SC had a voice training exercise with the “*Dragon Natural Language*” software before its associated introductory tutorial. This voice training exercise took five to seven minutes to complete.

Each session began with an introductory tutorial. These ten to sixteen slide presentations introduced the prototype, established its intentions, its limitations, what the prototype would say and do, how to use the prototype and give the user suggestions in how to respond. The figures below provide an introductory tutorial example that shows the slides of the SC.

Participants then used the SC, HFC and mobile HFC for 10 - 15 minutes each, completing the on-line questionnaire after each session. Researchers were sitting in the background while the participants interacted with the prototypes, and participants were video taped during their interaction. Researchers were able to intervene in case of catastrophic failure. Finally the participants were each interviewed by a researcher.

3 OBJECTIVE SPEECH AND DIALOGUE METRICS

Taking the metrics outlined in table 1, we collected standard timing information from each interaction – to establish baseline guides for the usability and ‘stickiness’ of each prototype. Turn and utterance durations (in seconds) are available for both systems, however because the systems work in different ways and timestamps have been produced at different points, the figures are not directly comparable. For example, the SC has an always-open microphone channel, whereas the HFC has a push to talk feature. For HFC, the user utterance length is calculated from the time when first click of Nabaztag button was received until the second, terminating click was received. System utterance length is calculated from moment when audio output started to the moment when it ended. Everything between those events is reported as delays. For SC, the system itself has calculated average system and user turn durations. Of these, system turn duration is similar to HFC system utterance

durations, while user turn duration is comparable to sum of the HFC user utterance duration and the two delays.

Vocabulary sizes and utterance lengths (in words) are available both based on ASR results and on transcriptions. Word error rate (WER) has been calculated using the standard formula; (Deletion Errors + Insertion Errors + Substitution Errors) / (number of words actually uttered by user). Regular dynamic programming string alignment has been used to calculate the errors. Concept Error Rate (CER) has been calculated by ignoring the order of recognised concepts, substitution errors are used only for cases where part of the recognised and actual concepts match. Substitution of a concept with Deletion Errors + Insertion Errors + Substitution Errors/ actual concepts: CER can be (significantly) > 1 when lot of concepts were inserted (in many cases it is 5 or 6). Currently CER is available only for HFC, as SC did not have any concept information on the log files of this evaluation phase. To calculate preliminary CER for the SC, transcribers add that information according to their best guesses.

Dialogues with SC had between 100 and 160 dialogue turns (sum of both user and system turns). Dialogue durations were between 9 minutes 20 seconds and 15 minutes 15 seconds. HFC dialogues had between 20 and 74 turns and lasted between 3 minutes 15 seconds and 12 minutes and 45 seconds. Since the set of tasks were given to the participants, these figures are only descriptive of the evaluation procedure.

Average length of user utterances varied between participants from 2.9 and 6.8 words for SC and between 3.0 and 8.3 words for HFC. It can be seen that there are significant differences in how verbose different people are. While the small dataset does not allow statistical testing, the utterance lengths used by one person with SC seem to be rather well in line with their utterance lengths with HFC. Comparing the actual utterance lengths with ASR results, ASR in SC recognises fairly closely the same amount of words as uttered, while HFC recogniser tends to recognise fewer words, i.e. makes a high degree of deletion errors, as might be expected for a trained, single user ASR< versus a large vocabulary, multi-user system. Average system utterance length for SC is around 14 words and for HFC 12 words.

Vocabulary size used by people with SC ranged between 33 and 131 words, while HFC resulted in vocabularies between 18 and 116 words. The average of these is 70 for SC and 55 for HFC. The larger vocabulary of SC dialogues is to be expected due to system’s more open questions; in fact, it is somewhat surprising that the vocabularies are so small.

Word error rates for SC range between 0.12 and 0.37. Many of the errors are small insertion errors, but there are cases, where larger segments are completely misrecognised. Word error rates for HFC range between 0.79 and 0.51, with one case where error rate was over 1 because of massive amount of rejection errors. While the word error rates of HFC are extremely high, concept error rates are somewhat smaller, between 0.33 and 0.65. These numbers are still high, but most errors are insertions of several concepts in some specific cases, while most of the time concept error rate was reasonably good. The rule-based grammars of HFC currently seem to have both insufficient coverage and precision, resulting in a recogniser that easily inserts incorrect concepts into the result.

4 SUBJECTIVE PERCEPTION OF THE COMPANIONS

Measures of how people related to the Companions were collected through on-line questionnaires. The SC consisted of forty questions that were answered on a 5-point Likert scale (strongly agree, agree, undecided, disagree, strongly disagree). The last ten questions were concerned with gathering feedback about the aesthetics of the interface, in order to inform subsequent designs. The first thirty were aimed at validating a model of companions [1] and at establishing a base line for further developments. Twenty-seven responses were collected.

The questions were organised around six themes:

Table 2: Six themes for the questionnaire

| |
|---|
| A. The behaviour of the Companion and what it looked like |
| B. The utility of the Companion |
| C. The nature of the relationship between participant and Companion |
| D. The emotion demonstrated by the Companion: |
| E. The personality of the Companion |
| F. The social attitudes of the Companion |

The HFC used the same set of questions, but allowed for people to provide additional comments to explain their choice. Eight responses were collected.

The mobile HFC used open-ended comments for a set of ten questions mostly concerned with the functionality and interaction style of a mobile companion. Eight responses were collected.

The Likert scales asked people to indicate whether they agreed or not with statements such as those shown in table 3:

Table 3: Examples of questions asked with Likert scale

| |
|--|
| The dialogue between the Companion and me felt natural |
| I thought the dialogue was appropriate |
| Over time I think I would build up a relationship with the Companion |

The full set of questions is shown below. The answers were scored as 1 for strongly agree through to 5 for strongly disagree. Much of the data was biased to the 'undecided' option, partly because the prototypes are still in their early stages and are not achieving the higher-level ambitions of Companions. However, some very strong opinions were elicited. Twenty-seven responses were received to the SC. Nine of those scored an average of over 3.0 and one scored an average of 1.8. All the others scored between 2 and 3.

On average people *disagreed* (average score 3.0 – 3.5 with the following statements:

Table 4: Disagreements with the SC

| |
|--|
| The dialogue between the Companion and me felt natural |
| I liked the behaviour of the Companion |
| Over time I think I would build up a relationship with the Companion |
| The Companion showed empathy towards me |
| The Companion demonstrated emotion at times |
| The Companion was compassionate |

They *strongly disagreed* (average score 3.5 – 4.0) with the statements:

Table 5: Strong disagreements with the SC

| |
|--|
| I felt I could correct the Companion when necessary |
| The Companion got to know me during the conversation |
| The Companion is rather like me. |

They *agreed* (average score 1.8) with the statement:

“The Companion was polite.”

For the HFC there was a similar response with people disagreeing (average score 3.0 – 3.5) with the statements:

Table 6: Disagreements with the HFC

| |
|--|
| The dialogue between the Companion and me felt natural |
| I felt I could correct the Companion when necessary |
| The Companion showed empathy towards me |
| The Companion surprised me at times |
| The Companion anticipated my needs |

People *disagreed* more strongly (average score 3.5 – 4.0) with the statements:

Table 7: Strong disagreement with the HFC

| |
|---|
| The Companion showed empathy towards me |
| The Companion demonstrated emotion at times |
| The Companion is rather like me. |

They *agreed* (average score 2) with the statement:

The Companion was polite.

5 CONCLUSIONS

This is an early attempt to characterise Companions as a new form of interaction between people and technologies. Suitable forms of evaluation are required if we are to move in this direction. The combination of speech and dialogue metrics with people's perceptions of relationship building is what we hope will provide a good combination to assess companionship. Existing dialogue metrics concentrate too closely on task driven dialogue [2,5], where for example a number of features are reduced to a single parameter maximising, for example, user satisfaction [10]. We are looking to draw new correlations between subjective measures of system performance, with observable objective metrics.

For this initial trial, we can already see some indicators of the directions developers might take to address issues of with companions – including addressing the naturalness of the overall interaction. It is hoped that advances in Machine Learning from dialogue corpora, dealt with by other research groups within the Companions project, will enable this to happen on a scale large enough to be effective for subsequent demonstrators.

Some additional factors we wish to address – using system metrics to guide development – include:

Conversation focus versus explicit (or implicit) task focus.

At present, both the systems are task focused. As the focus shifts toward successful conversation, we should expect, for example, initiative rates, words per user utterance and dialogue duration to increase. In particular, both systems have a high degree of system initiative. Users should be able to ask questions and provide more information than was asked for – information captured by dialogue structure and concept density measures. All interactions should maintain appropriateness, by some measure.

Relationship Building

We believe that we can measure this through focus groups, but also expect observable changes in raw data - for example, people would interact longer or more frequently with a companion with which they are forming a positive relationship. They may also offer increased amounts of information, use more informal language, or longer, less task-directed utterances. Furthermore, we expect that the use of language may change over time from a more formal to more informal interaction between human and companion.

Cooperation

Given the possibility to cooperate a companion to achieve a task, it should be the case that those working with collaborative, cooperative partners are able to achieve more, faster, than those without access to the technology. More generally, how do the companions facilitate the use and persistence of the models (of data, and of the user) that they acquire. Observations as well as log data would capture this type of behaviour.

Multi-modality

There are presently no specific metrics measuring additional modalities. We need to define to what extent extra modalities form part of the companion experience in the future. Included in a measure of modality is the difference in use across companion domains, and user groups. People will probably interact in different way with the companion due to domain, age, gender etc. Different interaction types between the human or groups of humans and the companion may occur. For example, an older person may interact differently during her/his ageing process.

Mobility

We have a mobile version of the HFC. We hope to see further integration between this and the home version, and consideration of a mobile version of the SC. Further evaluation should consider the mobile application as an extension of, rather than an additional version of, the home based systems.

REFERENCES

- [1] Benyon, D. R and Mival, O. (2008) Landscaping personification technologies: from interactions to relationships. In Proceedings of CHI2008, Extended Abstracts, ACM.
- [2] Bernsen, N. O.: DISC dialogue engineering best practice guide glossary for dialogue management and human factors. NISLab November 1999. URL: <http://www.disc2.dk>
- [3] Bickmore T. and Picard R. (2005) Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction (TOCHI), Volume 12 Issue 2
- [4] Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., and Bhogal, R. S. "The persona effect". In Proceedings of CHI 1997, 359-366.
- [5] den Os, E., and G. Bloothoof, 1998. Evaluating various spoken language dialogue systems with a single questionnaire: Analysis of the ELSNET olympics. Proceedings of the 1st International Conference on Language Resources and Evaluation, 51-5
- [6] Reeves B. and Nass, C. (1996) The Media Equation CSLI Publications; Stanford, CA
- [7] Ståhl, O., Gambäck, B., Hansen, P., Turunen, M. And Hakulinen, J. (2008). A Mobile Fitness Companion 4th International Workshop on Human-Computer Conversation, Bellagio, Italy, 2008.
- [8] David R. Traum, Susan Robinson, Jens Stephan Evaluation of multi-party virtual reality dialogue interaction, In Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004), pp. 1699-1702.
- [9] M. Turunen, J. Hakulinen, O. Ståhl, B. Gambäck, P. Hansen, M.C. Rodríguez Gancedo, R. Santos de la Cámara, C. Smith, D. Charlton, and M. Cavazza, (2008). "Multimodal Agent Interfaces and System Architectures for Health and Fitness Companions", 4th International Workshop on Human-Computer Conversation, Bellagio, Italy, 2008.
- [10] Marilyn Walker, Diane Litman, Candace Kamm and Alicia Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents . PDF . In Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics , ACL 97, 1997.
- [11] 9Wilks, Y. (2006) Artificial Companions as a new kind of interface to the future internet. Oxford Internet Institute, Research Report 13, October 2006