Dialogue, Speech and Images: The Companions Project Data Set

Yorick Wilks*, **David Benyon**†, **Christopher Brewster***, **Pavel Ircing**‡, **and Oli Mival**†

*Department of Computer Science, University of Sheffield, Sheffield, S1 4DP,
Initial.Surname@dcs.shef.ac.uk

†School of Computing, University of Napier, Edinburgh, EH10 5DT,
Inial.Surname@napier.ac.uk

‡Department of Cybernetics, University of West Bohemia, Pilsen, Czech Republic,
ircing@kky.zcu.cz

Abstract

This paper describes part of the corpus collection efforts underway in the EC funded Companions project. The Companions project is collecting substantial quantities of dialogue a large part of which focus on reminiscing about photographs. The texts are in English and Czech. We describe the context and objectives for which this dialogue corpus is being collected, the methodology being used and make observations on the resulting data. The corpora will be made available to the wider research community through the Companions Project web site.

1. Introduction and Context

It is widely agreed that research on dialogue is the in some ways the Cinderella of NLP: there is far less access to naturally occurring dialogue corpora on the web and elsewhere than for other, text-centred, areas of NLP. This has hampered the use of data-driven methods and machine learning for the development of dialogue systems. Researchers constantly go back, as a sort of Rosetta stone, to the ATT SWITCHBOARD corpus (Godfrey et al., 1992) in spite of all its limitations, rather in the way other areas of NLP was over-focussed on the Wall Street Journal Corpus in the 1990s.

The COMPANIONS project (Wilks, 2005) (http://www.companions-project.org) is attempting to develop new sources of corpora for dialogue systems, and new methods for deriving them. One of our initial demonstrators, the Senior Companion (SC), is aimed at the elderly, and thus we need to collect data for this domain because dialogue structures do not seem to transfer well from domain to domain, nor from specially selected group to subject group. The SC will initially cover the domain of subjects reminiscing about photographs and the families and friends in them. There are two important ideas in the plan we have begun to execute. First, that the initial implementation will be minimal, based on limited data gathered by WOZ methods (see below), so that the initial implementation itself can be used as a device to produce more data on a larger scale, by analogy with the ways in the which first sketch POS taggers and even dialogue act taggers have been applied to texts so as to gather further data for eye/hand inspection and subsequent machine learning. The project is also developing a second early prototype—the Health and Fitness Companion—-but there the data requirements are different as is the methodology of further data acquisition; so, although the prototypes will share modules, we shall concentrate in this paper on corpus data issues concerning the SC.

Secondly, the AI implementation philosophy behind the plan for the whole project is to have a Phase I where quick implementations are built and first stage machine-learning is done over corpora produced in the way just described. That phase should take the first 18 months of a 48 month project; in parallel with this is a 1-48 month Phase II. This is investigating more sophisticated ML methods and is designed to absorb and benefit from the results of the Phase I corpus projects. The motive here is to escape what one could call the paradox of AI development: if one implements quickly—the methodology that used to be called rapid prototyping—– the results are never thrown away in later versions and all the design limitations based on early choice are preserved, with bad long term consequences. If, on the other hand, one waits to the end of a project to implement, then that project is never properly evaluated or developed from; it is usually finished just in time for the final report. Our plan is to do both at once in an attempt to gain the benefits of both parts of the paradox and not the disadvantages.

The Companions project involved a number of institutions with different specialisations. It includes partners who specialise in NLP and related technologies, partners specialising in speech recognition and synthesis, as well as specialists in human-computer interaction. The requirements of each of these actors affected the collection of the data.

The rest of this paper is organised as follows: In Section 2, we describe the requirements from the perspectives of these actors. In Section 3, we describe the data collection procedure undertaken by Napier University (Data Set 1) and our future plans for disseminating this data set. The project also produced a different data set collected by partner AsAnAngel, and this will be described elsewhere. In Section 3, we describe how this data has been used to set up the initial prototype SC that is currently being used to generate more data.

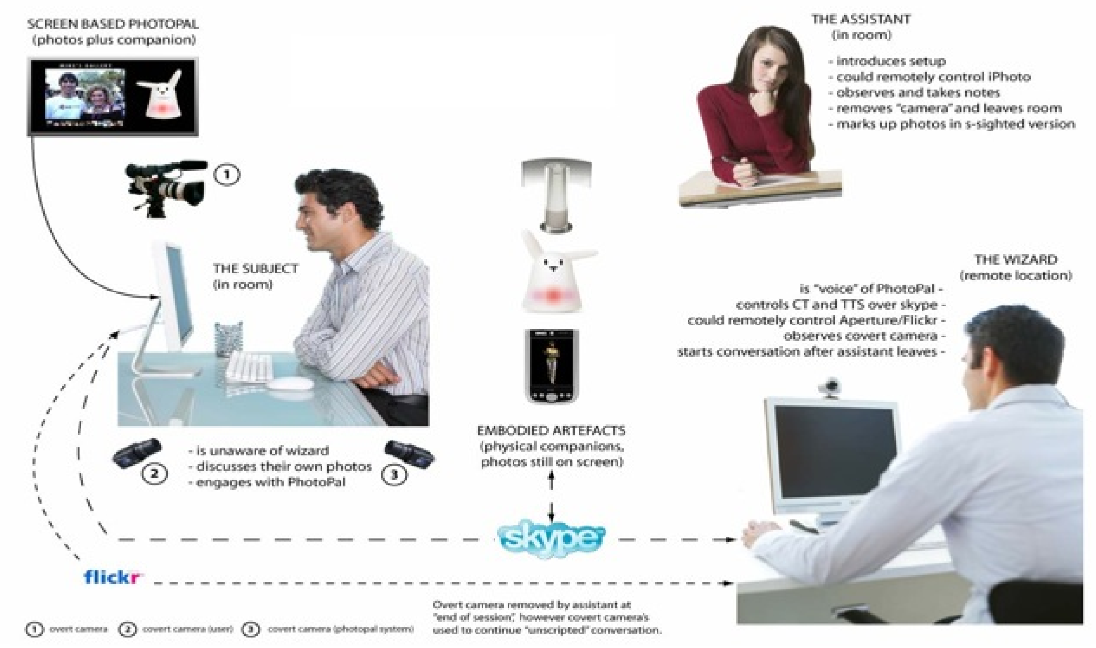
SCREEN BASED PHOTOPAL
(photos plus companion)
1
THE SUBJECT
(in room)
- is unaware of wizard
- discusses their own photos
- engages with PhotoPal
2
3
THE ASSISTANT
(in room)
- introduces setup
- could remotely control iPhoto
- observes and takes notes
- removes "camera" and leaves room
- marks up photos in s-sighted version
THE WIZARD
(remote location)
is "voice" of PhotoPal -
controls CT and TTS over skype -
could remotely control Aperture/Flickr -
observes covert camera -
starts conversation after assistant leaves -
EMBODIED ARTEFACTS
(physical companions,
photos still on screen)
skype
flickr
Overt camera removed by assistant at "end of session," however covert camera's used to continue "unscripted" conversation.
1 overt camera
2 covert camera (user)
3 covert camera (photopal system)

Figure 1: The WOZ1 setup used to collect data

2. Specifications and requirements

The Companions data described here was all collected using a slightly modified Wizard of Oz method. In a normal Wizard of Oz experiment, the user believes they are talking to a system while in fact the system is being controlled by a human being (Strauss et al., 2006). In our case the emphasis was on collecting naturally occurring dialogues relevant to the domain rather than obscuring the use of humans. For the SC, a number of subjects were asked to reminisce about photos. In early experimental stages, the photos were random publicly available images, but for the experimental set up to work properly the scenario needed people to reminisce about photos of personal importance to them. This added a further constraint to the process to which we will return below.

The experimental design was to have a person talk about their photos to the (WoZ) experimenter in another room. The photos were primarily about family and friends; gatherings such as birthdays and weddings were considered ideal examples. Guidelines were provided for the experimenter play the WoZ, including a contextual account of what the interviewer was trying to achieve.

The aim of the interviewer is to build up an image of what it is that the subject is talking about in their photos. This image is partial but, as in the game of 'Battleships' small pieces of information in conjunction with a model of the domain may be enough. The location of things in the image will also be useful. If the subject says that this is a photo of Sue and Jane, the system might ask which one is Sue and ask the subject to point at Sue with the mouse cursor or on a touch screen. At a later date, the system might use the data about Sue in the photo to ask if the subject is talking about this particular (brings up image and points) Sue. Such a mechanism would be a natural way to have the subject go back over old photos and hopefully enjoy the process of reminiscing about them. . The user should be encouraged to express feelings, attitudes memories, but turns should not be too long. The wizard, when driving the dialogue, should do so from the questions below - though not in any fixed scripted order and from the named entities recognised in the dialogue so far as provided by the user. — from the Companions WoZ guide.

The Wizard was instructed to choose from a standard set of questions examples of which include the following:

- How many people are in the picture?
- What are their names?
- Which side of the photo is each [name] on?
- Are they looking at us, or each other, or an object?
- What is their relationship?
- What date does it seem to be roughly?

One of the assumptions which has been borne out by subsequent work was that the dialogue system would have access to sufficient image processing data to allow it to "know" whether there were people in the picture or not and, if so, how many. The current demo used the OpenCV Library from Intel.

The set up for capturing the audio from the interviewed subjects needs to be also carefully specified, as the collected data will be used for training and/or adaptation of the automatic speech recognition systems tuned to the domain. It is desirable to use high-quality close-talking (or lapel) microphone and to keep the background noise at minimum.

The recorded speech then needs to be segmented (roughly) into sentences and carefully transcribed. The software tool Transcriber (Barras et al., 2001) is employed for this

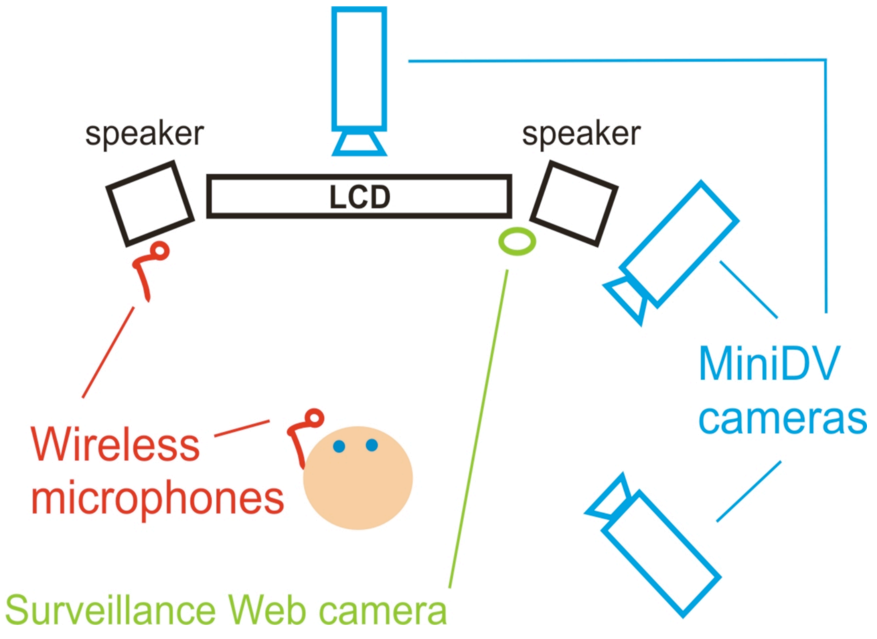

Figure 2: The Czech WoZ Recording Room Set up

purpose. The time alignment between speech and discussed photographs also needs to be stored for the dialogue-management purposes.

3. The Senior Companion Data Set

The original concept for the PhotoPal aspect of the SC is that, through having a conversation about photos with PhotoPal, the system can automatically tag a photo with some far more complex metadata than the simple user-selected feature system used widely in systems like FLICKR (http://www.flickr.com/). This would allow retrieval later on by any of the words, phrases or specific metatdata contained in the conversation, and we are also experimenting with coded RDF triples expressing semantic relationships, derived from the AKT project (http://www.aktors.org/akt/). In a classic WoZ experiment the participants are not meant to know that a human is playing the part of technology. However, the effort required on the part of the wizard can be quite considerable if the output is given from text to speech (TTS), in which case the wizard has to type in responses on many occasions (despite the wizard having a good interface), and this quickly becomes tiring. Accordingly we have experimented with collecting dialogues between participants and both an avatar wizard (WoZ1) and participants and a human wizard (WoZ2). As yet, there do not seem to be major differences between dialogues collected in these different ways.

In both WoZ1 and WoZ2 scenarios, photos are displayed one at a time. For some subjects this is not a particularly interesting activity to engage in and conversations can begin to tail off after approximately 20 minutes. A second stage prototype has therefore been developed that allows people to engage with groups of photos and this is expected to lead to more interesting dialogues that are more typical of how the interaction with a final system will be. We await the results of further experimentation with the multiple photos in further version of the PhotoPal SC.

By September 2007, 45 data collection sessions have been undertaken generating approximately 30 hours of transcribed dialogue, as follows:

- Participant Breakdown
  - 27 male participants (age range 22 73)
  - 13 female participants (age range 19 69)
- Locations
  - 7 in home sessions
  - 38 in Napier University labs
- WOZ Versioning
  - Version 1 (with avatar) 16 session
  - Version 2 (without avatar) 29 sessions
- Hardware Overview
  - Computer Internal Mic 9 sessions
  - Onboard Camera Mic 4 sessions
  - Belkin TuneTalk Stereo for iPod 32 sessions[1]

On the Czech side of the data collection, we have decided to stick to the set up with avatar wizard. We have implemented new user interface equipped with the Czech 3D talking head avatar developed at U. of West Bohemia (Železný et al., 2006). A dedicated room has been established for recording purposes its set up is sketched in Figure 2.

[1] All subsequent sessions have utilised this mic for ASR analysis purposes.

The subject sees just the photo being currently discussed and the avatar on the screen. Speech from both the person and the avatar[2] is captured by a high-quality Senheiser microphones and recorded to a computer placed in another room (in order to minimise noise), sampled at 22 kHz. The times when the displayed photo is changed are logged. The speaker is also recorded by three miniDV cameras simultaneously the video data are currently just being archived and are intended for future use in audiovisual speech recognition, emotion detection, gesture recognition, etc.
So far we have recorded 32 people, 20 female (avg. age 70.25 years) and 12 male (avg. age 71.67). The average length of the session is 54:46 minutes and average number of photos discussed is 8.47. Our "subjects" tend to spend a lot of time on a single photo (our wizards need to end the interview after approx. 55 minutes because of the miniDV tape length which is 1 hour). In contrast with the observations by Napier University, the conversation usually does not tail off quickly on the contrary, our human wizards often have to make serious effort to gently interrupt a stream of memories and make the subject move to the next photo. This is most probably related to the choice of subjects our first observations suggest that people that are already retired (who were the vast majority of our subjects) tend to reminiscence a lot more than the people who are still working. However, this fact merits further investigation.

4. Future Data in Companions

Elsewhere we have set out the initial and more sophisticated learning mechanisms we are applying to the Companions corpus to derive an initial Dialogue Management system (Pinto et al., 2008)[3], based on a stack and Dialogue Action Forms (DAFs) intended to capture overall dialogue context. Our methodology for obtaining fresh data from this initial implementation relies on the PhotoPal part of the SC being system initiate driven to a large extent, although the SC as a whole is very much mixed initiative with the user able at any time to change the topic and force putting a new DAF on the dialogue management stack. In the PhotoPal component we able to run the system repeatedly with the same images but different users and gain a wide range of possible responses to known questions, thus enabling the building of a wide response set, both for named entities, and for forms of question answer with which we are seeking systematically to augment the DAF structures.

5. Data Availability

All data referred to here including transcripts and audio files are available from the Companions web site; http://www.companions-project.org

6. Acknowledgements

This work was funded by the Companions project (http://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

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 2001. Transcriber:development and use of a tool for assisting speech corpora production. *Speech Communication*, 33:5–22. Special issue on speech annotation and corpus tools.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development . acoustics,. In *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, volume 1, pages 517–520.

Hugo Pinto, Yorick Wilks, and Roberta Catizone. 2008. The senior companion multiagent dialogue system. In *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 08)*.

P. Strauss, H. Hoffmann, W. Minker, H. Neumann, G. Palm, and S. Scherer et al. 2006. Wizard-of-oz data collection for perception and interaction in multi-user environments. In *International Conference on Language Resources and Evaluation (LREC)*.

Miloš Železný, Zdeněk Krňoul, Petr Císař, and Jindřich Matoušek. 2006. Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. *Signal Processing*, 86(12).

Yorick Wilks. 2005. Artificial companions. *Interdisciplinary Science Reviews*, 30:145–152(8), June.

[2] Naturally, only the speech from the interviewed person will be used for ASR training the speech from the avatar is nevertheless important for dialogue tagging.

[3] See also http://companions-project.org/research/deliverables.cfm