Evaluating Human-Machine Conversation for Appropriateness

Nick Webb[1], David Benyon[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,mival}@napier.ac.uk, preben@sics.se

Abstract

Evaluation of complex, collaborative dialogue systems is a difficult task. Traditionally, developers have relied upon subjective feedback from the user, and parametrisation over observable metrics. However, both models place some reliance on the notion of a task; that is, the system is helping to user achieve some clearly defined goal, such as book a flight or complete a banking transaction. It is not clear that such metrics are as useful when dealing with a system that has a more complex task, or even no definable task at all, beyond maintain and performing a collaborative dialogue. Working within the EU funded COMPANIONS program, we investigate the use of appropriateness as a measure of conversation quality, the hypothesis being that good companions need to be good conversational partners . We report initial work in the direction of annotating dialogue for indicators of good conversation, including the annotation and comparison of the output of two generations of the same dialogue system.

1. Introduction

The development of conversational agent technologies in the EU-funded COMPANIONS project[1] requires us to examine new models of dialogue system evaluation. Companions are targeted as persistent, collaborative, conversational partners, where the user may have a wide degree of initiative in the resulting interaction. Rather than singular, focused tasks, as seen in the majority of deployed dialogue systems, fully developed Companions can have a range of tasks and can 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 with information on who or what is in the picture. It may be that the definition of the task itself is one of maintaining a relationship, not something that can be measured using traditional metrics such as *task completion*. When devising an evaluation paradigm for such systems, we need to balance the completion of tasks (where measurable) with some measure of "conversational performance". The assumption in traditional dialogue evaluation is that the manner in which the system handles the conversation is covered by *user satisfaction*. That is, if the resulting dialogue is annoying or repetitive, we expect a corresponding drop in user satisfaction. However, user satisfaction is in some sense a composite score, covering the entire interaction. We have seen for example that poor text-to-speech component performance can have a disproportional effect on user satisfaction.

Instead, we want to develop a method of scoring conversational performance directly; measuring the systems capability to maintain a conversation based on the progression of the dialogue so far. We believe that this conversational performance can be measured in terms of appropriateness, as demonstrated by Traum et al. (2004).

The motivation for developing new evaluation techniques was the on-going development of Companion technologies, the first prototypes of which include a personal automated 'Senior Companion' (SC), that will learn about its owner's life story, needs, and preferences through free-ranging natural language dialogue driven by a collection of users' photographs, and the 'Health and Fitness Companion' (HNF) that allows the user to plan their day in terms of exercise (how they get to work, for example, via bicycle or car), leisure activities and diet. We evaluated three manifestations of the Companions concept, the 'Senior Companion' (SC) (Catizone et al., 2008), the 'Health and Fitness Companion' (HNF) (Turunen et al., 2008) and the mobile HNF (Ståhl et al., 2008), a version of the HNF designed to work on mobile platforms and PDAs.

In this paper, we discuss the use of appropriateness annotation for evaluating these systems. Appropriateness is a measure of each utterance made by the system, where human annotators score the level of appropriateness given the utterance's level of information and the progression of the dialogue so far. Measures of appropriate interaction can be used to penalise mechanisms that are seen as inappropriate, such as over-verification, strong, one sided initiative and repetitive behaviour. Several researchers (Simpson and Fraser, 1993; Hirschman and Thompson, 1996) have looked at using a mechanism of appropriateness of dialogue as a measure of an effective communication strategy. We principally explore the application of appropriateness as described in the work of Traum et al. (2004).

2. Related Work

There has been a significant amount of effort spent on evaluating spoken language dialogue systems (cf. (Danieli and Gerbino, 1995; Minker, 1998; Walker et al., 2002)), most of which relies on a combination of observable metrics and user feedback. Efficiency and effectiveness metrics include the number of user turns, system turns, and total elapsed time. For the "quality of interaction", it is usual to record Automatic Speech Recognition rejections, time out prompts, help requests, barge-ins, mean recognition score (concept accuracy), and cancellation requests. Note these

[1] www.companions-project.org

(62) Speaker User: *AppTag="Response Received"*: **tell me a joke**

(63) Speaker System: *AppTag="Appropriate Response"*: **The Software Development Process: 1. Order the T-shirts for the Development team, 2. Announce availability, 3. Write the code, 4. Write the manual, 5. Hire a Product Manager, 6. Spec the software (because writing the specs after the code helps to ensure that the software meets the specifications), 7. Ship, 8. Test (the customers are a big help here), 9. Identify bugs as potential enhancements and 10. Announce the upgrade program**

(64) Speaker User: *AppTag="No Response: Appropriate"*:

(65) Speaker System: *AppTag="Appropriate Question"*: **What is your relationship to Libby?**

(66) Speaker System: *AppTag="Inappropriate Response"*: **I'm sorry I didn't understand your relationship to Libby.**

...

(70) Speaker User: *AppTag="Response to System"*: **Could be as my friend**

(71) Speaker System: *AppTag="Filled Pause"*: **Thank you**

Figure 1: An interaction with the Senior Companion

are somewhat functional descriptors of quality of interaction.

The DARPA Communicator Program made extensive use of the PARADISE metric (Walker et al., 1997). PARADISE (PARAdigm for DIaLogue System Evaluation) was developed to evaluate the performance of spoken dialogue systems, in a way de-coupled from the task the system was attempting. 'Performance' of a dialogue system is affected both by *what* gets accomplished by the user and the dialogue agent working together, and *how* it gets accomplished, in terms of the quality measures indicated above. In other words, PARADISE aims to maximise task completion, whilst simultaneously minimising dialogue costs, measured as both objective efficiency of the dialogue (length, measured in total turns for example) and some qualitative measure. A consequence of this model is that often the dialogue quality parameters are tuned to overcome the deficiencies highlighted by the observable metrics, such as discussed in Hajdinjak and Mihelič (2006). For example, using explicit confirmation increases the likelihood of task completion, and so is often chosen, despite being regarded as somewhat unnatural in comparative human-human speech data.

3. Evaluation of Companions

We initially evaluated three manifestations of the Companions concept; the Senior Companion (SC), the Health and Fitness Companion (HNF) and the mobile HNF.

The mechanisms for evaluation were two-fold. Qualitative surveys were used to acquire subjective opinions from the users of the Companions prototypes, in conjunction with quantitative measures such as Word Error Rate and Concept Error Rate. We analysed 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.

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 ASR software before its associated introductory tutorial. This voice training exercise took five to seven minutes to complete. Each session then 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. Participants then used the SC, HNF and mobile HNF for 10-15 minutes each, completing the on-line questionnaire after each session. Participants were video taped and recorded during their interaction. Finally the participants were each interviewed by a researcher.

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 SC and HNF 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 HNF has a push to talk feature.

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.

Dialogues with the 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. HNF 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 HNF. It can be seen that there are significant

(16): Speaker System: *AppTag="Appropriate Question"*: **Alright. How about ordering lunch from a takeaway?**

(17): Speaker User: *AppTag="Response to System"*:

(18): Speaker System: *AppTag="Request Repair"*: **Sorry, I didn't understand. Can you re-phrase: How about ordering lunch from a takeaway?**

(19): Speaker User: *AppTag="Response to System"*: **That is too splendid**

(20): Speaker System: *AppTag="Appropriate Question"*: **So, you are going to take the bus part of the way and walk the rest of the way to work?and you are going to order takeaway for lunch?Are both of those correct?**

Figure 2: An interaction with the Health & Fitness Companion

differences in how verbose different people are. While the small dataset does not allow statistical testing, the utterance lengths used by one person with the SC seem to be rather well in line with their utterance lengths with HNF. Comparing the actual utterance lengths with ASR results, ASR in the SC recognises fairly closely the same amount of words as uttered, while HNF 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 HNF 12 words. Vocabulary size used by people with SC ranged between 33 and 131 words, while HNF resulted in vocabularies between 18 and 116 words. The average of these is 70 for SC and 55 for HNF. The larger vocabulary of SC dialogues is to be expected due to systems more open questions; in fact, it is somewhat surprising that the vocabulary is 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 HNF 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 HNF are extremely high, concept error rates are somewhat smaller, between 0.33 and 0.65.

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). Twenty seven responses were collected. The questions were organised around six themes:

- The behaviour of the Companion and what it looked like
- The utility of the Companion
- The nature of the relationship between participant and Companion
- The emotion demonstrated by the Companion:
- The personality of the Companion
- The social attitudes of the Companion

The HNF used the same set of questions, but allowed for people to provide additional comments to explain their choice. Eight responses were collected in total. More details of the metric based evaluation can be seen in Benyon et al. (2008).

4. Appropriateness Mark-Up

In order to capture appropriateness of dialogue, annotation of the resulting dialogue transcripts was required. Annotators used a scheme that splits the system and user utterances (here, utterance is used to mean a single unit of information in the dialogue sense, where a user turn can be made up of several such utterances each corresponding to a single piece of information) and codes each utterance with one of several annotations, as seen in Figure 3. For users, there are four grades of annotation, three of which come directly from the work of Traum et al. (2004); those utterances that elicit a response (RES); those where no response was received, and this was appropriate behaviour (NRA); and those where no response was received, and this was deemed inappropriate (NRN). Initial trials showed that annotators were often confused as to what to annotate a reply by the user to a system question, so we added a fourth category, response to system (RTS).

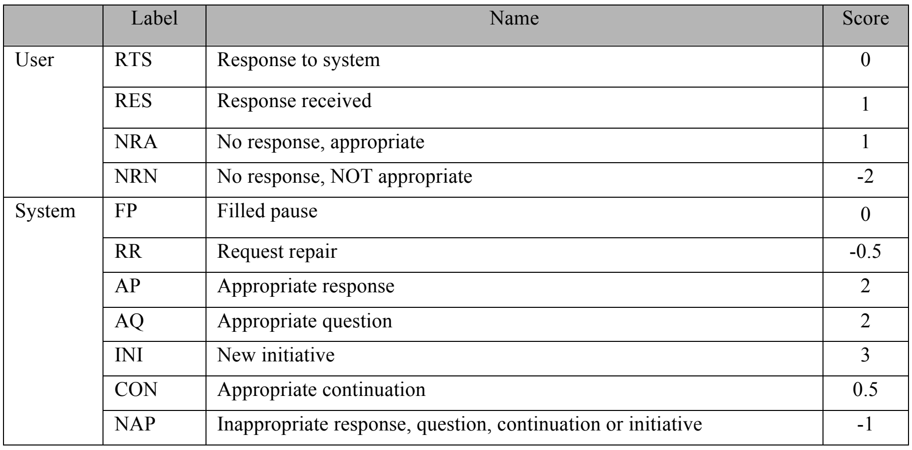
As this appropriateness annotation process is targeted at the role of the artificial agent in interactions, there are more options for annotating system responses. For agent utterances, there are seven categories. There are filled pauses (FP), requests for repair (RR), appropriate responses (AR), appropriate new initiatives (INI), appropriate continuations (CON) and finally inappropriate responses, initiatives or continuation (NAP). Again, annotators were confused as to how to annotate questions asked by the system that were appropriate, so we added the category appropriate question (AQ).

4.1. Example Annotations

For each of the categories listed above, an example is given, taken from our annotation guideline manual. In each case the utterance *italicised* is the utterance under discussion.

RTS: Response to system

The user gives an utterance that is a direct reply to a system question. It does not matter if the response is appropriate or not.



	Label	Name	Score
User	RTS	Response to system	0
	RES	Response received	1
	NRA	No response, appropriate	1
	NRN	No response, NOT appropriate	-2
System	FP	Filled pause	0
	RR	Request repair	-0.5
	AP	Appropriate response	2
	AQ	Appropriate question	2
	INI	New initiative	3
	CON	Appropriate continuation	0.5
	NAP	Inappropriate response, question, continuation or initiative	-1

Figure 3: Tags used to annotate appropriateness

Example:

SYSTEM: Who is in this picture?
USER: Alan

RES: Gets response

The user gives an utterance (a request or a directive) that gets a response from the system (that response need not be appropriate).

Example:

USER: I'd like to read the news
SYSTEM: Ok, here is the news

NRA: No response, appropriate

The user gives an utterance that doesnt get a response from the system, but this is appropriate behaviour.

Example:

USER: Please be quiet
SYSTEM: ¡blank¿

NRN: No response, not appropriate

The user gives an utterance that doesnt get a response from the system, but that is NOT appropriate behaviour.

Example:

USER: Whats your name?
SYSTEM: .....

FP: Filled pause

A statement, or content free filler, or an automatic response action. Contains neither a question, or new initiative.

Example:

SYSTEM: thanks

RR: Request for repair

When the system fails to understand input, and requests help, either explicitly (can you repeat that) or implicitly (by asking the same question again).

Example:

SYSTEM: Im sorry, I didnt understand, could you repeat that?

AP: Appropriate response

The system responds appropriately to some utterance or request, using knowledge or context of the dialogue.

Example:

USER: Who did I say is in this picture?
SYSTEM: I see Bob and Roger and David

AQ: Appropriate question

The system asks an appropriate question given its level of knowledge and the context of the dialogue.

Example:

USER: Thats my Uncle Bob
SYSTEM: And how old is Bob in this picture?

INI: Appropriate new initiative

When the system moves the dialogue situation along by introducing a new task or direction.

Example:

USER: I dont know anyone in this picture
SYSTEM: Ok, why dont we talk about another picture?

CON: Appropriate continuation

When the system makes extended contributions to the current context.

Example:

USER: This was taken in Rome.
SYSTEM: I love Rome.
SYSTEM: It's where I met my first wife.

NAP: Inappropriate response, continuation, question or initiative

When the system responds in a way inappropriate to the level of understanding or situation in the dialogue. Consider the dialogue context (i.e. what the system can be said to understand at this stage). Only consider what has been recognised, not what was actually said.

Example:

USER: I dont know the person in this picture, I never met them
SYSTEM: Ok, so when were they born?

4.2. Scoring

Each of the resulting annotations over the transcript is then scored. At this stage, we imported the scoring intuitions used by Traum et al. (2004), although it is not clear that these represent the most effective scores for Companions dialogues, something we shall explore in later work. We give the scoring values for each of the annotations, and the corresponding intuition that the scoring is meant to capture. First, filled pauses are graded as generally human-like, and good for virtual agents to perform, but don't add a lot (score:0). Appropriate responses and questions are very good (score:+2), but even better are initiatives that push the interaction back on track (score:+3). Extended contributions, where the system adds additional information to something initiated earlier, are fine (score:+0.5). Repairs and clarifications are bad in their own right (score:-0.5), but their use can still gain points by allowing subsequent appropriate response. For example, if it takes 2 dialogue moves to complete a repair (with a combined score of -1), that then leads to an appropriate response (and receives a score of +2), then we still reward this sub-part of the interaction with an overall score of +1. Finally, inappropriate responses are bad (score:-1), but no response is worse (score:-2). For those familiar with reward-oriented approach to dialogue modelling, it can be seen that such an evaluation methodology can be used to assign rewards to complete and partial dialogues.

5. Appropriateness Evaluation

For the first phase of our evaluation, we had 8 users complete the entire protocol, that is, interact with both the SC and the HNF. All participants were native English speakers without strong accents, whose ages ranged from 27 to 61. Of the participants, 2 were female, and 6 were male. Each participant had to complete four distinct tasks: Introductory tutorials; Using the prototypes; On-line surveys; and post-interaction interviews. Shortly after the initial evaluation exercise, we received an updated version of the Senior Companion (that we shall refer to as version 2), and we repeated the entire evaluation using this new version. This time there were 12 total participants, 9 male and 3 female, with ages from 21 to 38. The key differences between version 1 of the SC and version 2 were:

- version 2 interfaced with Facebook. This meant that users had access to their own photograph collections, and if any photographs had been tagged with the identities of individuals, these names were already known to the Companion.
- the A.L.I.C.E. chat-bot[2] was integrated into the Companion. When there was a misrecognition, for example the Companion asks for location information, but the NLP module fails to identify any matching input, the chat-bot would be called using whatever had been recognised as input. Often, the simple pattern matching of the chat-bot would result in meaningful information being retrieved from the web, and used to construct an on-topic comment or question. For more about this integration, please refer to (Field et al., 2009).

Everything else about the SC demonstrator remained the same, and scored around the same in a range of key observable metrics (such as WER and CER, for example). However, we were able to determine that version 2 of the SC elicited both different behaviour and different reactions from users than version 1. For example, the average number of words per utterance from the user increased from 4.27 with version 1 to 6.1 with version 2, a 43% increase. The feedback from user surveys also showed a significant improvement in user satisfaction with the system, with more users finding the Companion 'engaging' *and* indicating that the Companion 'demonstrated emotion' at times.

With subjective and objective evaluations complete, we have XML files containing all user-system interactions, and a sense that, over two versions of the same prototype, improvements in the interaction have been made. We wanted to apply the annotation scheme described in Section 4., and determine the score for each *dialogue* and for each system as a whole, to see if we could characterise errors and capture improvements in the dialogue. For this evaluation, all dialogues were annotated by a single user. However, to check preliminary consistency, one dialogue of the HNF and one of the SC was annotated by 3 additional annotators, with no training other than access to the annotation

[2] `http://alicebot.blogspot.com/`

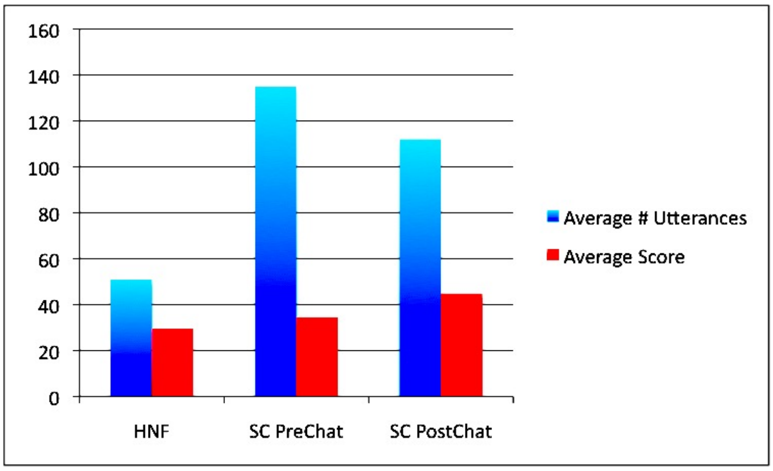


Figure 4: Average number of utterances, and average score, per system

guidelines. Several of the annotators had no prior experience with dialogue systems. We then computed inter-coder reliability among all annotators, according to Krippendorf's alpha. For the SC, we computed $\alpha = .80$, and for the HNF $\alpha = .73$. Both of these, particularly the score for the SC dialogues, are good indicators that the annotators are able to reliably encode the categories as described. The difference in α scores can be explained by the differences in utterance segmentation, with the HNF often having longer, more complex utterances.

To understand how the appropriateness annotations were applied, we discuss some examples with reference to Figures 1 and 2. In Figure 2, a dialogue with the HNF, the systems asks a question of the user (utterance 16), which is marked as appropriate for this stage in the dialogue. The user replies (17), but nothing is captured by the system. The system initiates a repair (18) to which the user replies (in 19). Not shown here, but in the original XML file, is the actual user utterance ('that is too expensive'), but what is shown here is what the system recognises. This is important to note: appropriateness captures the systems response to what has been understood, and *not* what the user actually says. In the case of mis-recognitions, we would expect the user to be primarily responsible for correcting any obvious errors. The dialogue concludes with an appropriate question (20). There is a case to be made that the formulation of the question in (20) is some sort of check, or confirmation, and that it may be appropriate to annotate this as such. We discuss this later in Section 7..

In Figure 1 there is a dialogue with the Senior Companion. This fragment begins (62) with a user initiating a request for a joke, which the system recognises and responds appropriately (63). The next user utterance is not recognised at all by the system (64), so instead the system returns to the task prior to the joke request, talking about some underlying set of photographs (65). However, the system *immediately* enters a loop; this cannot be seen explicitly in the data represented here, but the timing information in the XML files shows that there is less than a second between utterance (65) and utterance (66) and this is marked as inappropriate behaviour. Indeed, the system enters an error loop, as utterances (67) through (69) are repeats of utterance (66)and all are marked as inappropriate. Finally, in utterance (70) the user finally answers the original question and the system recognises this input, and thanks the user (71). Again, there is an ambiguity here; we may at some later date want to encode utterances such as (71) as an explicit use of politeness but for the moment we capture them as essentially meaningless filled pauses.

5.1. Analysis

Once all dialogues have been annotated, we use the scoring mechanisms outlined in Section 4.2. to calculate average scores for each system. A summary of the average dialogue score, and a comparison with the average number of utterances can be seen in Figure 4. There is some notion that for Companion-like technologies, longer conversations are better (a metric often referred to as 'stickiness', and at odds with goal-oriented systems, where shorter interactions are better), indicating a willingness of the user to converse with the technology, but of course the complexity of the underlying domain and it's open-ended nature also has a direct effect. As an alternative analysis, if we normalise the results by length of dialogue we see in Figure 5 that the resulting average score per utterance of the dialogue favours the HNF. This is also a potentially useful score, that indicates that there are more positive contributions made by individual utterances of the HNF.

Where comparison is most useful at this indicative stage is between two versions of the same Companion. From Figure 4 it is possible to see that with version 2 of the SC, the average length of the dialogue decreases, but the overall average score *increases*. In Figure 5 this is confirmed, as the average contribution per utterance in version 2 of the SC (the post chat-bot system) increases, from 0.26 to 0.4, representing a 54% increase per utterance. In order to understand in more detail what was happening with each prototype, we performed an analysis of the distribution of tags across the systems, as can be seen if Figure 6. We can see that with version 2 of the SC there is a significant increase in the number of appropriate questions asked, at the same time as a significant decrease in requests for repair.

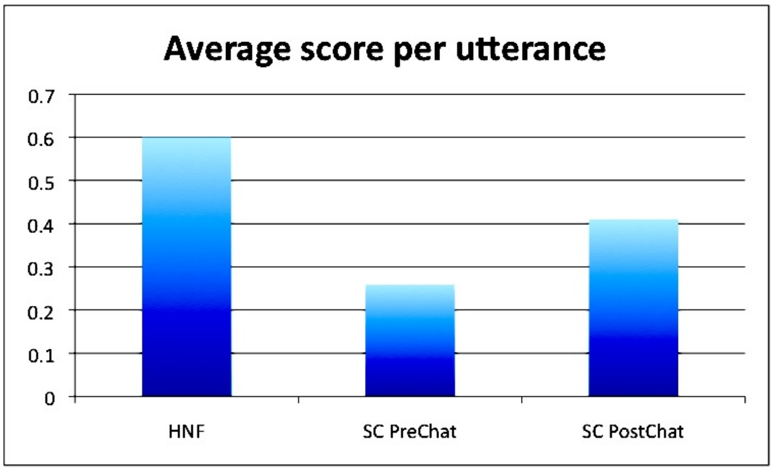


Figure 5: Average score of *each* utterance, per system

In order to understand in more detail what was happening with each prototype, we performed an analysis of the distribution of tags across the systems, as can be seen if Figure 6. We can see that with version 2 of the SC there is a significant increase in the number of appropriate questions asked, at the same time as a significant decrease in requests for repair. We can determine from the transcripts that this is due to the deployment of the chat-bot technology which at times performs a query over the internet, and replies with a relevant question or continuation. For example, if the user replies that the location of a photograph is "Edinburgh", asks if the user has ever been the the Royal Mile (a well known tourist site). Even if the information from the user is mis-recognised, the chat-bot can be used to ask a seemingly appropriate question given the context of the dialogue, resulting in less repair requests. One side effect of such a mechanism is that in addition to the increase in appropriate questions, there is an increase in *inappropriate* questions or statements, as sometimes the information retrieved from the web is incorrect. However, this is outweighed by the number of times the information is correct and is still seen by the users (judging by the subjective surveys) as appropriate conversational performance.

6. Conclusions

In this early phase of evaluating Companions prototypes, we have created an evaluation pipeline that has generated baseline objective and subjective performance measures. These measures are useful to show improvements over subsequent versions of Companions prototypes. For example, earlier we showed that there had been a 42% increase in the words per utterance coming from the user, and that there had been improvements in the subjective user feedback, when transitioning from version 1 to version 2 of the Senior Companion.

By annotating the resulting output files, we are able, with the appropriateness annotation, to mirror this improvement, which indicates that for future new versions of prototypes, we can collect some sample dialogues with the new system, annotate them, and hope to predict changes in user satisfaction., although this requires significant further exploration. We have also established a set of reliable annotation guidelines, and further created a tool, the Appropriateness Annotation Tool, that enables users to annotate dialogues with more ease.

7. Discussion

It is important to note that these annotation metrics are performed by hand, and the scoring based only on an intuition of appropriate dialogue behaviour. In a Companion scenario one might want to weight the measures in alternate ways. For example, we may give increased reward for extended contributions, when the system talks about pictures, and the user is in listening mode. Further, the current scheme conflates some issues, such as responsiveness and appropriateness of response, which we may wish to tease apart further. On the subject of granularity, we already mentioned the possibility of adding new tags (or dimensions of existing tags) to capture, for example, the use of emotion or politeness. More specifically, as we are interested in the inappropriate use of dialogue, we might want to expand the categories of annotation to include the use of inappropriate knowledge, or behaving repetitively.

Importantly, whilst we have these measures for the baseline of system performance, we have no measure of the possible upper bounds of performance. We propose to perform several Wizard of Oz (WoZ) experiments, with subjects determined to be our likely end user group. We will then perform another evaluation over data generated from these experiments, and refine the annotation scheme and scoring weights using this data as our target.

What we do have is a record of those sub-parts of dialogue that are regarded by annotators as inappropriate dialogue. These sections can be passed to developers, for them to determine which part of the prototypes can be updated to improve dialogue performance, something that may not be possible with global user feedback. In the case of mechanisms such as reinforcement learning, appropriateness measures can be used to score both whole dialogues and dialogue sub-structures.

Finally, there are other models of dialogue coherence, such as the work of Artstein et al. (2008) that could work in com-

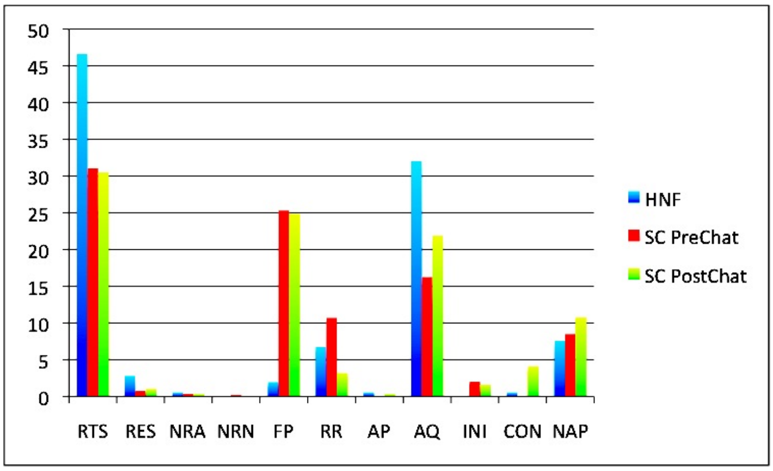


Figure 6: Distribution of tags (*as %*) across systems

bination with measures of appropriateness. Finally, all evaluations with Companion technology must include a longitudinal study, where we can measure users reactions and changing behaviours over time.

8. 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.

9. References

R. Artstein, S. Gandhe, A. Leuski, and D. Traum. 2008. Field Testing of an Interactive Question-Answering Character. In *Proceedings of the ELRA Workshop on Evaluation, at Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

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.

R. Catizone, A. Dingli, H. Pinto, and Y. Wilks. 2008. Extraction tools and methods for Understanding Dialogue in a Companion. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Morena Danieli and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford.

Debora Field, Roberta Catizone, WeiWei Cheng, Alexiei Dingli, Simon Worgan, Lei Ye, and Yorick Wilks. 2009. The Senior Companion: a Semantic Web Dialogue System. In *Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems*, Budapest, Hungary.

M. Hajdinjak and F. Mihelič. 2006. The PARADISE Evaluation Framework: Issues and Findings. *Computational Linguistics: Special Issue on Empirical Studies in Discourse Interpretation and Generation*, 32:263–272.

L. Hirschman and H. S. Thompson. 1996. Overview of evaluation in speech and natural language processing. In R. Cole, editor, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press.

W. Minker. 1998. Evaluation methodologies for interactive speech systems. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pages 199 – 206, Granada, Spain.

A. Simpson and N. Fraser. 1993. Blackbox and glassbox evaluation of the SUNDIAL system. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 1993)*, Berlin.

O. Ståhl, B. Gambäck, P. Hansen, M. Turunen, and J. Hakulinen. 2008. A Mobile Fitness Companion. In *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy.

D. Traum, S. Robinson, and J. Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1699–1702, Lisbon, Portugal.

M. Turunen, J. Hakulinen, O. Ståhl, B. Gambäck, P. Hansen, M.C. Rodrguez 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. In *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy.

M. Walker, D. Litman, C. Kamm, and A. Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Jan.

M. Walker, A. Rudnicky, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, R. Prasad, S. Roukos, G. Sanders, S. Seneff, and D. Stallard. 2002. DARPA Communicator Evaluation: Progress from 2000 to 2001. In *ICSLP*.